

Identity Deception Detection on Social Media Platforms

by

Estée van der Walt

Submitted in partial fulfilment of the requirements for the degree
Doctor of Philosophy (Information Technology)
in the Faculty of Engineering, Built Environment and Information Technology
University of Pretoria, Pretoria

September 2018

Identity Deception Detection on Social Media Platforms

by

Estée van der Walt

E-mail: <mailto:estee.vanderwalt@gmail.com>

Abstract

Whenever they interact on big data platforms such as social media, humans run the risk of being targeted by other malicious individuals. The protection of these humans is problematic, even though cyber-attack events have received much attention in public in order to create greater awareness. Cyber protection is difficult, largely due to the nature of these social media platforms (SMPs), as they allow individuals to create and use almost any persona they choose, with minimal validation. This, together with the sheer volume of data being generated, warrants the use of automated threat detection methods. The victims are targeted through different forms of cyber threats of which identity deception is but one example.

Identity deception is by no means a novel concept. The social sciences, more specifically psychology, have for many years attempted to understand the motive(s) behind human deception. More recently, research work aimed at finding fake or bot accounts has had some success. This study used the abundant knowledge about identity deception, in addition to the information already available by default on SMPs, to detect those humans who lie about their identity.

The study in hand presents an SMP research environment with a methodology and prototype that will assist with the automated detection of human identity deception on SMPs. This environment enables various supervised machine learning experiments. The first experiment used those attributes describing an SMP profile to detect identity deception with a final F1 score of 32%. The second experiment added additional features known to detect deceptive bots on SMPs with a final F1 score of 49%. The third experiment added further features from psychology, known to identify deceptive humans, with a final F1 score of 86%. A critical evaluation of the SMP attributes and engineered features reveals that age, name, and location contributed most towards identity deception as executed by humans in respect to other humans. The results show that human identity deception detection can be achieved by using SMP attributes and engineered features that describe the identity of the individual only, thus excluding SMP content that can be costly to construe.

The prototype furthermore includes an Identity Deception Detection Model (IDDM) that scores a human's perceived deceptiveness and intuitively explains the score. The IDDM not only indicates when a human is potentially deceptive but also highlights those attributes or features that were most prevalent in the conclusion. This aids investigators, like the police force, to not only identify potential deceptive humans, but to also make a more informed decision.

The results from this research make a significant contribution to the fields of both cyber security and the social sciences.

Keywords: Identity Deception, Cyber Security, Social Media, Big Data, Data Science, Social Science, Bots, Psychology.

Supervisor : Prof. J. H. P. Eloff

Department : Department of Computer Science

Degree : Doctor of Philosophy

Acknowledgements

I would like to thank the following people and institutions for their assistance during the completion of this thesis. Without them this work would not have been possible:

- Prof. Jan Eloff whose insights and guidance was invaluable. My research work matured immensely under his guidance. Thank you for many fruitful discussions around the research topic and always pushing my research to the next level.
- The HPI Future SOC lab in Potsdam, Germany who provided the resources used during this research. Their infrastructure enabled the research work to be completed at scale and in a timely manner. In addition, the lab assistants were always available for help and support when needed.
- Dr. Jacomine Grobler whose knowledge and additional insights at a later stage of the research guided some of the directions which gave the research work more depth.
- Isabel Claassen for the timely editing of this thesis and related work throughout.
- All my friends and family for their continued support throughout.
- My parents for their motivation and support from an early age to follow my dreams. They instilled in me a drive to become the best version of myself and to never give up.
- My grandmother for her enthusiasm and interest in my studies. Thank you for always asking how it is going with very little knowledge about the topic.
- Wendy Stephens who has been my pillar of strength throughout most of this process. Thank you for looking after my well-being and always understanding with unwavering love and support.
- Last but not the least, my Creator whom without any of this was impossible.

Contents

List of Figures	vi
List of Algorithms	ix
List of Tables	x
I Introduction, Significance of and Background to the Study	1
1 Introduction	2
1.1 Introduction	2
1.2 Research problem	6
1.3 Research questions	6
1.4 Research scope	7
1.5 Methodology used	7
1.6 Layout of thesis	9
2 Big data and Social media platforms – a cyber-security view	13
2.1 Introduction	13
2.2 Big Data	14
2.2.1 Big data case studies	15
2.2.2 Big data characteristics	16
2.3 Social media platforms	20
2.3.1 Social media data	21
2.4 Cyber security within social media platforms	25
2.4.1 Related research on cyber threats found on SMPs	26
2.5 Related research on identity deception on SMPs	28
2.6 Conclusion	32

3	Cyber Security and Identity Deception	34
3.1	Introduction	34
3.2	Deception	35
3.3	Defining an identity	36
3.4	Identity Deception	37
3.4.1	Who commits identity deception on SMPs?	39
3.4.2	Attributes and features used to detect identity deception on SMPs	40
3.4.3	Methods used to detect identity deception on SMPs	43
3.5	The requirements for the detection of identity deception by humans on social media platforms	47
3.6	Conclusion	48
4	Learning from bots and the social sciences	49
4.1	Introduction	49
4.2	Identity deception in bot research	50
4.2.1	Attributes and engineered features used to detect identity deception in bots	51
4.3	Identity deception in social sciences research	58
4.3.1	Motivation for considering the social sciences and psychology in particular	58
4.4	Identity deception in psychology research	59
4.4.1	Features used to detect identity deception in psychology	60
4.5	Proposed attributes and features to detect identity deception by humans on SMPs	62
4.6	Conclusion	64
II	Research design	66
5	Steps to assist in the automated detection of identity deception on SMPs	67
5.1	Introduction	67
5.2	Requirements for the automated detection of identity deception by humans on SMPs	68
5.3	The research approach	68
5.4	Steps towards detecting identity deception by humans on SMPs	69
5.4.1	Preparing the data	70
5.4.2	Experimenting with the data	80

5.4.3	Developing a model for identity deception detection	86
5.5	High-level design steps	89
5.6	Conclusion	91
6	A prototype for assisting the automated detection of identity deception	92
6.1	Introduction	92
6.2	The objectives of the prototype	93
6.3	The components of the prototype	94
6.3.1	Prepare	95
6.3.2	Discover	98
6.3.3	Detect	99
6.4	Conclusion	101
7	The research environment	102
7.1	Introduction	102
7.2	The research environment	103
7.2.1	Hardware	105
7.2.2	Database	107
7.2.3	Network	109
7.2.4	Software	109
7.3	Findings from the technical research environment considerations	115
7.4	Conclusion	118
 III The implementation of a prototype to develop and validate a model for assisting with the automated detection of identity deception		 119
8	Prototype: Prepare	120
8.1	Introduction	120
8.2	Gather the data	122
8.2.1	Results from exploratory data analysis	124
8.2.2	Findings that emerged from exploratory data analysis	128
8.3	Clean the data	131
8.3.1	Disregard non-human accounts	131
8.3.2	Remove irrelevant attributes	132
8.4	Label the data	137
8.4.1	Generating deceptive accounts	138

8.5	Engineer features	144
8.6	Prepare the data for machine learning	145
8.6.1	Discretisation of the attributes	146
8.6.2	Centring and scaling of the attributes	146
8.6.3	Testing the correlation of the prepared corpus	146
8.7	The ‘prepare’ component as a state transition diagram	149
8.8	Conclusion	150
9	Prototype: Discover	152
9.1	Introduction	152
9.2	The experiments to detect identity deception on SMPs	154
9.3	Supervised machine learning	156
9.4	Results from the supervised machine learning experiments	159
9.4.1	Experiment 1 – Using attributes from social media platforms ‘as is’	159
9.4.2	Experiment 2 – Using bot detection rules	164
9.4.3	Experiment 3 – Using knowledge from psychology	170
9.5	Comparing the results of Experiments 1, 2 and 3	176
9.6	Conclusion	178
10	Prototype: Detect	179
10.1	Introduction	179
10.2	Problems with using supervised machine learning to detect identity deception	181
10.2.1	Use entropy to explain identity deception detection results	184
10.3	Building an interpretable identity deception score	185
10.3.1	The IDDMLM sub-component	186
10.3.2	The IDDSM sub-component	188
10.4	Illustrating the working of IDDM	189
10.5	Comparing IDDSM with other interpretation methods	193
10.6	Conclusion	194
IV	Conclusion	195
11	Conclusion	196
11.1	Introduction	196
11.2	Revisiting the problem statement	197
11.3	Main contributions	201

11.3.1 Advancing the state of the art	203
11.4 Future work	204
Bibliography	206
A Glossary of terms and definitions	237
B Acronyms	239
C Ethical Clearance	242
C.1 Application for ethical clearance	243
C.2 Ethical clearance approval	249
D The FSOC proposal	250
D.1 HPI Future SOC resource example request	251
D.2 Technical research progress report example	254
D.3 Research poster example	258
E Publications and contributions	259
E.1 Journal papers	259
E.2 Conference papers	260
E.3 HPI Future SOC technical research reports	263
F Disclosure of the machine learning results	265
F.1 Introduction	265
F.2 Results from Experiment 1	265
F.3 Results from Experiment 2	272
F.4 Results from Experiment 3	279

List of Figures

1.1	The convergence of cyber security, big data, and humans	5
1.2	Graphic depiction of the layout of this thesis	12
2.1	Twitter account attributes [310]	22
2.2	Facebook account attributes	22
2.3	The convergence of cyber security, big data, and humans	26
2.4	Total papers published per year, per topic	30
2.5	Papers published about cyber threats on SMPs	31
2.6	Google trend search results since 2013 for cyber threats on SMPs	31
2.7	Keywords pertaining to social media threats	32
3.1	Deceptive role players on SMP and the focus of the current research	40
5.1	An explanation of variance vs bias [99]	73
5.2	Example of results returned from Google Face API [135]	76
5.3	Example of results returned from OpenStreetMaps [234]	77
5.4	Example of the results from the external names database [221]	78
5.5	Explaining the difference between ROC and PR	87
5.6	High-level design steps towards identity deception detection by humans on SMPs	90
6.1	The prototype – UML component diagram	96
6.2	Correlation between the requirements, steps and a prototype proposing to assist in the detection of human identity deception on SMPs	97
6.3	UML sequence diagram – ‘Prepare’ component	97
6.4	UML sequence diagram – ‘Discover’ component	98
6.5	UML sequence diagram - Machine Learning within the ‘discover’ component	99
6.6	UML sequence diagram – ‘Detect’ component	100
7.1	A proposed research environment infrastructure overview	104

7.2	HPI Future SOC lab storage used in total	107
7.3	A proposed future research environment infrastructure	117
8.1	High-level overview of the prototype: ‘prepare’ component	121
8.2	The ‘Prepare’ component	122
8.3	An example of gathered Twitter data	123
8.4	Tweets gathered from Twitter	124
8.5	Exploration of Twitter content data	127
8.6	Exploration of Twitter user data	130
8.7	Data cleaning: removing entries vs attributes	132
8.8	Corpus before and after data cleaning	133
8.9	The correlation between attributes found on Twitter describing a user’s account or identity	136
8.10	The correlation between attributes after data cleaning	137
8.11	Example results for the ‘Generate data’ API	139
8.12	Example results for the ‘Random User generator’ API	140
8.13	Comparing the appended accounts to the original Twitter user account corpus	143
8.14	The correlation between attributes after data labelling	144
8.15	The correlation between attributes after feature engineering	145
8.16	Results from binning the location	147
8.17	Results from binning the FRIENDS_COUNT attribute	148
8.18	The correlation between attributes after machine learning preparation . .	149
8.19	Data preparation as a state transition diagram	150
9.1	High-level overview of the prototype: ‘Discover’ component	153
9.2	The ‘discover’ component	154
9.3	The intention with iterative experimentation	156
9.4	Experiment 1 - Distribution of input data	160
9.5	Experiment 1 - Combined AUC results	163
9.6	Experiment 1 - Various-sized dataset results	164
9.7	Experiment 2 - Distribution of input data	166
9.8	Experiment 2 - Combined AUC results	169
9.9	Experiment 2 - Various-sized dataset results	169
9.10	Experiment 3 - Distribution of input data	171
9.11	Experiment 3 - Combined AUC results	174
9.12	Experiment 3 - Various-sized dataset results	175

9.13 Entropy results across all three experiments	177
10.1 High-level overview of the prototype: ‘Detect’ component	180
10.2 The ‘Detect’ component	181
10.3 Tree representation of the rpart model that determines identity deception	183
10.4 Explanation of the rpart tree	184
10.5 Detailed UML component diagram of the ‘detect’ component	186
10.6 Tweets for individual users (U)	191
10.7 The IDDSM sub-component	192

List of Algorithms

1	Training a machine learning model	84
2	The IDDMLM	187
2	The IDDMLM (continued)	188
3	The IDDSM	189

List of Tables

1.1	Research questions vs Research methodology	9
2.1	Examples of attributes found for top social media platforms in 2018	24
2.2	Cyber threats on social media platforms	28
3.1	SMP attributes vs Identity attributes	37
3.2	Using multiple attributes to engineer features for deception detection . . .	41
3.3	SMP attributes vs Cost classes defined by Cresci et al. [80]	42
3.4	Supervised machine learning algorithms used to detect bot and spam accounts	45
4.1	Attributes and features used in related work to detect bot accounts on SMPs	55
4.2	Detecting deceptive humans, given attributes and features from bot research	57
4.3	Identity features humans generally lie about – according to psychology . .	61
4.4	Attributes and features to detect identity deception by humans on SMPs	63
5.1	Additional features engineered and added to the corpus	79
5.2	Supervised machine learning algorithms used in this research	81
5.3	Hyperparameters per supervised machine learning algorithm	83
5.4	Theoretical depiction of a confusion matrix	85
5.5	Theoretical depiction of a confusion matrix with additional metrics	86
5.6	Identity deception detection model requirements catered for by the steps	90
7.1	Comparing the CPU core performance	108
7.2	Comparing operating system performance	110
7.3	Comparing hyperparameter performance	113
7.4	Comparing resampling fold performance	113
7.5	Comparing resampling repeat performance	113
8.1	Missing and unique values in the attributes of Twitter accounts	135

8.2	Origin of appended deceptive dataset attributes	141
8.3	Example of a generated deceptive account	141
8.4	Testing the validity of the introduced deceptive SMP accounts	142
8.5	The mean and standard deviation of all attributes	147
9.1	Machine learning algorithms used to develop a model for each experiment	157
9.2	Machine learning algorithm hyperparameters used across all experiments	158
9.3	Experiment 1 - Machine learning results	161
9.4	Experiment 1 - F1 scores over 30 repeats	162
9.5	Experiment 1 - F1 scores for various-sized dataset results	162
9.6	Experiment 1 - Entropy results	165
9.7	Experiment 2 - Machine learning results	167
9.8	Experiment 2 - F1 scores over 30 repeats	168
9.9	Experiment 2 - F1 scores for various-sized dataset results	168
9.10	Experiment 2 - Entropy results	169
9.11	Experiment 3 - Machine learning results	172
9.12	Experiment 3 - F1 scores over 30 repeats	173
9.13	Experiment 3 - F1 scores for various-sized dataset results	173
9.14	Experiment 3 - Entropy results	175
9.15	Overview of the best research per experiment	177
10.1	Results obtained from the IDDMLM model	190
10.2	Results obtained from the IDDSM model	192

Part I

Introduction, Significance of and Background to the Study

Chapter 1

Introduction

“The beautiful thing about learning is that nobody can take it away from you.” — B.B. King

1.1 Introduction

With Social Media Platforms (SMPs), anyone can contribute content towards the Internet whenever they want. SMPs make it very easy and intuitive for any layman to upload content or write about almost anything. By January 2018, 42% of the world’s population was actively adding data via SMPs to the Internet on a regular basis [180]. The mere fact that more people are using the Internet exposes more people to the wide range of cyber threats found on the Internet. SMPs are an important delivery vehicle of these cyber threats, as are emails [161].

Some of the many current cyber threats include the interception of data [352]; identity theft [175]; cyber grooming [92]; online bullying [7]; online paedophilia [48]; fake news [140], and cyber terrorism [113]. Take for example the case of the false images and rumours that were reported via Twitter, one of the many available SMPs, during hurricane Sandy [140]. The effect was the spread of unmerited panic, thus aggravating an already dangerous situation. It is also rumoured that fake news influenced the position that certain key role players took during the 2016 US presidential election [11]. In 2017, a 23-year-old British woman was jailed for grooming a 13-year-old boy via Facebook and later physically abusing the boy [27]. These examples show that cyber threats delivered through SMPs target humans in particular, whereas in the past, cyber threats were more likely to be directed at hardware devices and

infrastructure [65]. Those past attacks required great skill from malicious individuals, whereas SMPs exploit the vulnerabilities of the typical user; for example, it is nowadays possible to bully another individual anonymously and at a very low risk to the attacker [242]. These examples show that cyber attacks can target both large groups of people (e.g. the fake news spread about hurricane Sandy [140]) and individuals (e.g. the boy who was groomed via Facebook [27]). Most of the threats mentioned above have had deception in common.

It is very difficult to know what and whom to trust on SMPs. Lies on SMPs can be found in either the content or how the account holders present themselves to others [207]. A lie is when some fact is presented in a way which is not true [236]. On SMPs, for example, a person can lie about their age. When the lie, or the age as per the example, is accepted by someone else as the truth, it is known as deception [236]. Deceit is thus the consequence of a lie.

Much can be said about why humans lie. A short literature study of different social sciences (anthropology, sociology, psychology, and the like) [311] shows that humans lie about their image [284], name [311], age [103], gender [147], location [59], ethnicity [147], occupation [174], and many other things. Wang et al. [324] identify the details about which humans are bound to lie most by analysing the behaviour of past criminals. Wang et al. [324] found that humans potentially lie about certain details more than about others. Humans are also known to lie differently in different scenarios so as to be more effective in reaching their end goal [59]. Knowledge about the things humans lie about and how much they lie could help in identifying those details (also known as attributes) of the SMP account that could prevent/counter deception.

On SMPs, deception is unfortunately not restricted to humans alone. The accounts themselves could be non-human. Such accounts, commonly referred to as bot accounts, can for example be created at large, but be controlled by a single human. These bot accounts are created with various malicious intentions, for example aiming to elevate the position of a presidential candidate. This was noticed in the 2016 US election where the Russians were accused of controlling the messages sent from bot accounts on SMPs to elevate Donald Trump's public image [41].

To date, much research has been done to detect bot accounts on SMPs. For example, Yang et al. [344] show that the friend-to-follower ratio found in the meta data of an account holder on Twitter could be indicative of bots. Bots typically have very few friends but many followers. Li et al. [204] use clustering methods combined with behavioural patterns over time to detect bot accounts on YouTube. On a SMP like

YouTube, people are rewarded for the content they contribute in proportion to the amount of attention the content receives. By detecting behaviour (amongst clusters of accounts) that differs from normal expected human behaviour, these deceptive accounts can be identified, investigated, and removed by the SMP. Furthermore, Dickerson et al. [340], who propose natural language-processing techniques like sentiment analysis as an approach for detecting bot accounts on Twitter, showed success in each of their presented use cases. We also learn from their research that the analysis of content is very resource intensive and time consuming. According to Cresci et al. [80], using the attributes found to describe the SMP account, yields a degree of deception detection success similar to that achieved by methods that include the content as well. Cresci et al. [80] also showed how attributes from past research can be combined to engineer new features towards the detection of bots on SMPs.

Protection from deceptive humans themselves has, however, not been addressed much in research. This could largely be due to the difficulty of getting the volumes of sample data [350] required in order to really understand deceptive humans. When humans are targeted individually by another human, the attack is usually followed by detrimental consequences for the targeted individual, for instance fraud [340], or in extreme scenarios, death [89]. Past research towards deceptive bot detection also showed that using content as a mechanism of identifying deception is very resource intensive. For the purposes of this research, the researcher therefore focused only on the attributes describing the SMP account itself. These attributes describe the identity of the human. Lying or deceiving about any of these attributes is known as identity deception and implies that someone lies about who they are for some malicious purpose [306]. On SMPs, the cyber threat of identity deception is complex in that no face-to-face contact is required. In face-to-face communications, it is possible to visually validate whether the facts presented by a person about their identity, for example age and hair colour, are truthful or not [148].

Over the years, researchers have attempted continually to detect identity deception on SMPs as executed by humans. Current state-of-the-art research, like that undertaken by Alowibdi et al. [12], proposes a method to detect identity deception by only looking at a human's SMP account profile colour and name. They argue that certain genders prefer certain colours above others and this knowledge can aid in the detection of identity deception. Tsikerdekis et al. [305] claim that non-verbal attributes, like the time it takes to post content on Wikipedia, could aid in detecting identity impersonation. Ferrara et al. [113] use a known list of extremists to create a supervised machine learning model with which to detect similar-looking accounts. They use account attributes such as the

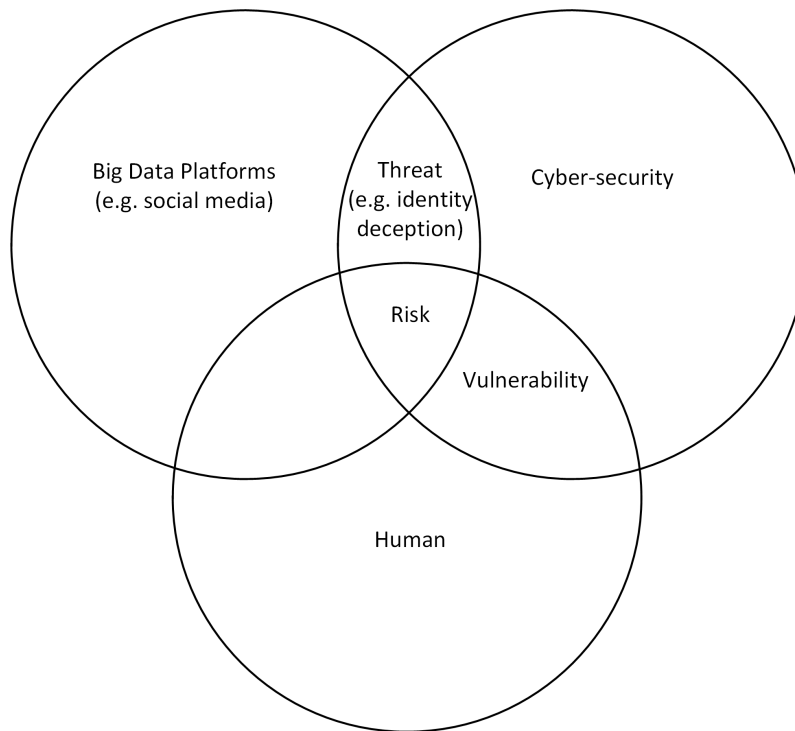


Figure 1.1: The convergence of cyber security, big data, and humans

number of friends, number of followers, and the length of the name in their machine learning model. Although their research yielded promising results, the attributes used were handpicked for the cases at hand.

Cyber security proposes to protect humans against various cyber threats, of which identity deception is an example. SMPs are an example of a big data platform [139] that exposes humans to cyber threats such as identity deception. Figure 1.1 shows the high-level components of cyber security and their relationships to big data platforms (e.g. social media) and humans. Cyber security proposes to lower the risk of threats (e.g. identity deception) by protecting vulnerable humans on SMPs.

The question remains whether all attributes on SMPs can be used towards the automated detection of identity deception by humans against humans. Which features contribute more than others? Can we learn from past research in the fields of social sciences and bots to engineer new features that may assist in the more accurate automated detection of human identity deception? Where do we get enough sample data, or do we even need sample data to assist in the automated detection of human identity deception on SMPs?

The present research has tried to answer these questions by first defining an appropriate environment for SMP research that uses a methodology and prototype that will assist

with the automated detection of human identity deception on SMPs. This environment proposes to enable various machine learning experiments. A critical evaluation of the SMP attributes and engineered features showed which of them contributed most towards identity deception as executed by humans on other humans. This knowledge should contribute towards the fields of both cyber security and the social sciences, not only to understand human identity deception better, but also to protect individuals from various cyber threats found on SMPs (which are maybe not as common as cyber threats from bots).

1.2 Research problem

Against the background of the previous discussion, the primary research problem is summarised as “assisting in the automated detection of identity deception by humans on big data platforms, and in particular on SMPs”. Secondary to that, various experiments were used not only to empirically evaluate the available SMP attributes in order to detect such identity deception, but also to engineer new features borrowed from bot detection and psychology to enrich the results. The results help to explain whether related work can assist in the automated detection of human identity deception on SMPs, and were also meant to establish which SMP attributes were viable and valuable in this process.

1.3 Research questions

The following research questions had to be answered to solve the stated research problem:

- Research Question 1: What are the cyber threats found on SMPs and why is it important to find a solution to the problem of identity deception by humans on SMPs as opposed to bots? Finding identity deception by bots on SMPs is out of scope for this research.
- Research Question 2: What attributes are available on SMPs that have the potential to be used for identity deception by humans?
- Research Question 3: What are the requirements for a model that will assist in the automated detection of human identity deception on SMPs and how can such a model be implemented?

- Research Question 4: Can features from related research in the detection of non-human or bot accounts and knowledge about deception from the social sciences contribute towards the detection of identity deception?
- Research Question 5: Can we explain the model results in a format that is interpretable without any prior knowledge of machine learning, to show which attributes and features were most valuable in the detection of human identity deception?

1.4 Research scope

This research reported on in this study was limited to identity deception by humans on SMPs. The detection of deceptive non-human or bot accounts did not fall within this scope, but research on the detection of bot accounts was consulted for any potential synergies or lessons learned. This research was also limited to Twitter, although the results were applicable to other SMPs as well. Furthermore the Twitter corpus was limited to minors only, although the results are applicable to other demographics as well. Deception over time (a human can lie differently from one day to the next) and deception between SMPs (a human can lie differently from one SMP to the next) were not considered for the purposes of this research. Lastly, besides proposing a prototype for automated assistance in the detection of identity deception by humans, further work towards integration with an existing SMP was not addressed in the current research. Comments were nevertheless made on how to address the current vulnerabilities on SMPs.

1.5 Methodology used

The following steps were taken to solve the primary research problem as well as the subsequent research questions stated in Section 1.2 and section 1.3.

The *first step* was to conduct a literature study on SMPs, big data, cyber security, deception, and identity deception. A discussion on the explosive increase in data led to defining what constitutes big data, why SMPs are seen as examples of big data, and the cyber threats faced in dealing with SMPs. Deception and identity deception were found to be examples of such cyber threats. A detailed literature review was presented with a critical evaluation of existing state-of-the-art research work in the field of SMPs, cyber

security, deception, and identity deception. The knowledge gained was used to create a list of requirements for a model to assist in the automated detection of identity deception by humans on SMPs. Finally, a literature review of current research in the field of social sciences – more specifically psychology – as well as on bot detection was presented so that the review would lead to a better understanding about why humans lie.

The *second step* was to present a research environment for implementing the requirements and adopt a scientific research approach towards the automated assistance of identity deception detection by humans on SMPs (as defined in the first step). A prototype with the necessary components to implement these model requirements was presented and it allowed for various experiments to be conducted. Finally, the researcher defined the research environment required for the proposed prototype to enable SMP research.

The *third step* involved each of the following components of the prototype in detail: preparing the data; discovering an identity deception model; detecting human identity deception on SMPs. The data was gathered from an SMP and prepared to enable various experiments. Several experiments were conducted to discover the best attributes and features for the automated assistance of identity deception detection as performed by humans on SMPs. The experimental results culminated in a ‘deception’ score, also referred to as the Identity Deception Detection Model (IDDM) per person. A further explanation followed as to why any one person was potentially perceived as being more deceptive than another. This was indicative of the attributes and features about which humans deceive others most on SMPs.

A final and *fourth step* was to summarise the findings and come to a conclusion on the success of the research.

Table 1.1 shows how these research steps map to the research questions.

Table 1.1: Research questions vs Research methodology

Research question	Research methodology step
Research Question 1: What are the cyber threats found on SMPs and why is it important to find a solution to the problem of identity deception by humans on SMPs as opposed to bots?	First step
Research Question 2: What attributes are available on SMPs that have the potential to be used for identity deception by humans?	First step
Research Question 3: What are the requirements for a model that will assist in the automated detection of human identity deception on SMPs and how can such a model be implemented?	First and second step
Research Question 4: Can features from related research in the detection of non-human or bot accounts and knowledge about deception from the social sciences contribute towards the detection of identity deception?	Third step
Research Question 5: Can we explain the model results in a format that is interpretable without any prior knowledge of machine learning, to show which attributes and features were most valuable in the detection of human identity deception?	Fourth step

1.6 Layout of thesis

This thesis consists of 11 chapters as depicted in Figure 1.2.

Chapter 1 introduces the thesis and indicates how the research was structured.

Chapter 2 introduces SMPs as big data platforms that display all the characteristics ascribed to such platforms. A detailed overview is given of all attributes available in current SMPs that describe an account holder, as well as their content and relationships with other account holders. A literature review reveals recent research performed on SMPs and looks in depth at research related to cyber security. The various cyber threats found on SMPs are identified. An additional high-level literature review, using the knowledge gained from over 60 000 published papers, shows how humans are left vulnerable on SMPs in the case of cyber-security threats like identity deception.

Chapter 3 discusses terms like ‘deception’ and ‘identity’ to understand their role in ‘identity deception’. Various aspects are examined that should be considered when one deals with identity deception on SMPs, such as the origin of identity deception, which attributes found on SMPs are prevalent in identity deception, and how to build a model to assist in the automated detection of human identity deception. The chapter concludes by presenting a list of requirements for a model that assists in the automated detection of identity deception by humans on SMPs.

Chapter 4 describes the features that distinguish one human from another. To enrich

the understanding of identity deception further, the research fields of social sciences and bots are presented as alternative sources of knowledge. So far, the research field of bots has proposed various attributes and engineered features to detect bots on SMPs. Equally, the field of psychology has done much regarding the topic of deception. A literature review of past research work on deception in the field of psychology reveals those features that humans are most prone to lie about. This knowledge gained from psychology and bots is applied to SMPs to detect human identity deception with a greater degree of success. A more targeted approach that uses only those attributes/features that we know humans lie about, should yield better results overall.

Chapter 5 defines the steps that are required of a model to assist in the automated detection of identity deception by humans on SMPs. A detailed discussion of each step is followed by a high-level design of the components required of a prototype to not only implement these steps, but also cater for the requirements expected of a model that can assist in the automated detection of human identity deception on SMPs.

Chapter 6 discusses the three components of the prototype. The first component prepares the data for the model. The second component uses the prepared data to discover a model by experimenting with various supervised machine learning algorithms and combinations of input data. The last component uses the most accurate model in the form of an IDDM to automatically detect human identity deception on SMPs. Unified Modeling Language (UML) will be used to discuss the relationships between the components and the flow of messages and data within each component in detail.

Chapter 7 defines what a research environment, in terms of hardware and software, should look like for the proposed prototype. The research environment should cater for the expected time required to detect identity deception in a big data environment, for the research work to be completed in a timely manner and, lastly, for a model that assists in the automated detection of human identity deception on SMPs.

Chapter 8 presents the first component of the prototype. The prepare component deals with data gathering, cleaning, labelling, feature engineering, and preparation for supervised machine learning purposes. The chapter also presents the results from this component and its sub-components. Finally, a state transition diagram shows how data evolves from one state to another as we implement the requirements expected of a model that assists in the automated detection of human identity deception detection on SMPs.

Chapter 9 presents the second component of the prototype. The discover component

uses machine learning algorithms to develop machine learning models aimed at identifying human identity deception on SMPs. This is achieved through various experiments. Each experiment proposes to improve on previous experimental results. Not only are the attributes found in SMPs used, but engineered features from related research towards the detection of bots and psychology are also added iteratively to achieve the proposed outcome.

Chapter 10 presents the third component of the prototype. The detect component explains why a human is perceived as being deceptive. This component introduces an IDDM that not only scores an SMP user's perceived deceptiveness, but also interprets the scores by using an interpretation method that includes the use of the Shannon Entropy equation [273].

Chapter 11 concludes the thesis with suggestions for future research work.

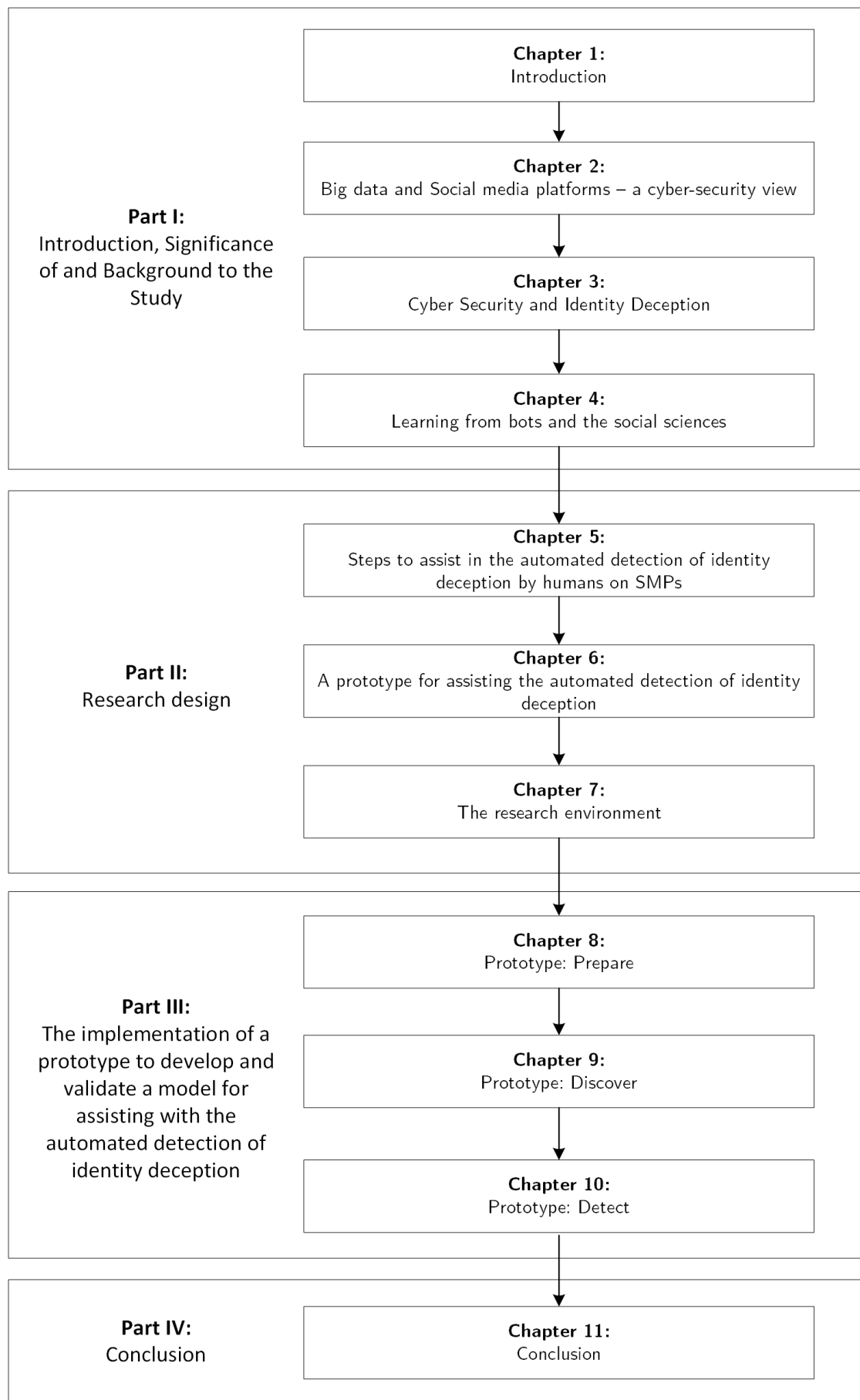


Figure 1.2: Graphic depiction of the layout of this thesis

Chapter 2

Big data and Social media platforms – a cyber-security view

“The goal is to turn data into information, and information into insight.” –
Carly Fiorina

2.1 Introduction

Data has exploded in volume over the past number of years. In addition, data is constantly being created and processed at higher velocities and in a variety of formats, such as videos and images. These three – volume, velocity and variety – are also known as the characteristics, or ‘3Vs’, of big data [195].

This chapter firstly describes big data by means of examples, followed by a further explanation of each of the 3Vs. This description of big data will serve as an introduction to Social Media Platforms (SMPs) (as an example of big data) by showing that the 3Vs are just as prevalent in SMPs. SMPs consist of data gathered from millions of users. A detailed analysis of the constituents or attributes that describe a single user – over the various SMPs deemed most used at the time of writing – will further illustrate the nature of data available on SMPs. This analysis will show whether the research performed on the attributes from one SMP can be relevant to another research study on a different SMP, and whether SMPs present new cyber-security challenges to humans. It will also contribute to the search for a solution that will assist in the automated detection of human identity deception on SMPs, as was discussed under the research problem in the previous chapter.

To critically evaluate the cyber-security challenges in more detail, the cyber threats found on SMPs are identified and an in-depth review is conducted of academic research available on cyber threats over the past four years. This will show how humans are left vulnerable on SMPs to cyber-security threats like identity deception.

2.2 Big Data

Big data deals with data and therefore it is first necessary to understand how data has evolved. Throughout history, data was gathered and preserved for the next generation. The earliest surviving papyrus scroll dates to 2400BC [212] and was used by the Egyptians to capture data. Since then, libraries have been filled with books and data has been captured in digital format. More recently, this data has been made available to people via the Internet, and the last decade has seen an unprecedented increase in the amount of data. Data is nowadays produced in real time and sometimes without human intervention, through sensors, digital cameras, machines, phones and many other devices. This data explosion was compounded by the Internet of Things (IOT) and SMPs, among others.

The IOT constitutes the idea that devices can be connected to the Internet via a unique identifier that both produces and consumes data, also referred to as ‘smart data’ [169]. Devices add to the ever-growing data explosion. According to Gartner, there will be nearly 25 billion devices on the IOT by 2020 [226], while ABI Research states that more than 30 billion devices will already be connected wirelessly to the IOT by 2020 [2]. SMPs, on the other hand, enable the ‘man on the street’ to create their own content, share their lives in words and videos, and blog about their thoughts without requiring any technical skills. As of mid-year 2017, Facebook had over 2 billion users, YouTube had 1,5 billion users, and Twitter had 328 million users [74]. For Facebook, this equates to humans generating 7 Petabytes of data each day [173] made up of, among other things, 2 billion images per month [270].

It is no coincidence that this increase in the amount of data correlates very closely with the evolution of the web and more specifically with social media (Web 2.0) and the IOT (Web 3.0) [250]. Web 2.0 enabled two-way communication on the Internet and thus allowed social media sites like Twitter and Facebook to be created during the 2004-2006 period [315]. Web 3.0, or better known as the semantic web, tries to make the web readable for the IOT [4]. Various research efforts are ongoing on the semantic web [123] [44] [15] [346] to allow for better communication between devices.

With not only the data volumes exploding, but the contributors to data also increasing rapidly, it is no surprise that data has evolved in various shapes and forms. The term ‘big data’ has emerged, among other things, due to this explosive growth. The term ‘big data’ was first coined in the 1990s but academically made its appearance in 1998 [96] to describe working with large volumes of data. According to Wikipedia [329], ‘big data’ defines data that is difficult to process with traditional data-processing applications. Gartner [125], on the other hand, describes ‘big data’ not in terms of its complexity, but rather according to its underlying characteristics, and notes that data is there to enhance insight and decision making.

The fact that we have too much data to consume is not new. The term ‘information overload’ was popularised in 1970 in a book by Alvin Toffler, called *Future Shock* [300], in which he describes a bleak future where information will grow and impede decision making. Even though we do generate much more information, technology has however managed to keep up by means of faster processing, research into better algorithms, and improved storage. The term ‘information overload’ has therefore evolved from ‘information glut’ and ‘data smog’ in the 1990s [276], and nowadays we refer to it as ‘big data’.

2.2.1 Big data case studies

To develop a better understanding of the characteristics of big data, the following case studies of its application were considered:

- In 2017, Coca Cola had 105 million Facebook friends and 35 million Twitter followers [215]. Coca Cola does not only analyse the messages posted by these people to deliver more targeted marketing, but also knows which products people refer to by using image recognition techniques. The company can handle large data volumes and is able to deal with data presented in different varieties or formats.
- Domino’s Pizza allows their customers to order food via a variety of channels, like Twitter, Pebble, and Amazon Echo [214]. More than 55% of orders are now generated online, as opposed to telephonic orders. The data generated from these and traditional channels comes from 85 000 different data sources [214]. The company does not only have to handle great volumes of data in a variety of formats, but its survival is dependent on customers receiving their orders on time. Dealing with huge amounts of data at high velocity is thus very important to

them.

- The whole identity of a person has become important for political campaigns [62]. For example, people's health concerns and even their car ownership could potentially indicate which political party the person would vote for. During the 2016 American elections, a company called Cambridge Analytica analysed more than 5 000 data points collected per person [295]. There is currently still an ongoing investigation into whether this data was legally obtained [179]. Cambridge Analytica overlaid consumer data on political data to create campaign messages targeted at all potential voters. They were also able to analyse responses from potential voters on social media in real time as the political campaign progressed. Although it is questionable whether Cambridge Analytica used legally obtained Facebook data, they dealt with high volumes and a vast variety of data, all leading to the success of their client, Donald Trump [295].

Each of the above case studies shows how each company improved its business in general while dealing with the volume, velocity, and variety of data. It is therefore imperative to explain each of these characteristics in more detail as each can be a determinant for a framework that describes big data.

2.2.2 Big data characteristics

In 2001, Doug Laney, an analyst from the Meta Group (now Gartner), produced a paper proposing a formal approach to data management [195]. Laney illustrated that dealing with data has challenges and necessitates a formal framework. He defined data management along three dimensions that have subsequently been referred to as the 3Vs, or characteristics, of big data, and that relate back to the case studies previously mentioned [88]:

- Volume – Data volumes are growing at a phenomenal rate and will continue to do so in the future.
- Velocity – Not only is more data being generated, but it is done at a faster pace. The demand to consume and make use of this data in a timely manner has become important.
- Variety – Data originates from many sources, all carrying data of value in various formats like text, video and photo.

Since the publication of Dough Laney's paper, additional characteristics have been added

from various places, all playing on words starting with a ‘V’. The list of additional characteristics includes Veracity, Volatility, Validity, Viability, Value, Variability, and Visualisation [255] [30] [181]. The researcher will in this study focus on the three original characteristics of big data, since they remain widely recognised as the core framework used to classify big data [321]. The three characteristics of volume, velocity, and variety will now be described in more detail.

2.2.2.1 Volume

As mentioned earlier, data volumes are growing at a phenomenal rate and will continue to do so in the future. The more data there is, the more challenging the algorithms required to analyse the data become and the more time it takes to maintain and ensure the quality of the data. Here are some noteworthy facts about data volumes today:

- In 2016, Netflix had roughly 3 500 instances hosting 1.3 Petabytes of data [46].
- In 2016, Airbnb had 11 Petabytes of data spread across two separate clusters [6].
- In 2017, the CERN datacentre, which monitors collisions of particles in the Large Hydron Collider (LHC), generated a Petabyte of collision data per second and stored over 200 Petabytes of that data for further analysis [121].

Traditionally, large volumes of data were dealt with by increasing the capacity of the infrastructure, for example by adding more hard disks to the same server [66]. However, a point is reached where adding additional storage or processing power to a single server does not provide a solution, and the volume of data is still too much. Some big data platforms have addressed the problem of dealing with mass volumes of data in the following ways:

- Frameworks like Apache Hadoop [19], which allow for the storage of large volumes of data across multiple servers through a distributed file system called HDFS.
- Cloud-based services like Amazon Web Services (AWS) [14] and Windows Azure [222], which allow data to be stored off-site at different pricing models. This allows the focus to be on the data and provides relief from the additional burden of handling server maintenance, archiving, additional storage, and scaling.
- Distributed non-relational database management systems like Apache Cassandra [194] and HBase [19], which allow data to be stored across multiple servers.

Many of the above proposed solutions are found in SMPs. Companies like Airbnb, for example, host some of their data off-site in the cloud on AWS [327]. Airbnb was able to reduce costs by 70% through the use of low-cost cloud storage solutions. Facebook, on the other hand, uses a distributed data storage approach to manage their volumes of data [327]. The different approaches show that every storage solution needs to be considered on a case-by-case base. For example, Facebook finds it much more cost effective to run their own dedicated data centres rather than to use cloud storage solutions [229]. Their volumes of data are just too large for cloud storage solutions to make sense. In their own data centres, they run a distributed architecture – one that allows for the storage of data across many low-cost servers.

2.2.2.2 Velocity

Not only is more data nowadays being generated, but it is done at a faster pace. The demand to consume and make use of this data in a timely manner has increased. Here are some noteworthy facts on data velocity today:

- In 2016, Netflix handled 8 million events, equating to roughly 24GB of data per second during peak times [46].
- In 2017, on average around 6 000 tweets per second were generated on Twitter [159].
- Instagram generated on average 4.2 billion likes a day during 2017 [299], equating to almost 50 000 likes per second.

Traditionally, data velocity was handled by adding more processors to servers, adding more RAM, or running processes in parallel. However, at some point a threshold is reached where the analysis cannot keep up with the amount of data being received, regardless of the latency. The problem of velocity has been addressed by some big data platforms in the following ways:

- With low latency, the data can be split across various servers. One common model is known as MapReduce that originated at Google [91]. With MapReduce, the initial analytical problem is mapped across multiple servers. After the calculations have been performed, another step collects all the answers from the various servers and reduces it to a single answer. This allows for faster parallel processing of batch data.
- With high latency, near real-time query tools are used like Cloudera Impala [20], Apache Dremel [272] or Spark Shark [272]. These tools achieve faster analytics

by using a form of column-based storage and scalable aggregation algorithms for computing query results in parallel.

- With high latency in-stream analysis, tools like Storm and S4 are coming to the fore [272]. These tools will analyse the data as it arrives and produce results as they become available. In contrast to low-latency batch processing, these tools will continue to analyse until instructed to stop.

Many of the above solutions are found in SMPs. Instagram [138], for example, makes use of the Lambda software framework [184], which caters for the requirement of doing both batch and real-time processing [138]. The architecture uses the following three layers to cater to the above three ways of handling velocity on big data platforms: batch, serving (real-time), and speed (in-stream) [85]. Facebook, on the other hand, has addressed the problem of velocity by creating their own near real-time query language called Presto [302].

2.2.2.3 Variety

In recent years, additional variations of semi-structured and unstructured data, also known as heterogenous data, have been seen that do not fit into a relational methodology. The challenge for big data with regard to the management and processing of heterogenous data seems to be the amount of manipulation and effort required to get data in a format that is usable, presentable and of value [279] [202]. Here are some noteworthy facts on data variety today:

- The Facebook ‘Like’ button does not have any information in itself – it is a button on a page. However, the action of the user selecting this option has a meaning that can vary in different contexts. The information can, for example, be used to know whether people like a certain product [98].
- According to an online blog, at the beginning of 2018 around 95 million images were uploaded to Instagram every day [29].
- It is predicted that by 2019, 80% of global internet traffic will consist of video content [143].

Traditionally, there were only a few options to handle heterogenous data. Objects such as images were either stored as large binary objects, or only the link of the image to an external location was stored. Some big data platforms have addressed the problem of data variety in the following ways:

- Database management systems like Postgres [249] give organisations the best of both the structured and unstructured world. Postgres has added support for the JavaScript Object Notation (JSON) data type. This allows for unstructured data (in the JSON format) to be stored in a relational model and queried via normal Structured Query Language (SQL) [308]. This is relevant to textual type data only.
- With heterogenous data, the data is in some cases kept in the form in which it was received. Database management systems like Hadoop Hive [19] allow users to query the data in its raw format, thus saving time by not first converting the data to a structured relational format.
- Heterogenous data is stored in a Content Delivery Network (CDN) that is aware of the region and where the data belongs physically [341]. This means that the CDN will only hold content relevant to that region. Following this approach allows users in that region to obtain data more quickly. Video providers like Netflix rely heavily on this technique [341].

Many of the above proposed solutions are found in SMPs. Companies like Netflix make use of CDNs extensively to ensure video content is delivered in a timely manner to the users in a particular region [341]. The meta data about the videos' viewing history is stored in Cassandra [47], while the content itself is kept on AWS [327].

2.3 Social media platforms

The previous section used examples from SMPs to show how the characteristics of volume, velocity, and variety are identified by big data. SMPs are therefore an example of big data. To find a solution to the research problem – i.e. assisting in the automated detection of human identity deception (as stated in Chapter 1) – the present research study will focus on SMPs going forward. This is because data for SMPs is readily available for research purposes [228] and can help answer questions pertaining to cyber threats when dealing with big data [217]. Many different types of SMPs exist [139], each characterised by user-generated content that allows users to connect to each other. The following are some examples of types of SMPs [217]: online social networking; blogging; wikis; media sharing; online reviews; news groups; microblogging; and geo-location services.

Online social networking sites like Facebook [109] allow users to *inter alia* share their

status, location, and who they are sharing an experience with (e.g. pictures and videos of events in their daily lives). Microblogging sites like Twitter [310] allow users to post, for example, short status updates that are instantly visible to other users. These posts can also include images or short videos, but the message text length is restricted. Media sharing sites like Instagram [165] allow users to upload their own pictures and share them with others. All these SMPs gather and store the content provided by their users.

2.3.1 Social media data

SMP data is mostly known to consist of the content added by each of the many users. These users are required to open an account with the SMP before they can start participating [109]. This data is generally referred to as meta data [281]. The meta data does not only identify the user, but also serves to distinguish them from another user. Having made a study of the meta data and content on SMPs, each piece of information can be seen as an attribute of that user. Take Twitter for example. On Twitter, the name and location of the user are examples of attributes describing the user. Figure 2.1 shows an example of some additional attributes available on Twitter that describe a single user.

Similar attributes are found on Facebook to describe a user. Figure 2.2 shows an example of some attributes available on Facebook. Both the Twitter and Facebook examples are presented in a language-independent format for exchanging information, namely JSON [82].

Although there are differences between Facebook and Twitter's attributes, there are also many similarities, such as the user's name and ID. Because of these differences and similarities, the researcher decided to use a mechanism that can assist in a structured way to develop a common understanding of attributes across multiple SMPs, namely a classical categorisation approach. A classical categorisation approach attempts to group objects based on their similarity [292] and also ensures that all attributes in one category are mutually exclusive of another [167]. This means that the category describes the intention of all attributes in that category and no attribute is more important than another within a category [167]. With this in mind, the researcher defined the following categories for attributes found on SMPs [314]:

- Attributes describing the account profile, for example the name of the account holder.

```

{
  "tweet": {
    "created_at": "Thu Apr 06 15:24:15 +0000 2017",
    "id_str": "850006245121695744",
    "text": "1\ Today we\u2019re sharing our vision for the future of the Twitter API platform!\nhttps://t.co/XweGn",
    "user": {
      "id": 2244994945,
      "name": "Twitter Dev",
      "screen_name": "TwitterDev",
      "location": "Internet",
      "url": "https://dev.twitter.com/",
      "description": "Your official source for Twitter Platform news, updates & events. Need technical help? Visit http",
    },
    "place": {
    },
  },
  "entities": {
    "hashtags": [
    ],
    "urls": [
      {
        "url": "https://t.co/XweGngmx1P",
        "unwound": {
          "url": "https://cards.twitter.com/cards/18ce53wgo4h/3xo1c",
          "title": "Building the Future of the Twitter API Platform"
        }
      }
    ],
    "user_mentions": [
    ]
  }
}

```

Figure 2.1: Twitter account attributes [310]

```

{
  "id": "12345678",
  "birthday": "1/1/1950",
  "first_name": "Chris",
  "gender": "male",
  "last_name": "Colm",
  "link": "http://www.facebook.com/12345678",
  "location": {
    "id": "110843418940484",
    "name": "Seattle, Washington"
  },
  "locale": "en_US",
  "name": "Chris Colm",
  "timezone": -8,
  "updated_time": "2010-01-01T16:40:43+0000",
  "verified": true
}

```

Figure 2.2: Facebook account attributes

- Attributes describing information about the account, for example the account creation date.
- Attributes describing the account's behaviour, for example the source a post was made from.
- Attributes describing the account's relationships, for example its followers.
- Attributes describing the content posted, for example tweets on Twitter.

Table 2.1 shows each of the above categories with the attributes available for some of the top SMPs at the time of writing this research in 2018 [64]. The list excludes any recognised messenger sites like Facebook messenger and Skype, since the researcher is interested in SMPs that show non-textual data as well. Even though YouTube is on the list of top SMPs [64], Google+ is included in Table 2.1 instead. YouTube allows its users to link their Google+ profile to their YouTube channel [348] and therefore the attributes describing the user's profile will be found on Google+ instead of on YouTube. It is interesting to note the large amount of overlap between the meta data of various SMPs in Table 2.1. Each SMP holds similar information about its users for each of the five attribute categories identified. This in turn means that research using data gathered from one SMP could potentially be applied to another. In the case of the research problem at hand, knowing the category of attributes where deception is most prevalent has many advantages. Not only could there be computational advantages in choosing the attributes of one category rather than another, but it is possible to potentially engineer new features from the existing attributes to describe the user in more detail – and thus, indirectly, to aid in deception detection. For example, by adding the age of the user, which was engineered from a profile image of that user, we know more about that user and could try to determine whether they lied about their age. Should the hypothesis that users lie about their age be proven on one SMP, the same could apply to another SMP, because they share similar information. If a method were then to be proposed to address the cyber threat of deception, it could be assumed that the proposed method will apply to SMPs in general.

Table 2.1: Examples of attributes found for top social media platforms in 2018

Content Type	Description	Facebook [109]	Google+ [134]	LinkedIn [206]	Twitter [310]	Pinterest [248]	Instagram [165]
Describing the user	The Identifier (ID) of the person's user account	id	id	id	id	id	id
	Their first name	first_name	name.givenName	first-name		first_name	
	Their last name	last_name	name.familyName	last-name		last_name	
	Middle name	middle_name	name.middleName				
	Full name		name.formatted		name		full_name
	Name to display	name	displayName	formatted-name	screen_name	username	username
	Age	age_range	ageRange				
	Birth date	birthday	birthday	date-of-birth			
	Profile picture	cover	image	picture-url	profile_image, original profile_image background_image profile_text_color profile_background_color	image	profile_picture
	Gender	gender	gender				
	Relationship status	relationship_status	relationshipStatus				
	Language	languages	language	languages	lang		
	Location	location	placesLived[].primary	location	location		
	Geo-location				geo_enabled, latitude, longitude		
Time zone	timezone			time_zone			
UTC offset				utc_offset			
Total number of posts	posts			statuses_count	counts	counts.media	
Total number of followers				followers_count	counts (per board)	counts.follows	
Total number of friends	friends.total_count		num-connections	friends_count		counts.follows_by	
Total number of public lists a user belongs to				listed_count			
Bio field	about	aboutMe	summary	description	bio	bio	
Describing their account	Authenticity of account	is_verified	verified		verified		is_business
	Updated time	updated_time			created_at	created_at	
Behaviour	List of devices	devices			source		
	Tagged by other users	tags			user_mentions		users_in_photo
	What content user liked	likes			in_reply_to		media.likes
Relationships	Friends	friends			friends		
	Groups	groups		following	followers		
	Listed		shared circles		lists		
	Family	family					
Content	Content specific fields	albums	urls[]	position, skills	tweets	boards	media
		feeds	organizations[]	job-bookmark	tweets.created_at	pins	media.created_time
		events	braggingRights	certifications			media.location
		photos	occupation	educations			
		videos	skills	courses			
				volunteer			
				publications			
		interests					
			honors-awards				

2.4 Cyber security within social media platforms

Many public examples can be found of cyber threats on SMPs. In February 2018, 13 Russians were charged by the United States Justice Department for subverting the 2016 political campaign. They created social media accounts as if they were American citizens with the assumed intention of creating discord in the democratic system through the content that they posted. In another example from 2017, women were groomed via Facebook, and then raped and killed [89] in South Africa. The victims were lured through a fake profile and given money to travel once trust was established. When they arrived at their destination, they met someone pretending to be a friend who would take them to the person they met on Facebook. What the victims did not know, was that the person whom they met was the person behind the fake profile and their attacker. The attacker thus presented a fake name and profile picture on Facebook.

In both cases, the attackers lied by changing some of their previously mentioned social media account attributes. Most of these cyber threats have fake IDs [307] or impersonation [92] as part of the threat. It has become very difficult to know who and what to trust online [352] [257]. Cheswick et al. [67] propose that cyber security can keep someone from doing something you do not want them to do on any electronic device. Cyber security is therefore the protection of users from threats like grooming, cyber stalking, phishing, or identity deception on SMPs. For the purpose of this research, identity deception will be referred to as a cyber threat, but it is also accepted that it could be seen as a risk when you have a poorly constructed account on a SMP and there are malicious people who want to impersonate you.

Whenever humans make use of SMPs, they expose themselves to a range of cyber threats. One of these threats could be another human, a malicious individual. Those people who use the SMPs have vulnerabilities that can be exploited and therefore they require the protection that cyber security can offer [67]. Humans are gullible and not always able to discern the truth from lies [265]. SMPs unfortunately allow humans, such as these malicious individuals, to deceive other humans [76]. The threat of the malicious individual, together with the vulnerable way in which humans use SMPs (e.g. the careless way in which they construct their account profiles), increases the risk of a threat materialising in the form of identity deception. The high-level components of cyber security and their relationship to big data platforms and humans are depicted in Figure 2.3. The figure shows how humans use big data platforms like social media. The figure also shows that these big data platforms pose a threat to whoever uses them,

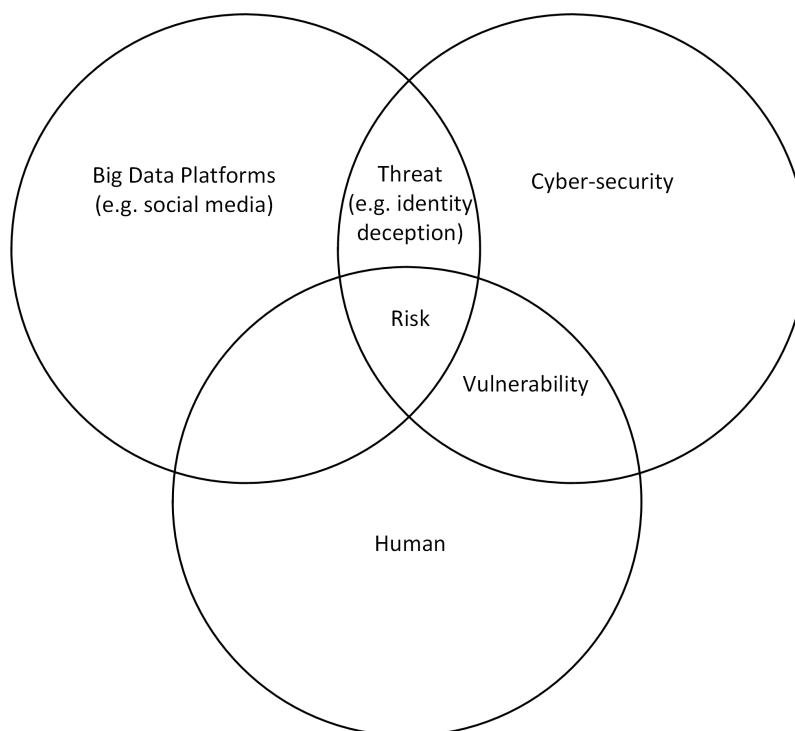


Figure 2.3: The convergence of cyber security, big data, and humans

leaving them vulnerable. When a threat targets a vulnerability, it creates the risk of the threat, like identity deception, being successful. Cyber security proposes to lower the risk of these threats by protecting vulnerable users on SMPs.

2.4.1 Related research on cyber threats found on SMPs

Having said the above, cyber threats on SMPs are generally far from understood and there is still much to explore going forward [246]. To understand what types of cyber threats, besides identity deception, can be found on SMPs, existing research published in Google Scholar [136] was consulted. Google Scholar [136] is an online search engine for scholarly research work from various disciplines. To find relevant cyber threat literature, a search was performed on Google Scholar to find research containing the words "taxonomy" and "cyber security" published since 2002. These taxonomies, which help in the systematic study of a field [162], apply to SMPs and define various cyber threats. Applegate and Stavrou [21] describe cyber security's impact on assets, operations, systems, and information. They do not focus on the threat itself in their taxonomy, but rather describe the asset being attacked. Howard [162] includes the attackers' motivation for the attack and notes that the attack can be motivated by status challenge, political gain, financial gain, or simply to damage another individual.

Cebula et al. [61] define a taxonomy for operational security risk that mentions the actions of people posing a risk. People can either do something they are unaware of, do something deliberate, or in some cases do nothing at all.

The problem with taxonomies, however, is that they are too general [150]. They describe, for instance, that there could be a threat but do not classify the threat itself. For this reason, the researcher performed a further literature review of research work that aims to classify cyber threats. The literature was obtained using "cyber", "threats", and "classification" as search words on Google Scholar [136]. The results of this investigation are summarised in Table 2.2. Hansman and Hunt [150] propose to extend cyber-security taxonomies with more dimensions to be able to classify the various threats found in cyber security. In their extended taxonomy they are able to classify cyber threats towards people and networks in more detail. Gharibi and Shaaibi [126] show that personal online attacks take many different forms. Perez [244] describes the top threats pertaining to Web 2.0. As mentioned earlier, Web 2.0 is known for the inclusion of SMPs. Williard [332] looks at threats specific to children on SMPs, while Fire et al. [114] divide cyber threats into two main categories: classical and modern. The modern threats are more specific to SMPs. Kirichenko et al. [185] list a range of threats on SMPs and focus on the approaches taken to protect against these threats. Patel et al. [240] identify threats on SMPs as either infringing on the security of the platform or on the privacy of the person using the platform. Trivedi et al. [303] show threats to be either related to people or to networks. Pradhan et al. [251] investigated related research towards finding solutions for various cyber threats specific to SMPs and found that clustering and classification techniques are used the most to detect malicious behaviour. Lastly, Acar [3] describes the numerous threats to children on the Internet, and more specifically refer to sexual extortion.

Given the aforementioned information, the researcher found that, in general, all threats are either some form of malware, they abuse some known network flaw, or they are personal in that they are aimed at a human or SMP account. Identity deception is found to be a cyber threat aimed at humans on SMPs. This research is particularly interested in cyber threats aimed at humans, since identity deception is found to be underlying to many of the other cyber threats. In the case of cyber bullying, for example, attackers usually change or hide their identity to avoid detection [122].

Table 2.2: Cyber threats on social media platforms

Cyber threats	Hansman et al. (2005) [150]	Gharibi et al. (2012) [126]	Perez (2011) [244]	Williard (2007) [332]	Fire et al. (2014) [114]	Kirichenko et al. (2017) [185]	Patel et al. (2017) [240]	Triveda (2016) [303]	Pradhan et al. (2016) [251]	Acar (2014) [3]
Malware					✓			✓		
virus	✓									
worms	✓					✓				
trojans	✓					✓	✓			
malicious scripts			✓		✓				✓	
Network attacks	✓									
denial of service	✓									
spam, brute force					✓		✓		✓	
insufficient authentication controls			✓							
data leakage						✓				
Physical/Personal attacks	✓									
identity theft		✓			✓		✓	✓	✓	
trolling (defamation)		✓								
flaming (a short-lived argument)				✓						
identity deception				✓	✓	✓				
cyber stalking		✓		✓			✓		✓	✓
cyber bullying		✓			✓		✓			✓
grooming (extremism, paedophilia, etc.)					✓				✓	✓
phishing	✓		✓		✓	✓	✓	✓		

2.5 Related research on identity deception on SMPs

To better understand what role identity deception plays as a cyber threat on SMPs, the researcher consulted the Institute of Electrical and Electronic Engineers (IEEE) and Institute of Engineering and Technology (IET) knowledge bases [163]. The reason

for consulting these two resources are that, together, they form the world's largest professional association. Together these institutions also produce more than 30% of the world's research work in the fields of electrical and electronics engineering and computer science [330]. The IEEE is based in the United States of America (USA), while the IET is based in the United Kingdom (UK). The IEEE and IET knowledge bases consist of many topics that include social network services (e.g. social media) and cyber security. The knowledge bases can be searched for keywords found in the meta data and text of the available research work. For this research, the knowledge bases were initially searched for the keywords "cyber security" and "social media". Although cyber security is known as many other things, like information security, the search was limited to cyber security for the scope of this research.

Since 2014, these institutions have together published 41 474 papers dealing with social media and 21 698 papers concerned with cyber security. The breakdown of topics in these published papers is presented in Figure 2.4. The figure shows how, as expected, the number of publications has increased each year and that more publications have been concerned with social media than with cyber security. As cyber security is relevant not only to SMPs, it was surprising that cyber security did not have more publications. The assumption is therefore that certain cyber threats were addressed more often than others.

To investigate which cyber threats were addressed, 2 718 papers that mention any of the personal cyber threats (see Table 2.2) in their title were extracted from the IEEE knowledge base. The amount of papers published on each topic gives an indication as to the degree in which these topics were the focus of prior research. Figure 2.5 shows that grooming, phishing and identity theft together account for more than 90% of all papers published about the cyber threats found on SMPs. Identity deception was found to feature in the title of only 35 papers. However, identity deception was an underlying concept found in all these cyber threats [204] [87] [92].

To understand whether similar patterns can be noticed in databases other than the IEEE, global web search trends from Google [133] were consulted. Google trends [133] is a public web service from Google that has been recording data since 2004 to show how often a search term is entered over time [68]. This is also known as the Search Volume Index (SVI), which can be expressed as the actual number of searches for a term, divided by the average number of searches for the same term. If the SVI for example shows '10' in 2017 and '20' in 2018, it means that the term has become twice as popular. The terms can also be compared to one another. Each search term's SVI is normalised so it

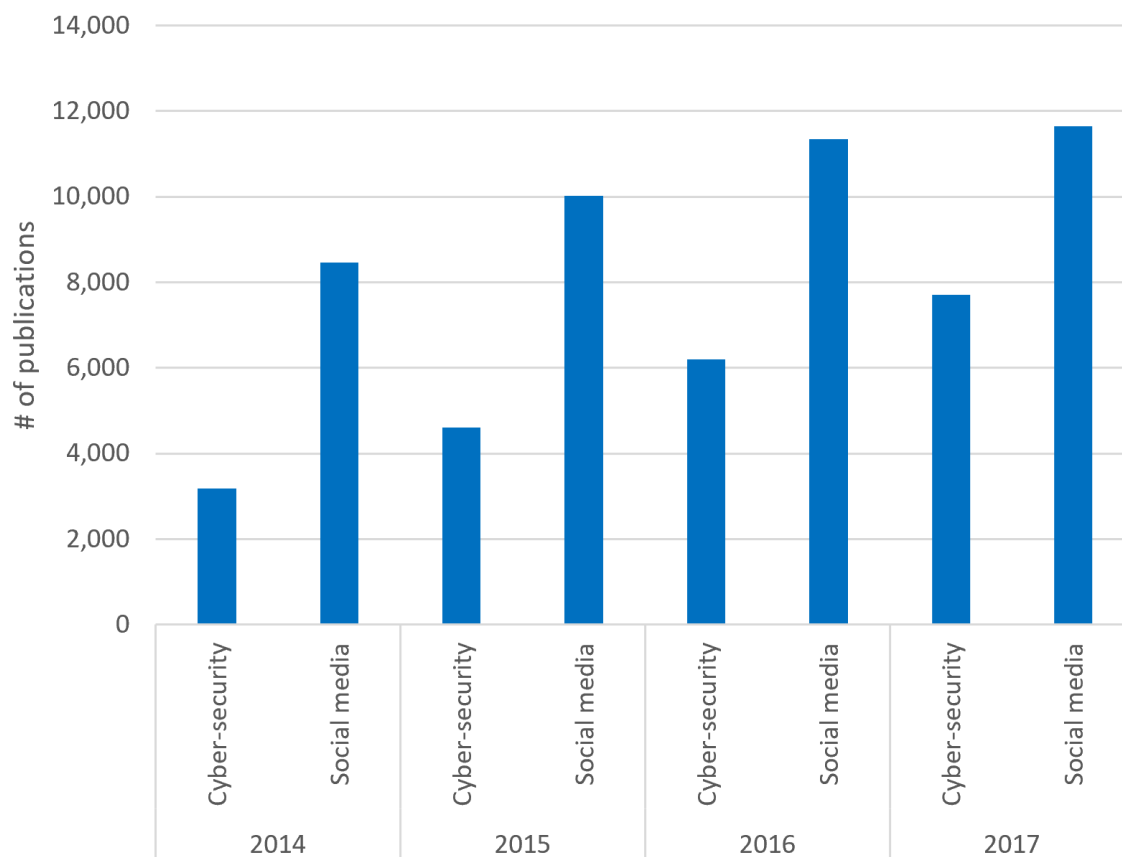
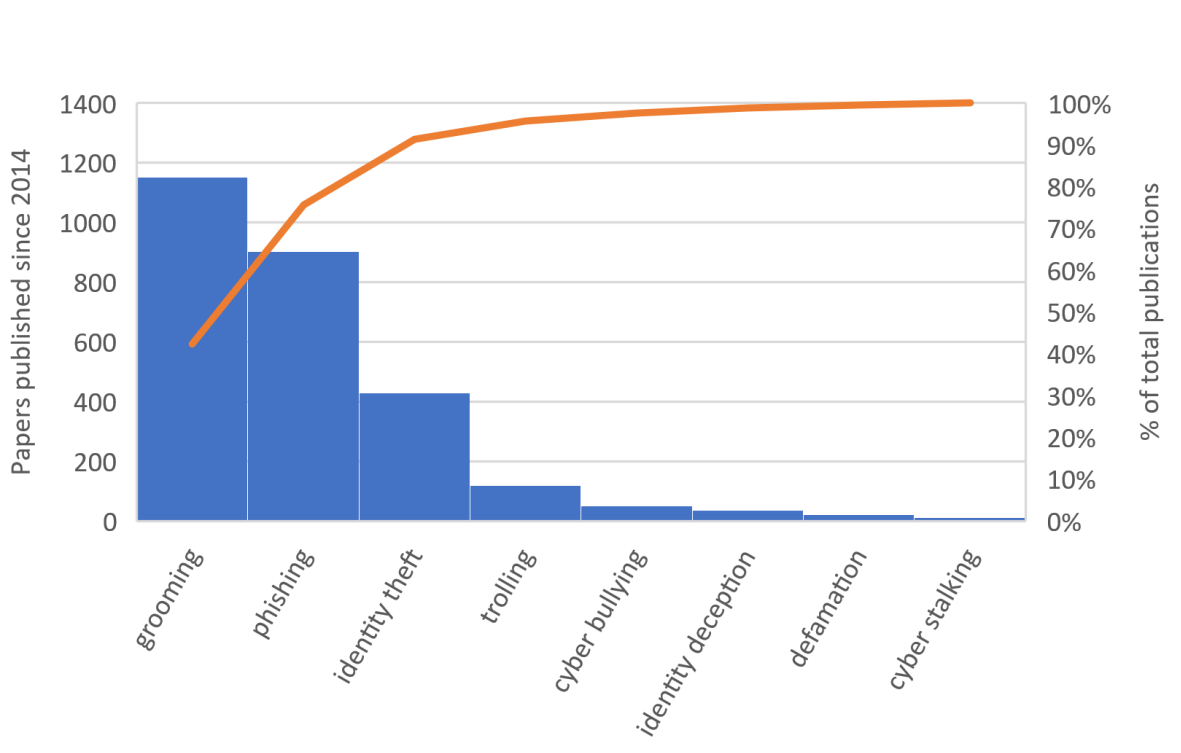


Figure 2.4: Total papers published per year, per topic

is comparable to the others. The researcher used Google trends to determine the current level of average interest for each of the cyber threats found on SMPs over the past five years.

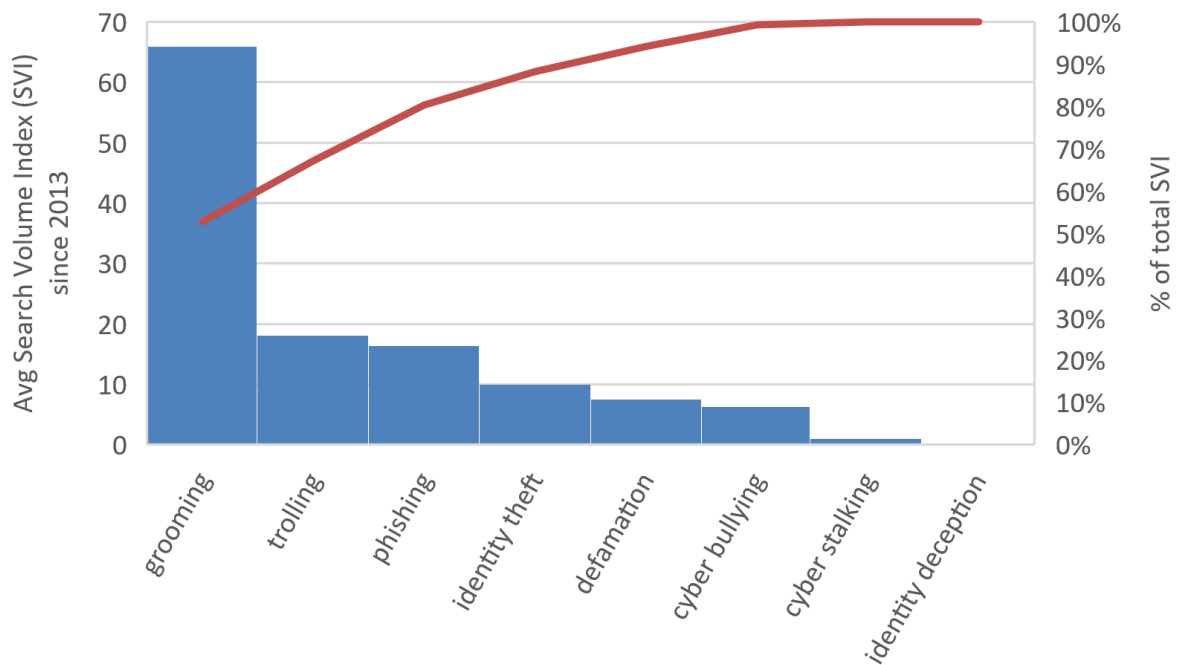
As depicted in Figure 2.6, grooming was of highest interest to the public, and alone it covered more than 90% of the total threats. The threat of trolling seemed to receive second-most attention in the public domain and identity deception the least. The fact that grooming is a well-searched topic on Google is not surprising. However, identity deception can be seen as a strategy used by attackers in many of the other cyber threats, including grooming [182]. Many different terms are used to describe identity deception, like fake identities, impersonation, social engineering, and masquerading. For these reasons, identity deception is not as often searched for in public circles than are the other cyber threats, but that does not mean it is less important.

To show the importance of identity deception, the abstracts of all 2 718 IEEE papers were presented as a single word cloud. A word cloud counts the frequency of words and presents the result in a visually appealing way, with more frequent words displayed



*The red line denotes the cumulative total publications

Figure 2.5: Papers published about cyber threats on SMPs



*The red line denotes the cumulative total SVI

Figure 2.6: Google trend search results since 2013 for cyber threats on SMPs



Figure 2.7: Keywords pertaining to social media threats

larger [158]. To ensure that only articles pertaining to social media were included in the word cloud, only those abstracts mentioning the word ‘social’ were included. The resultant word cloud for the top 100 key words is depicted in Figure 2.7. It clearly shows that ‘identity’ is among the top 100 key words and important in all cyber threats found on SMPs, even though it does not seem to have been the focus of the research itself.

Having said the above, it is clear that identity deception is a cyber threat. It was shown how identity deception is found to be underlying to other cyber threats as well. Since this leaves humans who are vulnerable to identity deception on SMPs particularly vulnerable to all types of personal cyber threats, this calls for research to be done on how to protect humans against identity deception on SMPs.

2.6 Conclusion

Big data has grown from a concept or buzzword to something tangible with measurable characteristics. SMPs show the same characteristics as are expected from a big data

platform. A literature review of available research work also showed how various SMPs share similar attributes and thus the research performed on one particular SMP could equally apply to another. The big data characteristics found on SMPs contribute to the exposure of humans to a variety of cyber threats. This exposure comes at great scale and complexity.

The next chapter presents the requirements of a model that will assist in the automated detection of identity deception, as executed by humans, on SMPs. The discussion will delve into deception and identity deception in much more detail – not only to understand these concepts and their role in cyber security, but also to show current research work on deception and identity deception in conjunction with SMPs.

Chapter 3

Cyber Security and Identity Deception

“There is nothing more deceptive than an obvious fact.” — Arthur Conan Doyle, *The Boscombe Valley Mystery*

3.1 Introduction

At the beginning of 2018, statistics showed that more than half of the world’s population (4 billion) use the Internet and more than a third (3.1 billion) use social media [64]. With the growth of Internet and social media usage, new abilities have been added with the intention of benefiting society. Some of these current benefits are the tracking of natural disasters [71]; the tracking of sports events [338]; the prediction of public crowd gatherings [38]; and detecting violations of the freedom of speech [290]. This contrasts with the various cyber-security threats that result from Social Media Platforms (SMPs) as described in Chapter 2. One such threat was identity deception.

This chapter discusses identity deception in detail to identify a set of requirements for a model that will assist in the automated detection of identity deception by humans on SMPs, which is the main research question. For this research, terms like ‘deception’ and ‘identity’ require further explanation. This is followed by a discussion on the various aspects that should be considered when one deals with identity deception on SMPs, such as the origin of identity deception; the attributes found on SMPs in which identity deception is prevalent; and how to build a model that assists in the detection of identity deception. To conclude, a set of requirements for a model that will assist in

the automated detection of identity deception on SMPs is defined.

3.2 Deception

Deception is defined as a “deliberate attempt, without forewarning, to create in another a belief which the communicator considers to be untrue” [323]. The Oxford English Dictionary (OED) defines deception as “the action of deceiving someone”, and deceit as “the action or practice of deceiving someone by concealing or misrepresenting the truth” [236]. Deception is executed for a purpose and can be achieved through various strategies.

Several studies represent deception based on its purpose [324] [177]. Wang [324] shows how deception has three main purposes, namely concealment, theft and forgery. According to Kashy [177], deception is performed for the purposes of manipulation, impression management, insecurity, socialisation, sociability, or relationship management. In general, deception is simply when the truth is misrepresented. Similarly, related research was found to represent deception based on the strategy that was used to deceive [261] [22] [56] [200]. Interpersonal Deception Theory (IDT) defines the following strategies of deception [56]:

- Falsification or changing of the facts
- Exaggeration, overstatement, or minimisation of facts
- Omission of important information
- Equivocation or presentation of vague information to leave a false impression

Truth Deception Theory (TDT), on the other hand, defines the following strategies of deception and closely resembles IDT: lies; omission; evasion; equivocation; and generating false conclusions from true information [200]. TDT assumes that people have a truth bias towards deception, which makes them insensitive towards detecting deception. IDT indicates that people are generally more sensitive to the deceit of others. Most recent research showed that TDT is superior to IDT in detecting deception [239]. The current research, however, focuses on the IDT strategies used for deception as defined by Buller and Burgoon [56], since our focus in this study is more towards detecting deception than detecting truth.

These deception strategies manifest in different forms on SMPs. For example, lies can be spread about opposition parties in elections [76], a person can lie about their

identity [306], or fake news can be published that creates unnecessary anxiety to the detriment of public safety [285]. For SMPs, this means that deception can potentially be found in many of the SMP attributes. For example, when an account holder is asked to provide a profile image for their account, they may well upload a profile image that is not representative of themselves [142]. It is also known that account holders purposefully change the location of their SMP account, to make others believe that they are somewhere else than where they actually are [12]. In a further example, humans are able to make untrue comments in the content they post to SMPs. Vosoughi et al. [322] found that false content dominates the truth on SMPs like Twitter. Some of this false content happens to be on those attributes identifying an SMP account holder [12].

The previously mentioned examples show that SMP accounts can be created from the onset with false information about the account holder or they can later be manipulated to make others believe that the attacker is someone they are not. These lies can be found in the SMP attributes that describe who they are, also known as their identity.

3.3 Defining an identity

Identity is defined as those features that describe who an individual is or any qualities that they display that can be used to distinguish them from another [324] [72]. Related research has been conducted into identifying deceptive people who lie about their identity. Goel et al. [131] used a graph-based algorithm to find similar accounts and showed that similar accounts share common characteristics like location, email domain, interests, friends, followers, and topics. Kim et al. [183] used the names of the groups people belong to, to detect the characteristics of those who belong to the groups. A group called ‘family’, for example, typically points to the fact that everyone in that group is related to another by blood. Another group named ‘running’, for instance, could show that everyone in that group is related by their common interest in running. Wang et al. [324] divided identity attributes into the following three groups: personal information; biometrical attributes that belong to an identity; and biographical attributes that change over time for an identity – for example where they have lived. Clarke et al. [72] divided identity attributes into the following groups: attributes describing appearance; name; the code that uniquely identifies you from another (for example an identification number); your social behaviour; knowledge; what you have; what you do; who you are; and your physical characteristics. A comparison between these identity attributes from related

Table 3.1: SMP attributes vs Identity attributes

SMP attributes	Goel et al. (2013) [131]	Kim et al. (2010) [183]	Wang et al. (2006) [324]	Clarke et al. (1994) [72]
Describing the profile	Location		Personal information	Appearance, name, code, who you are, physical
Describing the account	Email		Biometrical information	What you have
Behaviour	Mutual interests	Belong to the same group	Biographical information	Social behaviour
Relationships	Friends/Followers			What you do
Content	Topics			Knowledge

research and the attributes described as being available on SMPs is shown in Table 3.1. It shows that sufficient information is available on SMPs to describe an identity. Deception can occur on any of these identity attributes [12] [204] [142].

SMP account holders can lie about their name [307] upon opening an account. They can also continually tell lies by posting content that is untrue [73]. This last example is sometimes referred to as ‘rumours’ [140] and is prevalent in cyber threats like cyber bullying [122]. It also shows that the deception found on an account is not always a once-off event but can be a continual occurrence.

3.4 Identity Deception

Since it was shown in section 3.2 that deception is executed for a purpose, based on a given strategy, and that one of those strategies could be to lie about attributes on SMPs that describe an identity (as described in section 3.3), a more detailed investigation on identity deception is required. Identity deception occurs whenever deception is used to assume the identity of another. Identity deception is recorded to have happened as early as in the Old Testament of the Bible where Jacob donned his brother’s clothes to deceive their father into giving him the inheritance that rightfully belonged to his brother Esau [49]. With identity deception, a deceptive account is created for various purposes. The deceptive account’s purpose could include spamming another [39]; defaming the character of a company or person [122] [39]; inflating popularity [80] [141]; hiding so as to remain anonymous [175]; or recruiting/grooming for extremism [268]. These purposes relate to what Wang [324] said about deception, in that humans deceive to conceal, steal, and forge. To achieve these outcomes, SMP account holders have to lie about who they are. Identity deception thus occurs when anyone lies about those attributes that describe their identity. The researcher acknowledges that identity deception could also be performed for the purpose of anonymity as to offer a human protection or allow for freedom of speech amongst others. The scope of this research is however to find those

humans lying about their identity for malicious purposes.

To help lessen the threat of impersonation, fake identities, identity deception and the like, cyber-security laws and regulations have been proposed to protect humans on social media platforms. Examples of such laws are Communications of Decency Act (CDA); Child Online Protection Act (COPA); Children Internet Protection Act (CIPA); Deleting Online Predators Act (DOPA); and Children Online Privacy Protection Act (COPPA) [205]. CIPA, for example, forces filters to be implemented on online content to protect users from obscenity, pornography, and harmful content. CIPA is effective when fake identities present harmful content to SMP users but does not cater for situations where fake identities are, for example, used to groom a user on an online dating site, because their conversations and content are similar to other conversations expected on such a platform. DOPA, on the other hand, protects users by blocking certain sites and chat rooms. DOPA protects users from others who use impersonation on the blocked sites, but it does not prevent them from being targeted, through impersonation, on other permitted sites. The General Data Protection Regulation (GDPR) which became law in May 2018 [8], aims to protect a human's data and privacy by preventing others from storing such individual's information without their explicit consent. For cyber threats like identity deception, it means that the malicious attacker will find it increasingly difficult to find information on other humans they would wish to impersonate. However, this does not stop a malicious attacker from creating a fake identity that does not resemble that of an existing user.

Besides laws and regulations, several technologies have been proposed to assist in the protection of humans against identity deception on SMPs, for example plugins [256], Application Program Interfaces (APIs) [231], and software systems [106] [172]. A technology is the application of knowledge for a specific purpose [236]. These technologies differ in who they protect from, what deception they can detect, and the methods used to detect the deception. For these reasons, and to provide a structure of understanding as to how these technologies address the threat of identity deception on SMPs, the researcher proposes to delve into the following topics: who commits identity deception; the attributes used to deceive on SMPs; and what methods can be used to build a model that will assist in identity deception detection.

3.4.1 Who commits identity deception on SMPs?

When dealing with SMPs, there are various potential deceivers. SMP account holders can either be human, computer generated (also referred to as bots) or cyborgs [70]. A cyborg is a half human, half bot account [70] that is manually created by a human, and afterwards the actions on the account are automated by a bot. Several variations, created for various purposes, exist on bot and human accounts. For example, bots generated to gain information are known as social bots and bots created to mimic real accounts are known as spam bots [142] [320] [106]. Humans on the other hand, create accounts to, for example, influence the opinions of others [351], to damage the reputation of another [122], to fake illness [59], or to groom others [268]. Not all accounts show malicious intent and could be deceptive purely to remain anonymous [175].

Galán-García et al. [122] proposed to find human accounts by analysing the contents posted by the users of SMP accounts. Their research work used a single account, verified as belonging to a human, and manually selected related friends to construct a dataset for their research experiment. Their hypothesis was that a fake human account is usually followed by the real profile of the human behind the attack. This allows the attacker to hide their identity as everyone else sees communication between their real account and the attacking account without realising that these accounts are owned by one and the same person. Their research was successfully deployed to stop cyber bullying at a school. Tener et al. [293] presented research results in the form of a topology of offenders who use online media to commit crimes against minors. It was interesting to note the remark that most victims would have ‘run away’ from the offender if they had actually met them in real life. This indicates that identity deception was used in many examples of grooming to lure the non-suspecting victims. The results also proved that identity deception was present on SMPs. Klausen [186] investigated the influence of accounts controlled by terrorist groups and found that even though their dataset contained known terrorists, it was difficult to identify them.

Figure 3.1 illustrates human and non-human accounts. The focus of the current research, namely the detection of deceptive accounts created by humans (as opposed to bots or cyborgs) on SMP platforms for some malicious purpose, is highlighted in dotted lines. This focus is required due to the fact that humans, as opposed to bots, usually execute malicious acts and cyber threats like grooming, cyber stalking, and trolling [122] [126].

To identify identity deception by these humans on SMPs, it is necessary to look at how

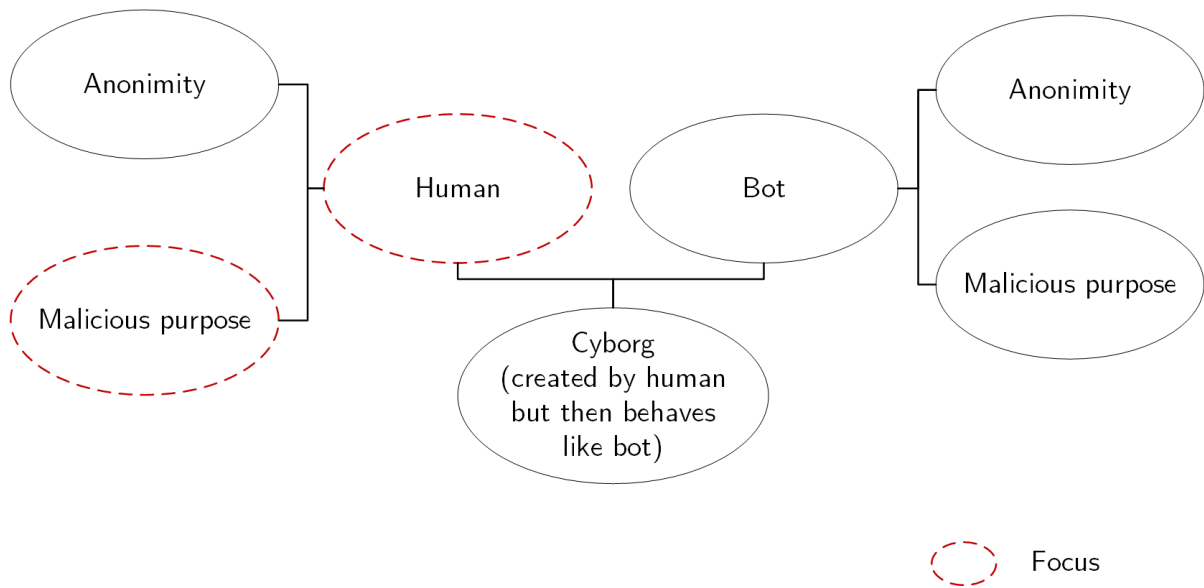


Figure 3.1: Deceptive role players on SMP and the focus of the current research

we use the information we have on these humans, namely the attributes that describe their identity.

3.4.2 Attributes and features used to detect identity deception on SMPs

When looking at how to assist in the automated detection of identity deception on SMPs, it is important to understand not only which attributes exist, but also which attributes can potentially contain false information and therefore have a bigger impact on identity deception. Some attributes, for example the date the account holder opened his/her SMP account, cannot be changed and therefore this attribute cannot be modified. However, this does not mean that these attributes should be disregarded when investigating identity deception. Consider for example that when combining the date the account was opened with information about a person's current age, a new feature such as 'compare age' could be engineered, which can be indicative of deception. Table 3.2 illustrates this concept, also known as feature engineering [99], with the help of examples.

Not only are the engineered features important, but related work in machine learning [80] [142] shows that the choice of attribute or engineered feature will influence the chance of successfully detecting identity deception on SMPs. Choosing for example only the opening date of an account provides very little information about the

Table 3.2: Using multiple attributes to engineer features for deception detection

Example	SMP attribute 1	SMP attribute 2	Newly engineered feature	Potentially deceptive?
1	Date when account was opened	Age	Compare age	
	20-Jan-16	45	43	No
	31-May-08	18	8	Yes, age less than allowed by SMP
	30-Jun-08	105	95	Yes, highly likely
2	Time Zone	Location	Distance time zone	
	UK	United States	8,000 km	Yes, far from expected location
	UK	UK	0 km	No
3	Profile image	Name	Compare gender	
	Female	Sarah Johnson	Match	No
	Female	John Smith	No match	Yes, the profile image shows an image of a female, but the name is indicative of a man
4	Name	Display name	Levenshtein	
	John Smith	JohnSmith	3	No, low Levenshtein distance
	Sarah Johnson	timothybates	25	Yes, high Levenshtein distance

deceptiveness of the user. But that, combined with the age of the user, could be quite valuable, as it can tell us how old the person was when they opened the account. This concept is also known as feature selection [247].

Conroy et al. [73], Booth et al. [50], Hauch et al. [153], and Appling et al. [22] all used linguistic features extracted from various SMPs to detect identity deception. Examples of such linguistic features are: repetitions [197]; sharing the same naming structure, for example ‘JohnSmith’ being very similar to ‘JohnSmit2’ [298]; number of URLs in a tweet [197]; and the number of hashtags in the content [113]. Dickerson et al. [95] on the other hand proposed to use sentiment features in the identification of identity deception. They extracted the sentiment from conversations on various topics found on an SMP. If the sentiment of one account differed from the average sentiment recorded for all conversations on the same topic, the account was classified as being potentially deceptive. Non-verbal features like the date the account was opened [298], the type of SMP [298], and profile update time [142] were useful where the information provided for an account was scarce. Network features, like accounts in the same domain [298], trending conversations [283], friends [142], and followers [142] were used to detect deception. Lastly, identity features such as gender [160]; location [12]; profile image [149]; age [307]; profession [307]; name [242]; and email [340] were proposed as indicators towards identity

Table 3.3: SMP attributes vs Cost classes defined by Cresci et al. [80]

SMP attribute category	Cost class	Example SMP features
Describing the profile	Class A (profile)	profile name, age
Describing the account	Class A (profile)	opened date
Behaviour	Class B (timeline)	retweets
Relationships	Class C (relationships)	friends, followers
Content	Class B (timeline)	tweets

deception.

Cresci et al. [80], however, showed how non-verbal features and identity features on their own are good enough to detect identity deception. Not only did they list the engineered features to detect bots, but they also grouped these features into what they called ‘cost classes’, or more specifically, the effort required to mine and engineer each feature for identity deception detection. The first class contained features engineered from the SMP attributes that describe the account or user profile. Examples of such features are the profile name and the date the account was opened. The second class contained those features engineered from information about the content a user posted, including behaviour over time. Examples of such features are the time a tweet was made and how many times tweets were forwarded or retweeted. The last class contained those features engineered from the relationships that one account has with another. Examples of these features are the friends and followers of an account. Table 3.3 shows the SMP attribute categories compared to the cost classes defined by Cresci.

The results from Cresci et al. [80] are important for the research at hand. They were able to demonstrate that by selecting features engineered only from Class A, accurate results that were almost as good as when all features were used, were still being produced. This means that for a lower cost, an accurate identity deception result can be achieved by only using features engineered from the account and user profile. It takes, for example, less time to only gather the data of a user profile compared to having to gather all the content they have posted as well. Some accounts could contain thousands of content items like photos. Even more so, it takes less time to gather information about the user’s friends and followers and their related content. Only the account and user profile SMP attributes and features will therefore be considered for this research.

3.4.3 Methods used to detect identity deception on SMPs

A method describes the way in which a goal will be achieved [236]. For the research at hand, the goal was defined as being able to assist in the automated detection of identity deception on SMPs by humans. Examples of the methods used in related research to achieve the same or similar goals are discussed next.

3.4.3.1 Rules

Rules are mostly reactive. Whenever a new threat of identity deception is identified, a new rule is added to counter such threat. The new rules can be as simple as a filter. Filtering techniques are common in email where senders are placed on a blacklist [155]. Similar filters to blacklist known malicious users have been proposed for SMPs [297]. Filtering, however, becomes very difficult when the identities used by spammers employ dynamically adaptive and automated strategies to circumvent the proposed methods. New bots, with different names, can be created at a scheduled time each day, which renders the current blacklist obsolete and unable to keep abreast of such attacks. This is just as true for human accounts on SMPs. Humans easily adapt themselves to avoid detection and, in the case of blacklisting, they simply create a new account and fake identity [80] as soon as the current detected account is blacklisted.

Besides filtering techniques, more complex rules have been established to identify fake accounts. Examples of such rules are specific words (such as ‘winner’) or combinations of words (such as ‘buy direct’) that are known to be used in messages that are spam [31]. These words are combined in a word dictionary [31]. If a message contains words from this word dictionary, it is regarded as spam. These same rules have been successfully applied to SMPs to detect fake accounts, like bots [39]. The problem, however, is that the word dictionaries quickly become outdated, and abbreviated words such as ‘rofl’ meaning ‘rolling on floor laughing’, are commonly found on SMPs. This is problematic in the sense that detection rules become outdated. More advanced rules were proposed on SMPs, such as pattern matching [142]. For example, if a fake account has been tweeting about three or more trending topics, or if a fake account took part in trending topics but is less than a day old, it can be classified as deceptive [192]. On Facebook, Fire et al. [115] used rules to score friends for deceptiveness, given their similarity. The similarity rule for two users was calculated as the sum of their common friends, chats, groups, posts, tagged photos, tagged videos, and family. They went further to assign weights in which ‘being family’, for example, counted 1000 times more than just ‘having a common friend’.

Also, ‘being tagged in the same photo’, for example, counted twice as much as ‘sharing common friends’. These rules had some success but were not deemed generic enough. Human behaviour is more random [254] and therefore more complex rules are required. Fire et al. [115] had better results with supervised machine learning.

3.4.3.2 Supervised machine learning

Research suggested supervised machine learning models that can detect fake accounts . For email spam detection, Tuteja [309] proposed supervised classification machine models such as Support Vector Machines (SVMs), decisions trees, Naïve Bayes and neural networks. For SMS spam detection [69], ten features (among others, SMS length) were engineered by Choudhary and Jain [69]. These features used supervised machine learning models like random forest, decision trees, J48, logistic regression, and Naïve Bayes to predict SMS spam with great success. Cresci et al. [80] proposed a supervised machine learning model based on the attributes describing the identity of an identity only, like the number of friends, the length of their name, and the time since the account was opened, to detect bots on SMPs. Gupta et al. [140] in turn suggested that behaviour, such as the frequency of messages and time of day, provides enough information to detect bots successfully through supervised machine learning models. Xiao et al. [340] proposed logistic regression, random forests, and SVMs to detect deceptive accounts. In their research, they combined basic distribution features (the average number of tweets of a user), pattern features (the number of operations required to change a value from one to another), and frequency features (the number of times a name is found in the corpus).

Regardless of the algorithm used, supervised machine learning models require that a label of the expected outcome be included in the corpus to build a model [227]. For this research, the label would indicate whether the account was deceptive or not. Table 3.4 shows an overview of the various supervised machine learning algorithms that were proposed to detect fake accounts. The table shows that the algorithms used for the detection of fake accounts covered various algorithm families. Machine learning algorithms were grouped into different families depending on their underlying math [112]. Various algorithm families had to be covered, since different algorithms were better at solving different problems. This is also known as the “no free lunch” theorem [112]. By trying various algorithms to build a model, the best model for a specific use case had a better chance of being found.

Table 3.4: Supervised machine learning algorithms used to detect bot and spam accounts

Machine learning algorithm	Algorithm family	Research detecting identity deception								
		Dickerson (2014) [95]	Tsikerdekis (2017) [304]	Gupta (2013) [140]	Fire (2014) [115]	Cresci (2015) [80]	Galán-García (2015) [122]	Soman (2014) [283]	Bu (2013) [55]	Xiao (2015) [340]
Adaptive boosting	Boosting	✓	✓		✓	✓				
Naïve Bayes	Linear	✓		✓	✓	✓				
J48 library from Weka	Tree			✓	✓	✓	✓			
K Nearest Neighbours	Clustering					✓	✓			
Neural Network	Neural Network							✓	✓	
Random Forest	Tree	✓	✓		✓	✓	✓			✓
SVM with Radial Basis Kernel	SVM	✓	✓			✓		✓	✓	✓
Gradient boosting	Boosting	✓								
Rotation forest	Tree				✓					
Logistic regression	Linear					✓				✓
Sequential minimal optimisation	SVM						✓			

3.4.3.3 Semi-supervised machine learning

A semi-supervised machine learning algorithm uses a labelled and unlabelled dataset to create a model [354]. Semi-supervised machine learning works well where very little labelled data is available. The algorithm will run the information from the unlabelled dataset through several different methods to refine and enrich the labelled dataset during the training of a model [157]. Various semi-supervised machine learning algorithms exist that include *inter alia* co-training methods [157]; self-training methods [253]; support vector machines [157]; generative models [253]; and graph-based methods [253].

Semi-supervised machine learning has been used in related work to detect fake accounts. Ebrahimi et al. [104] compared a one-class support vector machine model (semi-supervised model) to a Naïve Bayes supervised machine learning model and showed how the one-class SVM detected fake accounts used to groom others with greater success and better accuracy than the supervised machine learning model. A public dataset that contained information about past examples of such threats was used in their research. Their work was interesting in that their dataset contained an

example of a minority class, similar to the research at hand. Li et al. [204] used a graph-based semi-supervised machine learning method to detect fake social engagement, for example ‘likes’ on videos originating from fake accounts on YouTube. Sedhai et al. [271] proposed their own new semi-supervised method to detect spam on Twitter. Their algorithm was based on a self-training method that learned how to detect blacklisted domains, duplication domains and spam.

Semi-supervised algorithms must be able to leverage the unlabelled data to create a model. This however requires the labelled dataset to be representative of the population. For this study on SMPs, it was not practical to mine the minority class consisting of fake accounts [104], as there was no certainty that an account was indeed deceptive [174].

3.4.3.4 Unsupervised machine learning

With unsupervised machine learning, the data is unlabelled and grouped based on similarity [227]. The corpus’s result or outcome is initially unknown. Unsupervised machine learning was successfully applied by Guel al. [137], Wiet al. [335], and Yahyazadeh and Abadi [342] to detect bots in network attacks. Many bots are for example used to send the same message to a targeted machine on a network [335]. The research mentioned showed how clustering, which is a common unsupervised machine learning method, can be used to detect bots. Clustering successfully detects bots in network attacks since these bots usually share similar characteristics, such as the same domain, IP, and network traffic flow. The same can be said for when bots are used on SMPs to spam users. Xiao et al. [340] showed how the same message will usually be sent on the same date, and that these attacks or campaigns can be detected through clustering. They were able to detect spam campaigns with an accuracy metric called Area Under Curve (AUC) of 0.98. The same can unfortunately not be said for fake human accounts. Fake human accounts usually target specific individuals with specific messages [312], therefore clusters cannot easily be formed for such attacks. For this reason, unsupervised machine learning was not considered to solve the research problem at hand, although the features identified as valuable in the detection of deception will be explored in the next chapter. These features could complement a model that assists in the automated detection of identity deception by humans.

3.4.3.5 Reinforcement learning

Venkatesan et al. [319] presented a reinforcement proof-of-concept model. Their model continually updates itself, based on its past performance. If no further improvements can be made at a predetermined time, the model remains in its current state. This decision is made by considering various new options, such as using new combinations or weights of features, to protect against fake accounts. Some of these options might even be sub-optimal. If an improvement can be made, the model will be adjusted and remain in its new state till the next evaluation. Arif et al. [24] used a similar iterative process that takes the most important features during every iteration to build a model that can detect spam on SMPs. Reinforcement machine learning models require feedback from the environment, such as whether the new model worked well or not. This is not readily available on SMPs, as the SMP would need to let the model know how successful it was at detecting deception. Additional development is required to acquire this feedback from users and even then, their feedback may be found to be biased or opinionated. Therefore, reinforcement learning was not considered for solving the research problem at hand.

3.5 The requirements for the detection of identity deception by humans on social media platforms

Based on the previous defined terms and related work, the researcher defined the following generic requirements for a model that will assist in the automated detection of identity deception by humans on SMPs:

- Identify identity deception by humans – non-human accounts are disregarded.
- Find those humans who are deceptive about their identity for malicious purposes.
- Given IDT, focus on finding deception rather than on finding the truth. This means that attributes and engineered features will be indicative of deception as opposed to the truth.
- Use only attributes for defining a user account that are available on SMPs.
- Ignore content posted by users on SMPs.

- Ignore attributes that do not contribute to human identity deception detection.
- Features should be engineered such that they complement the automated detection of identity deception. The correct feature selection will lead to better results.
- Develop a supervised machine learning model.
- Display the ability to evaluate various machine learning models.
- Be able to use labelled data.
- Ensure that the machine learning model results are reproducible.
- Ensure that the machine learning model results are interpretable.
- Ensure the automatic detection of identity deception by humans on SMPs.

3.6 Conclusion

Various reasons for identity deception nowadays exist on SMPs. Identity deception is achieved by lying about some or all the SMP attributes. Examples of SMP attributes indicative of identity deception are name, location, and content. SMP attributes can be used ‘as is’ to detect identity deception, or they can be transformed into newly engineered features that illustrate deception in a more conclusive way than the original attributes. The scope for finding better engineered features towards the automated detection of identity deception is broad. This is also a well-known problem in machine learning in general, where the choice of features (or feature selection) can have a huge effect on the accuracy of the trained machine learning models. In the case of the current research, this implies that the right attributes and features will result in the successful detection of identity deception on SMPs.

Various attributes and features were already identified in related research. Much can be learned from research work in detecting fake accounts, such as bots and those used in spam cyber threats. It should also be considered that other fields might contribute to finding better features to assist in the automated detection of human identity deception. Consider, for example, the field of social science, and the fact that deception has long been studied in the social sciences, especially in psychology [103]. This, combined with features engineered successfully in the past to detect bot accounts, will be discussed in the next chapter.

Chapter 4

Learning from bots and the social sciences

“Research is to see what everybody else has seen, and to think what nobody else has thought.” – Albert Szent-Gyorgyi

4.1 Introduction

This chapter discusses two fields in which identity deception examples are abundant to gain knowledge into how research from these fields could complement a model that assists in the automated detection of identity deception by humans on Social Media Platforms (SMPs). The previous chapter showed that the correct choice of attributes and engineered features is required in a model that will identify identity deception on SMPs by humans.

Computer-generated accounts, also referred to as bots, are known for their deception on SMPs [340]. These fake SMP accounts hide their identity for malicious purposes like spamming [287], terrorism [338], creating false rumours [140], and political manipulation [274]. Research such as the work of Yang et al. [344] and Stringhini et al. [287] is shown to have the ability to detect identity deception, as presented by bots on SMPs. Their research proposes various SMP attributes, like the number of followers [344]; the number of messages [287]; engineered features such as the friend-to-follower ratio [344]; and URLs-extracted-from-message [287]. The field of social sciences, known for studying humans and their relationship to the world and people around them [43], also addresses identity deception by humans. Social sciences,

and especially the field of psychology, have studied various aspects of human deception, such as why humans lie [94] and whether such lies can be detected [107], as well as those identity features humans are deemed to lie about the most, such as name [284] and gender [103].

This chapter will discuss related work from the fields of bot detection and psychology, with the focus on identifying those attributes and engineered features that are most indicative of identity deception. The chapter shows how the attributes and features from each field complements a model proposed by the research at hand to assist in the automated detection of identity deception by humans on SMPs.

4.2 Identity deception in bot research

‘Bots’, short for the word ‘robot’ [342], have been defined as being *inter alia* automated programs [70], autonomous entities [317], or automated agents [128]. These definitions highlight that bots differ from humans in that they are able to perform activities without any intervention. For the purpose of this research, we refer to bots found in SMPs as opposed to bots found in other areas (e.g. the motor vehicle industry) [232].

Just as humans are a threat to other humans on SMPs, bots are a threat to humans too. It is a known fact that many humans often befriend people they do not know [115] and therefore they can also easily become the victim of a bot [287]. Bots are known to be widespread on SMPs [340] [106]. It is important to find these bot accounts, as they weaken the credibility of the SMP [340] and present many cyber threats to the humans using these platforms [317]. The following are examples of cyber threats executed by bots on SMPs:

- For *identity deception*, bots create fake identities and hide the fact that they are non-human to increase their importance [344]. On an SMP like YouTube, bots are used to increase the popularity of specific targeted videos [204] by giving reviews on content from fake accounts. In YouTube, the more a video is seen, the more money is paid to the producer of the content. There is thus a monetary gain, should one video be more popular than another.
- For *identity theft*, bots are known to mimic famous people by creating a profile similar to that of the target. An example of such identity theft is of a bot that was created in 2017 and that retweets content from Donald Trump and changes the formatting of the content to look like it could be official statements made by

the president himself [243].

- Bots are used in *grooming*, such as online extremism [113] and political rallying [41]. Such bots post fake content to exaggerate the need for political change and create in humans a desire to join these extremist forces. The bots also post fake content, as if they were from real humans, to convey mass support for a candidate or policy. This hype has an effect on human opinions and could also change the outcome of (for example) an election [41].
- Bots that spread *false rumours* [140] are comparable to cyber bullies; the target of the attack is just a large group instead of a single person. An example of false rumours, as spread by bots, was found during hurricane Sandy [285], when fake images about the threat level of the hurricane caused unwanted pandemonium among citizens. These bots hide their identity but make their claims as if these came from a human. Another example of rumours being spread [211] shows how bots can influence the price of Bitcoin by posting more positive or negative comments about the crypto currency online. These bots hide their identity to avoid detection.
- Bots are used to spam humans with false advertising messages [344]. These false advertisements could be a front for *phishing and fraud* by requesting personal information from a human with the intention to then steal their money [340].

Research has done much to propose attributes and engineered features that could identify bot accounts, given these cyber threats. As the boundaries between bot and human SMP accounts are not clear [317], this same research could complement a model that tries to detect deceptive humans. Research to detect bot accounts not only shows which attributes and engineered features are used, but in some cases the results also reveal which features are found to be most indicative of identity deception. Related bot research will be discussed next, followed by a summary in Table 4.2.

4.2.1 Attributes and engineered features used to detect identity deception in bots

Attributes and engineered features can be found that address the problem of bots in general and that do not focus on a specific threat. Gilani et al. [128] found that bots retweet more, and their tweets contain more URLs, whereas humans contribute more novel content to the SMP. Their research shows that behavioural attributes (like replies)

and demographic attributes (like account age) could identify bots. Gurajala et al. [142] found that bot accounts share similar profile images as well as friends-to-followers ratios. Their work shows that bots have fewer followers and that bot profiles are updated in batches, as opposed to humans who would update their profiles independent from another profile. Stringhini et al. [287] collected data from Facebook, Twitter and MySpace to attract bots used for dating, spam and fraud campaigns. Once they had a corpus, features were extracted and the behaviour was monitored. They identified different targeted campaigns by grouping together those accounts that share similar content. The results showed that bots deceive the same way across different types of cyber threats.

There is also research that focuses specifically on detecting bots involved in the cyber threat of spam. Spam refers to a situation where the target is bombarded with unsolicited messages [287]. Li et al. [204] looked at fake accounts generated to inflate popularity on an SMP like YouTube by generating more good reviews or ‘likes’ for a specific video. By using graphs to represent relationships between the SMP users and the videos, they were able to identify that some users consistently comment on and review the same videos. This similarity was found to be indicative of a potential bot. Yang et al. [344] proposed to design more robust features to detect spamming on Twitter that originates from fake accounts. They engineered features stemming from relationships (like similarity), from the account profile (like the number of followers), from content (like shared URLs), and from behaviour (like the tweet rate). The results showed that the friend-to-follower ratio, the profile creation time, and the source of the content were significant to detect bots. Benevenuto et al. [61] engineered over 60 features dependent on the content or the relationships of the SMP account holder. They were able to detect bots, with the results showing that the following features were most indicative of deception: the fraction of tweets with URLs, the age of the user account, the average number of URLs per tweet, and the fraction of followers to friends.

Some research work targets specific cyber threats from bots. Ferrara et al. [113] identified accounts involved with extremist campaigns and their work proposed 52 engineered features. The results showed that the ratio of retweets to tweets, the average number of hashtags adopted, the sheer number of tweets, and the average number of retweets generated by each account, were indicative of a bot extremist campaign. Xiao et al. [340] proposed to identify clusters of bots that share the same creation date, IP, and location. Their approach was different in that the engineered features were based on the whole group as opposed to a single account. Their assumption was that all accounts in a group would be fake. The results showed that

clusters of fake accounts used, for example, similar names. Fire et al. [115] proposed a method to protect a human's security and privacy on Facebook by restricting potential fake accounts from accessing personal information. They used engineered features like whether users belong to the same family or whether users share common chat messages. Their solution manifested as a browser plugin that would warn people of any connections they already have with other potential fake accounts. Egele et al. [106] studied how to find compromised accounts on SMPs, in other words one that belonged to a human but has since been taken over by a bot account to perform its malicious activities. These bots abuse the fact that trust has been built up with the human account and that they can use this trust to either spread a rumour, request money, or groom another individual. The detection model that they eventually proposed, used anomaly detection to highlight when a user's behaviour over time suddenly changes.

Work from Cresci et al. [80] evaluated past research to detect non-human accounts on SMPs. They used related work from both academia and commerce in their evaluation. The research result was a comparison across the various methods, which included rule sets as well as proposed attributes and features to detect non-human accounts on SMPs. An example of a rule set is that all accounts with more than 30 followers will be regarded as a non-human account. By using three public social media datasets, their experimental results showed that the following were the top three rules for identifying a non-human account on SMPs:

- An account having 30 followers
- An account having more than 30 tweets
- An account that has never tweeted another user directly

They found that attributes used to describe a user or his/her profile were easier (lower cost) to gather, yet as effective as other attributes (e.g. the content the user posts).

Varol et al. [317] define SMP bots to fall into three main categories. The first contains bots that can be active but have few followers. This category contains, for example, bots used to spam others. The second category consists of bots automated by applications. Bots can for example interact with their users and assist them with the answers to basic questions without getting a human involved [208]. The last category contains more sophisticated bots that can mimic human behaviour that is used to sway public opinion, for example. All of these categories are of importance to the research at hand, as most bots have to hide their identity [304] to achieve their intended purpose, just

like humans. In their study, Varol et al. [317] further extracted over 1 000 features from Twitter account holders. They used these features, together with supervised machine learning to detect the categories of bots mentioned before with an accuracy of 60% and higher. More interestingly, though, is that they grouped the features into the following sets to determine if one feature set was more important than another: user meta data; friends; content; sentiment; network; timing. Similar to the research work from Cresci et al. [80], Varol showed how user meta data attributes and features engineered from these attributes are by themselves well suited for detecting identity deception by bots.

All of the above bot research has been summarised in Table 4.1. The table shows which SMP attributes were used in each study with a mention of those specific attributes and engineered features that were found to be most indicative of identity deception by bots on SMPs. The table shows how research addressing specific threats, like that of Ferrara et al. [113], Xiao et al. [340], Fire et al. [115], and Egele et al. [106], only used data from specific SMP attribute categories. This was expected, as humans lie differently for different purposes [311]. Work from Cresci et al. [80] and Varol et al. [317] used all SMP attribute categories, as both were concerned with the detection of bots in general. Their research is of interest to the problem at hand, as it proposes a model to detect the identity deception of humans on SMPs in general. Both Cresci and Varol noted that SMP attributes describing the relationships and content of an account did not contribute much towards deception detection, therefore these SMP attributes will also be omitted from the research at hand.

Based on the work of Cresci et al. [80] and Varol et al. [317], Table 4.2 illustrates the attributes and engineered features that will be considered for the research at hand. The researcher found that the number of friends and number of followers can be combined into a friend-to-follower ratio as proposed by Stringhini, 2010 [287], Yang, 2013 [344], and Benevenuto, 2010 [39]. The researcher also omitted the ‘number of digits per screen’ proposed by Varol et al. [317], due to it being very specific to bots [340]. The researcher however proposes to include a new engineered feature to determine the URLs in the profile, as this seems to be an indicative factor of fake accounts as indicated in the research by Gilani, 2017 [128], Stringhini, 2010 [287], and Benevenuto, 2010 [39].

Table 4.1: Attributes and features used in related work to detect bot accounts on SMPs

Related research	SMP attribute category					Top attributes and features identified in research
	Describing the profile	Describing the account	Behaviour	Relationships	Content	
Gilani (2017) [128]		✓	✓	✓	✓	# of URLs in tweets # of retweets per tweet
Gurajala (2016) [142]	✓	✓		✓		created time updated time profile image # of friends # of followers
Stringhini (2010) [287]			✓	✓	✓	friend-to-follower ratio message with URL vs total messages message similarity likelihood to have picked friends from list # of messages # of friends
Li (2016) [204]	✓	✓	✓			# of likes time of like profile name
Yang (2013) [344]		✓		✓	✓	friend-to-follower ratio age of the account holder source of the content
Benevenuto (2010) [39]		✓		✓	✓	ratio of tweets with URLs Age of the user account Average number of URLs per tweet Fraction of followers per followees
Ferrara (2016) [113]	✓			✓	✓	# of posted tweets Ratio of retweets / tweets # of hashtags # of retweets

Table 4.1: Attributes and features used in related work to detect bot accounts on SMPs (continued)

Related research	SMP attribute category					Top attributes and features identified in research
	Describing the profile	Describing the account	Behaviour	Relationships	Content	
Xiao (2015) [340]	✓	✓				email address profile name IP address account open date
Fire (2014) [115]				✓	✓	tagged photos tagged videos are-family
Egele (2013) [106]		✓			✓	account open date message source, language and topic links in message direct user interaction similarity to friends
Cresci (2015) [80]	✓	✓	✓	✓	✓	has name, image, location, biography # of friends # of followers age
Varol (2017) [317]	✓	✓	✓	✓	✓	screen name length # of digits in screen name user name length time offset default profile default picture account age # of unique profile descriptions profile description lengths # of friends, followers, favourites, tweets, retweets, mentions, replies, retweeted

Table 4.2: Detecting deceptive humans, given attributes and features from bot research

Attribute / Engineering feature	Cresci (2015) [80]	Varol (2017) [317]	Considered for human identity deception detection?
Engineered feature	has name		yes
Engineered feature		name length	yes
Engineered feature		# of digits in screen name	no (very specific to bots [340]) but included # of URLs instead
Attribute	image	default picture	yes
Engineered feature	has location		yes
Attribute	biography	default profile	yes
Attribute	# of friends	# of friends	yes (combined with followers)
Attribute	# of followers	# of followers	yes (combined with friends)
Engineered feature	age	account age	yes
Attribute		Time offset	yes
Engineered feature		# of unique profile descriptions	yes
Engineered feature		Profile description lengths	yes
Attribute		# of favourites	no (not available for SMP user in this research)
Attribute		# of tweets	yes
Attribute		# of retweets	no (need content)
Attribute		# of mentions	no (need content)
Attribute		# of replies	no (need content)
Attribute		# retweeted	no (need content)

4.3 Identity deception in social sciences research

To gain further knowledge about engineering features indicative of identity deception by humans on SMPs, a second field was identified in which human deception is present. The field of social sciences is known for studying humans and their relationship to the world and the people around them [236]. This includes, among others, how deception – or more specifically, lies told by people – affect the world and the people around them.

4.3.1 Motivation for considering the social sciences and psychology in particular

The social sciences disciplines include but are not limited to the following: anthropology; archaeology; economics; human geography; jurisprudence; linguistics; political science; psychology; public health; and sociology [343] [237]. Social science is, in its broadest sense, the study of society. Social sciences propose to study the way people behave and influence the world around us. The social sciences can also help explain how our own society works – ranging from why people cannot find work or how economies grow, to what people buy, or what makes them happy [43]. Social sciences provide information to the government, for example for city planning, and to non-governmental organisations, for example for retail advertisement [78]. The field of social sciences also shows many examples of human identity deception [129].

In anthropology (the study of human behaviour in societies [236]) primates have been studied to understand what motivates deception, for instance [90]. These primates would pretend to be sick when they are not, to avoid doing a task they do not want to. This same behaviour applies to humans. In economics (the study of how humans deal with wealth [236]), people are prepared to lie about their identity when money rewards are offered [130]. In history, written materials are used to understand the behaviour of humans in the past. Here we can find many examples of identity deception. In war, humans are sometimes forced to assume a false identity, like in the case of the Trojan horse [334], to infiltrate the enemy. Jurisprudence, or legal theory [236], defines laws on how deception by humans should be dealt with [118]. In linguistics [154], various cues are proposed as indicators of deception. In the case of political science (the study of political activity and behaviour [236]) the effects of humans lying about their identity for example during political campaigns are investigated [230]. In psychology (the study of the human mind and its functions [236]), various issues are investigated, such as who

lies [177], why a human lie [94], what people lie about [147], and whether we can detect such deception [107]. In public health, deception during health trials is investigated [216]. Sociology is in turn concerned with the study of human society over time and ponders, for example, the effect of long-term deception on society [127].

From all these social science disciplines, linguistics and psychology stand out as those disciplines closest to the research at hand. Linguistics can be used to find cues to detect deception, whereas psychology identifies those attributes that humans are most prone to lie about. For the purposes of linguistics, content is required as part of the research [154]. For this research, content will not be considered to assist in the automated detection of identity deception by humans, as Cresci et al. [80] and Varol et al. [317] found that identity deception can be detected through user profile and account information with comparable accuracy. For these reasons, only the field of psychology will be considered further for the research problem at hand.

4.4 Identity deception in psychology research

Much of the past research on identity deception among humans has been psychological in nature and opposing views have been proposed on why humans lie [314]. For example, Halevy et al. [145] believe that most humans are honest most of the time, whereas DePaulo et al. [94] are adamant that most humans lie daily, to varying degrees, but mostly in small quantities. Ferrara et al. [113] believe that the act of deception is deliberate and intended to further a specific goal, such as to recruit other humans for terrorism. Deception harms trust [199].

According to Rubin [262], humans are not good at discovering deception, as they are biased towards the truth. Dando et al. [84] and Ekman et al. [108] believe the same and state that humans are unqualified to make judgements on the truth. Rong et al. [260] show that incentives such as reward schemes can result in lies becoming more prevalent. Leal et al. [196] propose that if humans were to declare their honesty up front, they will be less prone to lie. Whilst researchers will continue to debate about when and why humans lie, consensus remains that the act of lying is present.

4.4.1 Features used to detect identity deception in psychology

Past research in psychology was considered to identify those features about which humans are most likely to lie; more specifically features pertaining to their identity. It is assumed that human nature will prevail on SMPs and that humans will continue to lie, regardless of the medium of communication. Table 4.3 summarises the conclusions derived from related research by showing the various identity features humans lie about. Evidently, humans lie most often about their image, name, location, age, and gender.

From a psychopathological perspective, Stanton et al. [284] explored whether personality can explain deception such as changing one's name or image online. They found that feelings of inadequacy and self-dissatisfaction often lead to deception. Caspi and Gorsky [59] explored the emotions experienced during deception by using input from different demographics like location, age, gender, marital status, and occupation. They found that identity roleplay and privacy concerns were the main reasons for humans being deceptive. From an online perspective, Hancock [147] depicted identity-based and message-based deception as two main types of digital deception. He presented a detailed review on why and how humans lie and concluded that deception on online platforms could be more difficult to detect than face-to-face deception. Utz [311] defined the most common types of deception to be gender switching, identity concealment, and attractiveness deception. He also showed that these deceptive actions could be ascribed to different motivations.

Online dating deception has been the focus of attention of various researchers. For example, Toma et al. [301] investigated whether humans present themselves truthfully in their online dating profiles. They found that people deliberately deceive, and concluded that deception on certain identity features such as image, location, age, and gender, are more prevalent. Hancock and Toma [149] did similar research on online dating deception but focused on the images presented on these online dating profiles alone. They found that although users often present deceptive pictures, they try to remain authentic as far as possible. For example, users tend to present an image of their younger self.

Besides online deception, identity deception also occurs in other areas such as job interviews and criminology. Jupe et al. [174] investigated whether verifiable detail provided during a job interview could successfully distinguish humans telling the truth from those who lie. Wang et al. [324] considered past criminal records and compared the data provided by the criminals with the true data. The knowledge they gained

Table 4.3: Identity features humans generally lie about – according to psychology

	Image	Name	Location	Ethnicity	Age	Gender	Pseudonyms	Marital status	Occupation	Education	Activities	Interests	Appearances
Stanton (2016) [284]	✓	✓											
Jupe (2016) [174]									✓	✓			
Hancock (2007) [147]			✓	✓		✓	✓						
Donath (1999) [100]						✓							
Drouin (2016) [103]					✓	✓					✓	✓	✓
Caspi (2006) [59]			✓		✓	✓		✓	✓				
Utz (2005) [311]		✓			✓	✓	✓						✓
Tener (2015) [293]					✓								
Toma (2008) [301]	✓		✓		✓	✓							
Hancock (2009) [149]	✓												
Ho (2016) [160]						✓							
Wang (2006) [324]		✓	✓		✓		✓						
Al-garadi (2016) [7]													
Bergen (2014) [40]	✓				✓	✓							

offered a framework to indicate the identity features about which these criminals were most likely to lie. It was found that criminals most frequently lied about their name.

As is evident from the results presented in Table 4.3, humans in general lie most often about their image, name, location, age, and gender. These same lies are potentially present on SMPs as it has been established that humans lie on SMPs [103]. The researcher proposes the following engineered features to assist in the automated detection of identity deception by humans, specific to when they lie about their image, name, location, age, and gender:

- The age of the account and the age of the user can be used to determine the age of the user when the SMP account was registered. The legal age for opening a Twitter account, for example, is 13 [310]. Discrepancies found could indicate potential deception that warrants further investigation. The age of the user can be determined by using a technology like Google’s Vision Application Program Interface (API) [135]. This API extracts faces and their age from any given image by using Google’s own proprietary machine learning models for those accounts

that have images. Extracting the account opening date from the age of the profile image results in a number that represents the user's calculated age when they opened their account.

- For name deception, a mathematical distance formula such as the Levenshtein algorithm [201] [198] can compute the difference between a screen name and registered username. Since malicious users usually hide behind a pseudonym [311], it is expected that these users' screen name and registered username will not match.
- In some SMPs like Twitter, the geo-tag of the last tweet is stored for geo-enabled users [310]. Users also give their location in textual form and their time zone from a predefined list when they register an account on an SMP like Twitter [310]. The location and time zone can be updated again at any future time. By using a geo-location lookup technology like the ggmap library in R [176] to retrieve the geo location from the textual information and time zone, two additional features can be engineered – one comparing the geo-tag with the location and another comparing the location with the time zone. The Haversine distance formula [316] [277] can be used to determine the distance between the two points. It is expected that this distance should be close for those users telling the truth.
- The Google Vision API [135] can also be used to determine whether the face shown on a profile image is male or female. Combining this knowledge with a name database (of male and female names) creates a feature that can compare the gender of an image to the gender of a name. It is expected that these should match for users who are telling the truth.

For the purposes of the research at hand, these proposed engineered features will be combined with the attributes and features identified in the related bot research discussed earlier.

4.5 Proposed attributes and features to detect identity deception by humans on SMPs

Table 4.4 shows all the combined attributes and features from related research work in the fields of bots and psychology that will be useful to solve the research problem at hand and to assist in the automated detection of identity deception by humans on

Table 4.4: Attributes and features to detect identity deception by humans on SMPs

Attributes and features	Origin	Constructed from these SMP attributes
ACCOUNT_AGE_IN_MONTHS	Bot	created_at
AGE	Psychology	created_at, profile_image
GENDER	Psychology	name, profile_image
DISTANCE_LOCATION	Psychology	location, latitude, longitude
DISTANCE_TZ	Psychology	location, time_zone
DUP_PROFILE	Bot	description
FOLLOWERS_COUNT	Bot	followers_count
FRIENDS_COUNT	Bot	friends_count
FRIENDS_VS_FOLLOWERS	Bot	friends_count, followers_counts
GEO_ENABLED	Bot	geo_enabled
HAS_IMAGE	Bot	profile_image
HAS_NAME	Bot	Name
HAS_PROFILE	Bot	description
LISTED_COUNT	Bot	listed_count
NAME	Psychology	name, screen_name
NAME_LENGTH	Bot	screen_name
PROFILE_HAS_URL	Bot	description
TWEET_COUNT	Bot	status_count

SMPs.

Next follows a brief description of each attribute or engineered feature:

- ACCOUNT_AGE_IN_MONTHS – the number of months since the account was opened.
- AGE – the age of the user when they registered their SMP account.
- GENDER – this value will be true if the person’s name matches the gender detected from the profile image, otherwise this value will be false.
- DISTANCE_LOCATION – the distance in km between the geo-tagged location and the location stated by the user.
- DISTANCE_TZ – the distance in km between the location and the time zone stated by the user.
- DUP_PROFILE – whether the current account has a similar profile description as another user in the corpus.

- `FRIENDS_VS_FOLLOWERS` – the ratio of friends vs followers.
- `FOLLOWERS_COUNT` – the number of followers recorded for a user.
- `FRIENDS_COUNT` – the number of friends recorded for a user.
- `GEO_ENABLED` – a value indicating whether an account is enabled to store its location in terms of longitude and latitude.
- `HAS_IMAGE` – shows whether a profile image has been defined for an account or whether the account is still using the default SMP image as its profile (feature is constructed as a binary indicator).
- `HAS_NAME` – shows whether the name could be found in a name database (feature is constructed as a binary indicator).
- `HAS_PROFILE` – shows whether the account has a description or not (feature is constructed as a binary indicator).
- `NAME` – the Levenshtein distance [201] [198] between the screen name and registered username.
- `LISTED_COUNT` – the number of public lists the account belongs to is recorded.
- `PROFILE_HAS_URL` – shows whether the account’s description contains an URL or not (feature is constructed as a binary indicator).
- `TWEET_COUNT` – the number of tweets posted by the account.
- `NAME_LENGTH` – the number of characters contained in the screen name or pseudonym of the account.

4.6 Conclusion

Literature dealing with deception is currently available from a wide range of research fields. Most notably, the fields of bot detection and psychology lend themselves to the research at hand, as both fields deal with identity deception.

In the field of bot detection, research has focused on finding those fake accounts on SMPs that originate from non-human accounts. Such research has made use of the various attributes available on SMPs to achieve this goal. Researchers who propose how to find bot accounts – regardless of whether they are a threat to humans – agree that those attributes and engineered features describing a user profile and account on SMPs are

sufficient to detect the fake accounts. Based on this knowledge, the researcher presented a list of attributes and engineered features to complement a model that proposes to assist in the automated detection of identity deception by humans on SMPs. Additional engineered features from the field of psychology were added to this list, in order to increase the accuracy of the proposed model. It became evident that humans were prone to lie most about their image, name, location, age, and gender. The researcher showed how features could be engineered for each of these factors, using the attributes available on SMPs.

The next three chapters will describe a research environment in which to implement such a model by using the identified attributes and engineered features from this chapter as input.

Part II

Research design

Chapter 5

Steps to assist in the automated detection of identity deception on SMPs

“Research is formalized curiosity. It is poking and prying with a purpose.”
-Zora Neale Hurston

5.1 Introduction

The previous four chapters discussed big data, Social Media Platforms (SMPs), the cyber threats found on SMPs, and identity deception as a cyber threat on SMPs. It was shown how various attributes are available on SMPs that define an identity. Attributes describing a user profile or account were found to be sufficient to detect identity deception by bots on SMPs, with the added benefit of these attributes being more easily gathered than (for example) the content a user posts or the relationships they have with other users.

Additional identity features found in the field of psychological research showed that humans lie most about their image, name, location, age, and gender [284] [147] [324]. The researcher proposed that these features, combined with the attributes and features identified in bot-related research, be engineered to aid in the automated detection of identity deception. Furthermore, various methods (such as supervised machine learning) have been proposed in related research work as a method to detect identity deception. The requirements for a model to assist in the automated detection of identity deception

were subsequently defined.

To implement such a model, this chapter firstly introduces a quantitative research approach that involves various steps – better known as the research design [189]. This research design not only presents the approach taken to implement the previously defined requirements, but also provides a controlled experimental environment in which the various identified attributes and features towards identity deception detection can be accurately assessed. Each step followed in the research approach (aimed at describing a model that assists in the detection of identity deception by humans on SMPs) will be discussed in detail in the remainder of this chapter. The chapter culminates in providing a high-level design that presents a blueprint for the components required to implement a prototype of this model.

5.2 Requirements for the automated detection of identity deception by humans on SMPs

Various requirements of a model that can be expected to assist in the automated detection of identity deception have been proposed in Chapter 4. To define the make-up or components of such a model, these requirements should be taken into consideration. The researcher considered various research approaches to gain an understanding of the components that are required, and to ensure that these components will meet the requirements. A research approach is a methodical process, in the form of steps, that is used to solve a research problem at hand [189]. In the present case, the problem is to detect human identity deception on SMPs. The research approach will be discussed next in an effort to know which steps are to be taken and how these steps cater for the requirements.

5.3 The research approach

Research can in general be classified as following one or more approaches [189]. It is important to understand the type of approach as this can help to understand what steps will be used during the research at hand. Kothari et al. [189] defined research approaches as follows:

- Descriptive and/or Analytical – With descriptive research, only the current state

is described. Analytical research on the other hand will not only gather data but analyse the data further in an attempt to make assumptions or further suggestions. The research at hand is considered analytical, as data will be gathered to detect identity deception by humans on SMPs.

- Applied and/or Fundamental – Applied research can be used in everyday life to solve immediate problems, whereas fundamental research is more theoretical in nature. The research at hand is of the applied type, as it proposed to solve the risk of identity deception by humans found on SMPs.
- Conceptual and/or Empirical – Conceptual research is based on made-up numbers, whereas empirical research is based on true facts that were observed and can be confirmed. Empirical research sometimes includes samples from a population. The research at hand is empirical in nature, as samples from Twitter (an example of an SMP) were gathered.
- Quantitative and/or Qualitative – Quantitative research is based on actual numbers, whereas qualitative research is based on measuring quality of some kind that can potentially not be expressed in numbers. The research at hand can be said to be quantitative, as it proposes to build a measurable model that can detect identity deception by humans on SMPs.

Many researchers have defined their research approach to be either quantitative and/or qualitative [333] [145]. According to Kothari et al. [189] these are the two most common approaches found in research. For this reason, and the fact that a measurable model will be proposed to solve the problem at hand, this research will be approached as quantitative. In addition to being quantitative, the researcher will execute various experiments to find the most suitable model for identity deception by humans on SMPs. The same experimental research design has been used in the field of psychology [105] and bots [219].

5.4 Steps towards detecting identity deception by humans on SMPs

An experimental research design has various steps which include preparing the data, experimenting with the data by using various methods, and developing a model that can assist in the automated detection of human identity deception on SMPs based on the

results obtained.

The next section describes these steps in more detail.

5.4.1 Preparing the data

The preparation of data for the experiments concerned involves gathering, cleaning and labelling the data (a requirement for supervised learning), as well as further preparation steps that are specific to machine learning. The main output from this preparation step is to provide the data in a format that is easy to experiment with.

5.4.1.1 Gather the data

In order to conduct this research, the researcher required data to experiment with. There are various ways to gather data in general, for example, by observations [10]; interviews [168]; questionnaires [345]; computation [10]; scraping [1]; or data mining [271]. Since the identified research problem is specific to SMPs, data from an SMP was required. Mining data from SMPs is quite common nowadays, and various Application Program Interfaces (APIs) are available to do so [109] [134] [206] [310] [248] [165]. For the current research experiment, the Twitter API [310] was used to gather data. This SMP was chosen for the following reasons:

- Out of the top six social media platforms mentioned earlier, Twitter is the only platform where no consent is required to gather their data. With Twitter, it is possible to gather data from anyone without their consent, whereas Facebook (for example) follows an approach where a friend must first accept your request before you are able to gain access to their data.
- Twitter was identified as an SMP that is used in many research papers across various disciplines [54] [333] [142].
- The data gathered from the Twitter API [310] is deemed sufficient for research [228].

There are three different ways to gather data from Twitter: Twitter's Search API, Twitter's Streaming API, and Twitter's Firehose [38] [228]. The Search API retrieves historic tweets, whereas the Streaming API and Twitter's Firehose capture tweets in real time. A single request to the Twitter API can return up to 3 200 tweets. The Search

API and Streaming API are however limited to 180 requests per 15-minute interval [310]. The Streaming API is also capped at 1% of the total real-time tweet stream [228]. This means that only 1% of the data being requested at a given time will be returned by the Streaming API. Furthermore, Twitter's Firehose service is a paid service with no rate limits imposed. Since the end of 2017, Twitter has changed its Search and Streaming APIs to return data for different time periods and content that is dependent on a standard, premium, or enterprise subscription basis. The data for this research was gathered before the imposed subscription restrictions came into effect.

The sheer volume of data on Twitter and the rate limits imposed have made mining all account data since Twitter's inception in 2006 unfeasible and impractical for the current research [314]. For the purposes of the corpus, the researcher chose to limit the data to a demographic known to be the target of deceptive users. Minors are susceptible to cyber bullying [122], extremist recruitment [186], and grooming [182], among others. The corpus was limited to accounts that used the words 'school' and 'homework' through Twitter's Streaming API, as these are words used widely by minors [269]. The friends and followers of these accounts were also gathered using the Twitter Search API, because it is known that friends usually have similar friends [75] – in this case, more minors. The result is a corpus of data, obtained from an SMP, that contains the profiles and content of SMP account holders.

5.4.1.2 Clean the data

Since the current research is focused on addressing deception by human users, an attempt was made to rid the corpus of non-human accounts included in the initial gathered corpus. The originally gathered corpus included content from bots and humans, since humans either followed these bot accounts, or bot account themselves tweeted about 'school' or 'homework'.

Much research has been done to detect bot accounts [76] [256] [142], and in most cases the research either proposes rules or machine learning to detect these types of accounts. To this end, research presented by Cresci et al. [80], and described in Chapter 3, was found most promising. Their research evaluated related work for three different rule-based approaches that were able to detect bots. In the current study, the researcher combined the related work and then applied the top three rules (which had an accuracy of over 75% in detecting bots) to clean the corpus. The reason for using the rule-based approach was its simplicity and accuracy (above 75%), which was close to the accuracy achieved by

machine learning approaches [219] [113]. The rules concerned were the following:

- The account must have more than or equal to 30 followers.
- The account must have more than or equal to 50 tweets.
- The account must have replied to at least one direct tweet from another user.

Besides cleaning the corpus from bots, Twitter verified the authenticity of some accounts. The users initiated and requested the verification themselves [310]. These accounts would not pose a threat in respect of identity deception and were therefore removed.

Finally, certain attributes were removed from the corpus due to the following reasons:

- An attribute that has a strong correlation with another. Such an attribute would not have any added benefit to the detection of identity deception [349].
- An attribute that will introduce variance, if included. Variance is the tendency to learn random things unrelated to the problem [99]. An example would be if the corpus includes data unrelated to the problem being solved, for example the Identifier (ID) assigned to an account by Twitter [314].
- An attribute that will introduce bias, if included. Bias is the tendency to learn the same wrong thing [99]. An example would be if the background image attribute was empty for the gathered corpus and therefore assumed to be always empty – which is not the case [314].

Variance and bias are collectively referred to as overfitting [345]. Figure 5.1 illustrates the relationship between variance and bias based on dart board results as an example [99]. The centre of the dart board represents the correct detection of identity deception. Each X represents how near or far the final result was from the correct detection, per account. For example, with high bias and low variance, the wrong result is given consistently. With low bias and low variance, the correct result is consistently given. To detect potential identity deception, the aim should be to achieve low bias and low variance.

Based on this explanation of bias and variance, the following attributes in Twitter could be removed and the reason for removal is shown in brackets:

- Where the attributes were unique to a specific account – for example, the ID, name, and account description (variance).
- All remaining zero variance attributes were removed, in other words data with a remarkably high ratio of uniqueness – for example, longitude, latitude, and location

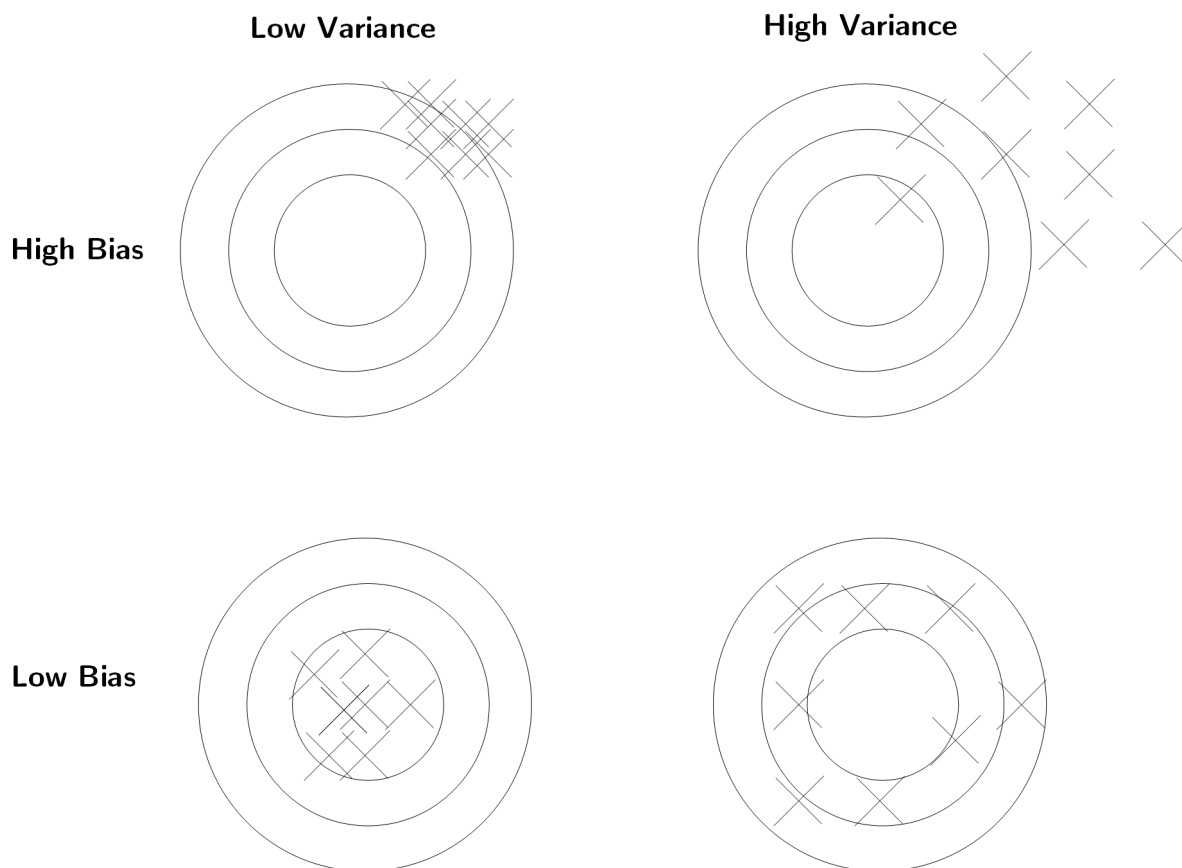


Figure 5.1: An explanation of variance vs bias [99]

(variance).

- Data that is mostly empty in the corpus – for example, background image and background colour (bias).

5.4.1.3 Label the data

Supervised machine learning requires a labelled dataset for model development purposes [190]. For this reason, known deceptive accounts were appended to the original corpus. Finding examples of deceptive human accounts does, however, pose a challenge. Zafarani and Liu [350] suggest manual crowd-sourcing mechanisms, like Amazon’s Mechanical Turk [13], to label accounts and classify a predicted outcome. There are also known means for user groups to oust malicious accounts, such as accounts linked to terrorism [113]. Peddinti et al. [242] used labelled datasets from their own previous research work to identify sensitive accounts referring to, for example, topics about pregnancy and paedophilia. None of these were viable options to be used in the current research, due to the absence of known deceptive human accounts

and expertise required to manually identify deceptive human accounts correctly as a group. As it would be impossible to collect confirmed deceptive accounts in the real world on the Twitter SMP, the researcher was compelled to fabricate deceptive accounts [314]. Further arguments for appending deceptive accounts were based on

- ethical reasons to respect the privacy requirements of Twitter, and
- ethical research policies that limit what can be reported on regarding actual social media users and the perception of their deceptiveness.

It was expected that there might be some deceptive accounts in the gathered corpus but the researcher believed the effect would be negligible. Most humans have been shown to be honest most of the time [145]. Over 15 000 known deceptive accounts were appended and represented almost 10% of the cleaned corpus. Halevy et al. [145] found that 5% of people tell 40% of all lies. With the introduction of 10% fabricated deceptive accounts, it was assumed that most lies would be catered for.

The fabricated deceptive accounts were generated using two random human data generator APIs from the internet [25] [178]. Section ?? discussed how humans lie most often about their image, name, location, age, and gender. Therefore, the researcher confirmed that each injected deceptive account was deceptive in respect of each of these attributes by applying rules to test for deceptiveness. An example was to ensure that the name used for an account and its pseudonym was never the same. Another was that the age detected in the user's image was different from their actual age. By ensuring that all attributes, as per psychological identity deception research results, were deceptive, each account was created to be as deceptive as possible – even though humans might lie only about some of these attributes in the real world.

Further manual intervention was required to complete the remaining attributes which the APIs could not do. An example of manual data was the number of friends and followers of an account. Values were chosen such that they were similar to what was observed within the bounds of the current gathered corpus. The deceptive accounts were classified as 'deceptive' and the original corpus as 'trustworthy'.

In the absence of deceptive human accounts and for the sake of the validity of the research, it was decided to align the fabricated deceptive accounts as far as possible with the data contained in the original corpus, to make the research results as realistic as possible. Moreover, the following two statistical tests were employed to validate that the appended deceptive accounts were still representative of the original gathered

corpus [314]:

- Wilcoxon Signed Rank test, also known as the Mann-Whitney-U test [189].

This test compares the sum of ranks, or indirectly the medians, of two sets of distributions. If the means are similar, the data can be assumed to be from the same population. This test does not require data to be normally distributed or sample sizes to be the same [213]. The test does however require the sets of distributions to be independent [213]. For example, evaluate the distribution of one attribute, like ‘number_of_friends’ for both the ‘deceptive’ and ‘trustworthy’ corpus. If both distributions are found to be similar, they are believed to represent similar data. If all attributes individually pass the Wilcoxon test, both datasets can be said to be from the same population, which in this case was Twitter.

- Pearson’s chi-square test of independence [189].

This test assumes that subjects in a single population are classified similarly. It shows when attributes in the population are correlated, and thus from the same population. The test works well when samples were generated at random, the attributes were categorical, and the resultant categories were greater than 5 [218]. For example, evaluate the correlation between one attribute, such as ‘number_of_friends’, for both the ‘deceptive’ and ‘trustworthy’ corpus. If the attribute is found to be highly correlated, it can be said that it contains similar data. If all attributes succeed in the Pearson’s Chi square test, it can be said that both datasets are from the same population, in this case Twitter.

For both tests, the level of significance was set at 5% – a commonly used practice in hypothesis testing of similar research [33] [177] [189].

In the current research, the appended deceptive accounts were not only representative of data found in Twitter, they also had to be actually deceptive. For this reason, the researcher analysed past research from the field of psychology by highlighting identity attributes humans are known to lie about. It was shown in Chapter 4 that humans lie most often about their image, name, location, age, and gender, and by applying rules to test for deceptiveness, the study confirmed that each deceptive account was deceptive in respect of each of these attributes. An example was to ensure that the names used for an account and its pseudonym were never the same, and also that the age detected in the user’s image was different from their actual age. By ensuring that all attributes, as per psychological identity deception research results, were deceptive, each account was created to be as deceptive as possible – despite the fact that humans might lie only about some of these attributes in the real world [314].



Figure 5.2: Example of results returned from Google Face API [135]

5.4.1.4 Engineer additional features

Additional features were engineered and added to the corpus as part of the data preparation. These attributes and features stemmed from related work in the fields of bots and psychology. The features were engineered from the combination of existing SMP attributes, as per Table 4.4, or by making use of additional external data or mathematical methods.

External data was retrieved using Google’s Face API [135], the OpenStreetMap API [234], and an external names database [221]. For the Google Face API [135], the profile image of the account was sent to the API through a Representational State Transfer (REST) call requesting the age, gender, head pose, as well as facial attributes such as smile, facial hair, and glasses. An example of the returned data from the Google Face API is illustrated in Figure 5.2. For the OpenStreetMap API [234], the location was also sent via a REST call. An example of the returned data is illustrated in Figure 5.3. Finally, an external names database [221], containing over 40 000 first names and their expected gender, was imported. An example of the available data is shown in Figure 5.4 where ‘F’ denotes Female, ‘M’ denotes Male, and ? indicates that the name could be either Male or Female.

In addition to the external data imported to prepare the data, various mathematical methods were used to construct new features. The Haversine mathematical formula [259], for example, calculates the distance between two pairs of longitude and latitude coordinates and takes the curvature of the earth into consideration. The formula has been used before to calculate how far tweets travel in Twitter [316] and the distance between astronomical objects [277]. The haversine formula is defined as [259]:

Requests

```
https://nominatim.openstreetmap.org/search?q=135+pilkington+avenue,+birmingham&format=xml&polygon=1&addressdetails=1#
https://nominatim.openstreetmap.org/search/135%20pilkington%20avenue,%20birmingham?format=xml&polygon=1&addressdetails=1#
https://nominatim.openstreetmap.org/search/gb/birmingham/pilkington%20avenue/135?format=xml&polygon=1&addressdetails=1#
```

```
<searchresults timestamp="Sat, 07 Nov 09 14:42:10 +0000" querystring="135 pilkington, avenue birmingham" polygon="true">
  <place
    place_id="1620612" osm_type="node" osm_id="452010817"
    boundingbox="52.548641204834,52.5488433837891,-1.81612110137939,-1.81592094898224"
    polygonpoints="[[[-1.81592098644987','52.5487429714954'],[-1.81592290792183','52.5487234624632'],...]]"
    lat="52.5487429714954" lon="-1.81602098644987"
    display_name="135, Pilkington Avenue, Wylde Green, City of Birmingham, West Midlands (county), B72, United Kingdom"
    class="place" type="house">
    <house_number>135</house_number>
    <road>Pilkington Avenue</road>
    <village>Wylde Green</village>
    <town>Sutton Coldfield</town>
    <city>City of Birmingham</city>
    <county>West Midlands (county)</county>
    <postcode>B72</postcode>
    <country>United Kingdom</country>
    <country_code>gb</country_code>
  </place>
</searchresults>
```

Figure 5.3: Example of results returned from OpenStreetMaps [234]

$$hav\left(\frac{d}{r}\right) = hav(\Phi_1 - \Phi_2) + \cos(\Phi_1)\cos(\Phi_2)hav(\lambda_1 - \lambda_2)$$

where:

d is the distance in kilometres between the two points,

r is the radius of the earth (6 371 kilometres),

Φ_1, Φ_2 : latitude of point 1 and latitude of point 2, in radians

λ_1, λ_2 : longitude of point 1 and longitude of point 2, in radians

where:

$$\text{radians} = \text{degrees} * PI/180$$

The Levenshtein distance mathematical formula [198] was also used in this research to calculate the difference between two names. The distance is determined by counting how many deletions, insertions, or substitutions it would take to make one string look like another. The Levenshtein distance formula has for instance been used in previous research to determine the distance between names used on criminal records [325] and the difference between identities in general, including their social behaviour [201]. The Levenshtein distance between two strings, a and b , with lengths, i and j respectively, is defined as follows:

$$D(i, j) = \min\left\{D(i-1, j-1) + \gamma(A\langle i \rangle \rightarrow B\langle j \rangle),\right. \\ \left. D(i-1, j) + \gamma(A\langle i \rangle \rightarrow \Lambda),\right. \\ \left. D(i, j-1) + \gamma(\Lambda \rightarrow B\langle j \rangle)\right\}$$

There are many alternatives to distance calculations, like the Hamming distance [146]

?	Lou
M	Louai
F	Louann
M	Louay
F	Loubna
F	Louella
M	Louie
M	Louis
F	Louisa
F	Louise
F	Louiselle
F	Louisette
F	Louisiane
?	Louison
F	Louiza
M	Louk
F	Louka
M	Loukas

Figure 5.4: Example of the results from the external names database [221]

and Jaccard similarity index [166]. For this research, the Levenshtein difference was used due to its previous success in detecting the difference between identities [325] [201].

Knowing the external data and mathematical formulas used by the researcher to prepare the data, the construction of each proposed new feature is described in Table 5.1.

5.4.1.5 Convert data for machine learning

Before any machine learning models for identity deception can be developed, the data must be in the correct format. Most machine learning models expect data to be discretised, centred, and scaled [191]. Discretisation implies that numerical data is converted to categorical data. An example is if the number of friends is grouped into bins of 500. The result would be accounts falling into the ranges of 0–500, 501–1000, and so on. All nominal values are then centred. For centring, the sample mean is subtracted. For example, if the mean of the ‘number_of_friends’ is 1 500 for the total corpus, this value will be subtracted from each account for their respective ‘number_of_friends’. Lastly, these centred values are divided by the standard deviation. This ensures that all input is similarly scaled and will not introduce bias if the values for other inputs are higher. An example is where the number of friends is on a different scale initially from the number of tweets or posts. The proposed scaling method ensures that machine learning models will treat both inputs as equally

Table 5.1: Additional features engineered and added to the corpus

Feature name	Origin	Engineered feature constructed as follow
ACCOUNT_AGE_IN_MONTHS	Bot	The number of months an account has been open can be calculated by extracting the current date from the account-opening date.
COMPARE_AGE	Psychology	The age of a user is extracted from their profile image via the Google Face API [135]. By subtracting the number of months an account has been open from the user's age, a new feature indicating the age of the user when they opened the Twitter account can be engineered.
GENDER	Psychology	The gender of the user is retrieved from the profile image via the Google Face API [135]. In addition, the user's gender is extracted via an external names database [221], and comparing this information indicates whether the extracted genders match or not.
DISTANCE_LOCATION	Psychology	Besides Twitter tracking the geo-location of a user, a user indicates the location in textual form in Twitter. The geo-location of the textual location is retrieved via the OpenStreetMaps API [234]. The Haversine mathematical formula [259] is applied to calculate the distance between the two geo-locations.
DISTANCE_TZ	Psychology	Besides Twitter tracking the time zone of a user, a user indicates the location in textual form in Twitter. The geo-location of both the time zone and the textual location is retrieved via the OpenStreetMaps API [234]. The Haversine mathematical formula [259] is applied to calculate the distance between the two geo-locations.
DUP_PROFILE	Bot	The profile of a user is compared with the profiles of other users to determine if it is a duplicate.
FOLLOWERS_COUNT	Bot	This value is available in Twitter as an SMP attribute.
FRIENDS_COUNT	Bot	This value is available in Twitter as an SMP attribute.
FRIENDS_VS_FOLLOWERS	Bot	Calculated as the ratio of friends to followers.
GEO_ENABLED	Bot	A boolean value indicating whether Twitter is tracking the geo-location of a user.
HAS_IMAGE	Bot	A boolean value indicating whether the Twitter user has provided a profile image other than the default.
HAS_NAME	Bot	A boolean value indicating whether the Twitter user has provided a name.
HAS_PROFILE	Bot	A boolean value indicating whether the Twitter user has provided a profile description.
LISTED_COUNT	Bot	This value is available in Twitter as an SMP attribute.
LEVENSHTEIN	Psychology	The Levenshtein difference between the user's name and display name provided.
NAME_LENGTH	Bot	The number of characters in the name provided by the user.
PROFILE_HAS_URL	Bot	A boolean value indicating whether the Twitter user has provided a URL in their profile description.
STATUS_COUNT	Bot	This value is available in Twitter as an SMP attribute. It indicates the number of Tweets a user has sent.

important [314].

The finally prepared data is used next in various experiments – including supervised machine learning to assist in developing a model that can detect identity deception by humans on SMPs.

5.4.2 Experimenting with the data

The identification of identity deception can be classified as a binary classification problem, since all data defined is described by one of two classes: ‘trustworthy’ or ‘deceptive’. The ‘deceptive’ accounts will be those that are manually generated. Due to the lack of machine learning to detect identity deception in social media by humans, it is deemed that examples of bot and spam detection are close to the research at hand. Bot and spam detection resembles a similar problem, which results in a binary answer. The researcher acknowledges that should a labelled dataset consisting of a ‘trustworthy’ and ‘deceptive’ set of accounts not have been available, unsupervised machine learning algorithms could have been explored. For the purpose of this research and as stated in Section 3.4.3, only supervised machine learning algorithms were considered. Human identity deception is usually unique [312] and clusters might not be as prevalent as expected by unsupervised machine learning algorithms. Table 3.4 showed the various supervised machine learning algorithms proposed by the research for spam and bot detection. Bases on these results, the algorithms most used were chosen for the research at hand.

Table 5.2 provides an overview of the supervised machine learning algorithms that were used in this research approach. The eight algorithms that were covered in this research comprised all currently known classification techniques, namely logic, instance, perceptron, statistical and vector based. [190]. Logic-based algorithms, such as decision trees, use rules to develop a final model, while instance-based algorithms, like clustering, depend on the proximity of other instances in the training data set. Perceptron-based algorithms, like neural networks, have various inputs that are changed by weights to produce an outcome that can be used in subsequent calculations. Statistical-based algorithms, for instance Bayesian types, are based on an underlying probability model. Lastly, vector-based algorithms, like support vector machines, map data on a higher dimension to determine boundaries for classification.

Next follows a short description of each chosen algorithm.

- Adaptive boosting – A type of ‘Ensemble Learning’ algorithm in which multiple learners are employed to build a stronger learning algorithm. The algorithm starts with a base algorithm, for example decision trees, and then iteratively improves its model by accounting for the incorrectly classified examples in the training set [117].
- Bayesian generalised linear algorithm – This is a flexible algorithm based on normal linear regression where the error distribution does not have to be normal [53].

Table 5.2: Supervised machine learning algorithms used in this research

Machine learning algorithm description	R Caret library name [191]	Algorithm Family	Classification technique
Adaptive boosting	Adaboost	Boosting	Combination
Bayesian generalised linear	bayesglm	Linear	Statistical-based
J48 library from Weka	J48	Tree	Logic-based
K Nearest Means	kknn	Clustering	Instance-based
Neural Network	nnet	Neural Network	Perceptron-based
Random Forest	rf	Tree	Logic-based
Recursive partitioning tree	rpart	Tree	Logic-based
SVM with Radial Basis Function Kernel	svmRadial	SVM	Vector-based

- J48 library from Weka – This is a variant of the C4.5 decision tree that can handle continuous and discrete variables and that defines *inter alia* weights to different features [241].
- K Nearest Means – A simple clustering algorithm [122]. Voting, based on k neighbours closest to the current node, determines the class of the current node.
- Neural Network – Simulates the neurons in the brain. The algorithm represents a set of nodes, organised in layers, interconnected, and the input of nodes is dependent on the outputs of others [353].
- Random Forest — Creates an ensemble of decision trees [210]. The random forest algorithm builds many variations of a decision tree by using different combinations of input and parameters. These decision trees are compared with one another in an effort to determine the most optimal solution.
- Recursive partitioning tree – A variant of a decision tree. The dataset is split recursively based on the largest possible reduction in heterogeneity of the predicted variable [296].
- SVM with Radial Basis Function Kernel – A form of linear regression [57]. In its simplest form, it derives a hyperplane that maximises the separating margin between the classes [5].

The selected attributes and engineered features are given as input to the machine learning algorithm to develop the best model with which to detect identity deception by humans on SMPs. This process is also referred to as ‘training’ [190]. Before developing any machine learning algorithm, the skewness of the dataset is considered [157]. Data has a skewed distribution when one class is in minority. This is the case for this research, as only a small number of deceptive accounts were generated.

Various means exist to handle such scenarios of skewed data, among others: under-sampling; over-sampling; Synthetic Minority Oversample Technique (SMOTE); and Random Over-Sampling Examples (ROSE) [339] [83]. With under-sampling, the majority class is reduced to the size of the minority. With over-sampling, the minority class is duplicated to produce as many samples of data as the majority. SMOTE and ROSE are both forms of over-sampling but instead of duplicating the minority data, new values are generated. With SMOTE, the new values are generated choosing random points between the two classes, whereas with ROSE these points can come from outside the current scope of the classes.

The over-sampling method, SMOTE, was chosen for the research at hand due to the following reasons:

- Under-sampling would have reduced the corpus significantly, as the trustworthy corpus would have been reduced to the size of the deceptive accounts. By discarding known data, variance could be introduced in the algorithms.
- SMOTE is known to randomly generate data based on the distribution of existing data and to produce more realistic values than ROSE. This is better than normal over-sampling where the deceptive accounts would just be repeated until an overall equal class distribution was achieved. It also reduces chances of bias in the model.

In addition, decisions were made whilst developing these machine learning models. Each algorithm was developed by applying 10-fold, 3-repeat cross validation resampling as suggested in similar research [16] [187]. This means that the data was divided into ten equal sets where nine sets were used to develop the model and one was used to test the accuracy of the model. This process was repeated three times with the average accuracy taken as the final result for the developed model. Cross validation ensured that the model was not developed using only one class, as various samples from the dataset were tested over many iterations.

Each machine learning model also has its own parameters, known as hyperparameters [28], as shown in Table 5.3, which can be adjusted to optimise the development of the model [28]. The best parameters were chosen, using a grid search or brute force approach taking the best precision-recall result as final answer [191] [113] [340] [115]. For this research, grid search was preferred due to the fact that the R Caret library [191] provides default hyperparameters that generally are the best for each machine algorithm [191]. The precision-recall curve was preferred to the Receiver Operator Characteristic (ROC) as research has shown that ROC curves

Table 5.3: Hyperparameters per supervised machine learning algorithm

Machine learning algorithm description	R Caret library [191]	Hyperparameter	Hyperparameter description
Adaptive boosting	Adaboost	nIter (#Trees)	The maximum number of iterations after which boosting is terminated. If a good model is found sooner, the boosting will terminate earlier.
		method (Method)	Real AdaBoost returns a probability of class membership, whereas Adaboost.M1 is an extension of multiclass classifications.
Bayesian generalised linear	bayesglm	-	-
J48 library from Weka	J48	C (Confidence Threshold)	At what point should the tree be pruned.
		M (Minimum Instances Per Leaf)	Minimum number of leaves in the tree.
K Nearest Means	kknn	kmax (Max. #Neighbors)	Max number of k
		distance (Distance)	Parameter of Minkowski distance
		kernel (Kernel)	The kernel function is used to weight the neighbours according to their distances.
Neural Network	nnet	size (#Hidden Items)	The number of items in the hidden layer.
		decay (Weight Decay)	The regularisation parameter to avoid over-fitting.
Random Forest	rf	mtry (#Randomly Selected Predictors)	The number of entries randomly chosen at each decision point to determine best split.
Recursive partitioning tree	rpart	cp (Complexity Parameter)	It is the amount by which splitting that node in the tree improves the relative error. If it does not improve, the splitting will stop.
SVM with Radial Basis Function Kernel	svmRadial	C (Cost)	It is a hyperparameter that controls how much incorrect classifications are penalised.

could hide bad performance in highly skewed distributions [264].

Algorithm 1 shows the pseudocode when applying all these concepts to develop a machine learning model.

Once a machine learning algorithm has been developed, it is referred to as a machine learning model. In the next section, the performance of the features need to be measured after the model was used to detect identity deception. This is done to understand which features were used more in the model and thus contributed more towards the detection of identity deception.

5.4.2.1 Presenting the machine learning results

A confusion matrix that was used to measure binary classification machine learning models [80] [307] [22] is depicted in Table 5.4. The confusion matrix has four categories [170]:

- True positives (TP) – examples correctly labelled as deceptive
- False positives (FP) – examples incorrectly labelled as deceptive
- True negatives (TN) – trustworthy accounts correctly labelled as trustworthy

Algorithm 1 Training a machine learning model

Oversample the data using SMOTE
 Split the data into 75% training data and 25% test data
 Use the 75% training data to:

for Each hyper parameter **do**
 for Each repeat (3 times) **do**
 for Each fold (10 times) **do**
 Break the data into 10 folds
 Train the algorithm on 9 (10-1) folds' data
 Test the result on the remaining fold
 Keep record of accuracy as per precision-recall curve
 end for
 Calculate average accuracy for all folds run
 end for
 Save the model producing the best accuracy, tested against 25% test data
end for

Use the model with the best accuracy, tested against 25% test data

- False negatives (FN) – deceptive examples incorrectly labelled as trustworthy

From these categories, various performance metrics were derived:

- Precision – the ratio of correctly predicted deceptive cases [81]

$$P = \frac{TP}{TP + FP}$$

- Recall – the ratio of deceptive cases that were indeed predicted as deceptive [81]

$$R = \frac{TP}{TP + FN}$$

- Sensitivity – the same as recall [81]

$$Sens = \frac{TP}{TP + FN}$$

- Specificity – the ratio of trustworthy cases that were indeed predicted as trustworthy [81]

$$Spec = \frac{TN}{FP + TN}$$

Table 5.4: Theoretical depiction of a confusion matrix

		Predicted	
		Deceptive	Trustworthy
Observed	Deceptive	True Positive (TP)	False Negative (FN)
	Trustworthy	False Positive (FP)	True Negative (TN)

- Accuracy – the number of correct results in relation to the total amount of classifications [124]

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

- F1 score – the F1 score is the harmonic mean between precision and recall [256]

$$F1 = 2 \cdot \frac{P \cdot R}{P + R}$$

- Cohen’s Kappa metric – indicates if the observed value was better than a random value chosen by chance [170]

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

- ROC curve – depicts the true deceptive prediction rate as a function of the false deceptive prediction rate [111]
- Precision Recall (PR) curve – depicts the relationship between precision and recall [264]
- Area Under Curve (AUC) – used in the context of both the ROC curve and PR curve. In the case of ROC, it indicates the chance of a random true deceptive prediction result being achieved over a false deceptive prediction. For PR, on the other hand, it indicates the chance of getting the decision right.

Table 5.5 shows the confusion matrix extended with these additional metrics.

For the research at hand, the Accuracy, Kappa, F1 score, and AUC were considered to evaluate the models. In skewed distributions, it is known that Accuracy and ROC-AUC suffer [220] [170]. The F1 score is suggested as an alternative [170], especially when the imbalanced data has been balanced using a sampling technique [286]. The F1 score and ROC-AUC metrics are in addition often used in research aimed at detecting spam and bot accounts to determine the effectiveness of the machine learning models [113] [340] [115]. More recently, PR-AUC has been recommended as a better alternative to ROC-AUC [86] [264] as it does not account for

Table 5.5: Theoretical depiction of a confusion matrix with additional metrics

		Observed		
		Deceptive	Trustworthy	
Predicted	Deceptive	TP	FN	Sensitivity/Recall $\frac{TP}{TP+FN}$
	Trustworthy	FP	TN	Specificity $\frac{TN}{FP+TN}$
		Precision $\frac{TP}{TP+FP}$		Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$

the true negatives. In this case it means correctly finding those users who are trustworthy, which was not important for this research. Figure 5.5 shows the difference between the calculations of PR-AUC and ROC-AUC.

5.4.3 Developing a model for identity deception detection

The final step is to develop a model based on the outcome of the executed experiments. The findings will be communicated by proposing an Identity Deception Detection Model (IDDM) [314] that is structured to consist of two sub-components:

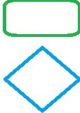
- The Identity Deception Detection Machine Learning Model (IDDMLM).

The IDDMLM makes use of the results produced while trying to experiment with supervised machine learning models and features that can detect human identity deception on SMPs. The machine learning models are of particular interest to the IDDMLM because the latter determines whether a new SMP user, one that was not used during experimentation, is deceptive or trustworthy.

- The Identity Deception Detection Score Model (IDSM).

With the IDDMLM, an SMP user was determined to be deceptive or trustworthy. The IDSM component goes further and takes this prediction, together with the entropy information that was calculated during earlier experimentation, to produce a view on the contribution of the attributes and features used in the prediction. The entropy indicates how much can be learnt about an SMP user's deceptiveness, given a specific attribute or feature. The IDSM model provides an opportunity to explain why an SMP user was deemed deceptive or trustworthy. The IDSM results from SMP users can also be compared to understand why one user is, for example, deemed to be more deceptive than another.

		Observed		
		Deceptive	Trustworthy	
Predicted	Deceptive	TP	FN	Sensitivity/Recall $\frac{TP}{TP+FN}$
	Trustworthy	FP	TN	Specificity $\frac{TN}{FP+TN}$
		Precision $\frac{TP}{TP+FP}$		Accuracy $\frac{TP+TN}{TP+TN+FP+FN}$



▭ ROC
◊ PR

Figure 5.5: Explaining the difference between ROC and PR

5.4.3.1 Model interpretation

The requirement of the IDSM component is to interpret the results from the IDDMLM. This is important for a number of reasons. *Firstly*, the interpretation explains how the model determines whether a human is deceptive or trustworthy. With this knowledge, a human who looks at these results can apply common sense to discern whether the model’s decisions seem reasonable. Take for example the case of whether the model would predict someone as being deceptive, given their hair colour. Human intuition indicates that this feature is probably not a good indicator of deceptiveness and the model most probably requires some more development. *Secondly*, the interpretation allows an agency, like law enforcement, to act on the results, and it highlights those humans who are most deceptive, with reasons as to why this decision was reached. This gives law enforcement a starting point not only to find these individuals, but also to prioritise their investigation effort. *Thirdly*, this information can be used by SMPs themselves to improve their platform. If an SMP, for example, knows that location is an indicator of deception, they could implement additional measures to verify a person’s location.

Various methods have been proposed in related work to interpret model results for supervised machine learning. Saabas [263] in essence reverse-engineered the random forest algorithm to produce a method, call ‘tree interpreter’. His method shows how the contribution of each feature, for each decision tree in the forest, can explain the final decision reached. Saabas’s method [263] caters only for decision trees and requires each individual decision tree’s result to be produced. Olah et al. [263] interpret the images recognised by a neural network machine learning model by showing how each layer of the neural network visually evolves. This method is only specific to images and neural networks.

Ribeiro et al. [257] propose a method called ‘LIME’ (Local Interpretable Model-Agnostic Explanations) and interpret the decision for each individual point in a dataset rather than to look at the contribution of engineered features to the global dataset as per the work from Saabas [263]. For this research, for example, the LIME method would take one SMP user and iteratively change the weights (contribution in this case) of the attribute or feature values of that user. The LIME model then measured the effect of these changes, through a linear model, to reach the outcome of the decision and to explain the results of that decision to the SMP user. The risk of using the LIME model is that local explanations might not be possible for complex predictions that cannot be linearly modelled. With game theory, the Shapely value [275] presents an alternative to LIME. Lundberg and Lee [209] proposed that by using Shapely values to change feature weights, an explanation can be achieved that resembles human intuition more closely. Shapely values also suffer from interpreting complex models as was described with LIME.

Model interpretation research shows that current interpretation methods can be used in the following ways to explain the results obtained from a model that detects identity deception by humans on SMPs:

- To interpret the reason for a single SMP user being deceptive (local) or deceptiveness by humans on SMPs in general (global).
- To interpret the results from text and images.
- To interpret the results from various machine learning algorithms, with some methods being model-agnostic.

Related model interpretation research furthermore indicates that interpretation is only calculated after the degree of deceptiveness of an SMP user was predicted. This is expected. However, the computational overhead associated with the interpretive calculations differs. LIME, for example, requires many iterations of models to be developed for an individual SMP user to explain their deceptiveness. This overhead needs to be considered in an environment like SMPs, where data is produced at high velocity. Explanations are required for actions that need to be taken promptly by (for example) law environment to protect individuals from being targeted.

Due to the computational overhead associated with related interpretation methods mentioned, the researcher looked towards the use of entropy as a measure with which results can be explained. Entropy is calculated during the development time of a model and readily available to be applied to interpret the results from the supervised machine

learning model. In general, entropy refers to uncertainty. Claude Shannon introduced information entropy already in 1948 [273], when it was initially applied to information compression by determining the quantity of information that can be discarded during transmission before a message becomes irretrievable [314]. The more uncertain an event, the more data it will contain in general [258]. Entropy is based on information theory and represented by the following formula [273]:

$$H(A) = - \sum_{i=1}^n p_i \log_2 p_i$$

Entropy is usually associated with ‘information gain’. $G(B|A)$ is used to measure how much information is gained by knowing that B is a subset of $A(B \subset A)$ [273].

$$G(B|A) = \log_2 \frac{1}{p(B)} = -\log_2 p(B)$$

Entropy and information gain are used not only in certain machine learning model decisions, but also to understand how much information is contained within the engineered features that were given to a specific machine learning model. This gives valuable insight into knowing which attributes contribute more towards identity deception detection. If one attribute, for example, has an entropy value of 50 and another 25, the attribute with an entropy of 50 is perceived to contribute twice as much as the other during the development of a human identity deception detection model. Entropy is more known to be applied to feature selection in machine learning [349] [122] [142] than to model interpretation, as proposed by the researcher.

5.5 High-level design steps

Figure 5.6 summarises all the steps mentioned for a model to assist in the automated detection of identity deception by humans on SMPs in a high-level design. The high-level design also indicates where the results of executing each of these steps will be discussed in following chapters.

Table 5.6 shows how, by following these mentioned steps, the requirements expected of a model that proposes to detect identity deception by humans on SMPs are met. The identity deception detection model adheres to the combination of the requirements catered for by each of the underlying steps.

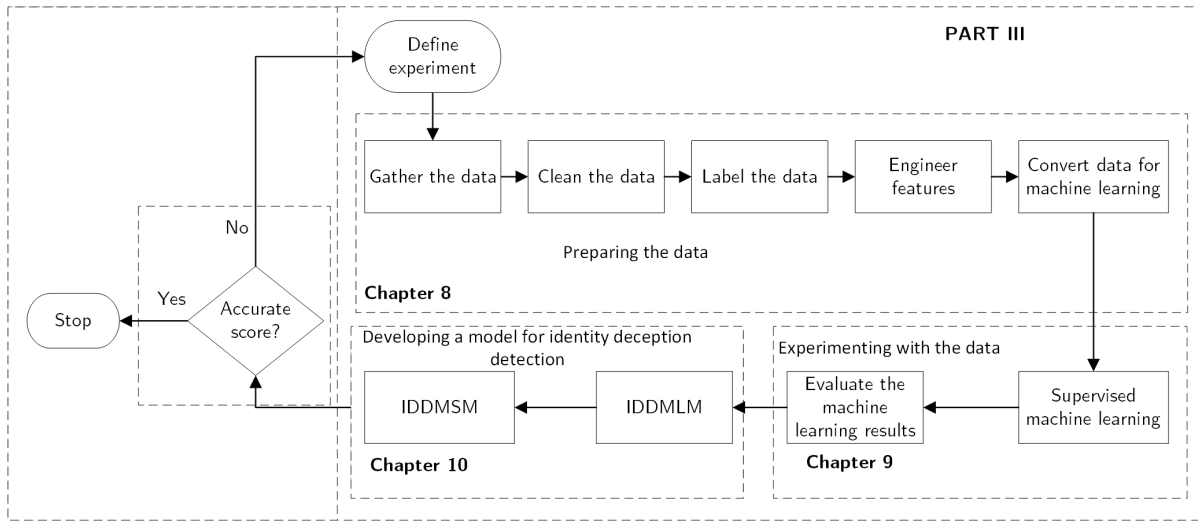


Figure 5.6: High-level design steps towards identity deception detection by humans on SMPs

Table 5.6: Identity deception detection model requirements catered for by the steps

Steps to assist in detecting human identity deception on SMPs	The requirement expected of a model that detects human identity deception on SMPs
1. Preparing the data	Find those humans who are being deceptive about their identity for malicious purposes. Ignore content posted by users on SMPs. Identify identity deception by humans; non-humans accounts are disregarded. Ignore attributes not contributing to human identity deception detection. Given Interpersonal Deception Theory (IDT), focus on finding deception as opposed to finding the truth. This means that attributes and engineered features will be indicative of deception as oppose to the truth. Use only attributes defining a user account that are available on SMPs. Features should be engineered such that they complement the detection of identity deception. The correct feature selection will lead to better results.
2. Experimenting with the data	Develop a supervised machine learning model. Evaluate various machine learning models. Make use of labelled data.
3. Developing a model for identity deception detection	Ensure that the machine learning model results are reproducible. Ensure that the machine learning model results are interpretable. Given a machine learning model, the detection of identity deception by humans on SMPs should be automated.

5.6 Conclusion

A research approach should be systematic, logical, empirical, and replicable [189]. The steps defined by the researcher provided a sequence that, if followed, would assist in the automated detection of human identity deception on SMPs. The steps were not only presented in a logical sequence but resulted in empirical values that could be evaluated to demonstrate the success of the proposed research approach. Furthermore, the details of each step were described in this chapter to support replicability for verification, implementation, as well as future research purposes.

The described steps satisfied the requirements expected from a model that can assist in the automated detection of identity deception humans on SMPs. The next chapter describes a prototype that implements this model, using the steps identified in this chapter.

Chapter 6

A prototype for assisting the automated detection of identity deception

“Design is not just what it looks like and feels like. Design is how it works.”
– Steve Jobs

6.1 Introduction

In the previous chapter, the researcher proposed steps to assist in the automated detection of identity deception by humans on Social Media Platforms (SMPs). These steps consider the requirements expected of a model proposing to detect human identity deception on SMPs. For example, the model is required to detect deception from humans only and therefore the data is cleansed from non-human data during the preparation step. The steps also allow for experimentation with various attributes and features describing the identity of a human. Additional knowledge gained from bot and psychology research fields suggests additional new features that can be engineered to increase the success of a model that proposes to assist in the automated detection of human identity detection on SMPs.

This chapter describes how these steps can be implemented by means of a prototype. Firstly, the goals and objectives of the prototype are defined. Thereafter the components of the prototype, their relationship with the steps, and their interaction with one another are presented by using, among other, a model employing the Unified Modeling Language

(UML) notation. Furthermore, each of the individual prototype components is discussed in detail by means of UML sequence diagrams. The UML sequence diagrams show the flow of messages and data within each component. The components and underlying interactions within each component present a prototype that can be implemented to assist in the automated detection of human identity deception detection on SMPs.

6.2 The objectives of the prototype

A prototype is an initial model of a product, that is built with the requirements of such a model in mind, to test a concept [331]. In this case, the presented prototype proposes an initial solution that can assist in the automated detection of human identity deception on SMPs. The proposed prototype has the following objectives:

- By implementing the steps through the prototype, each step's *purpose* is validated [102]. For example, without the machine learning preparation step, supervised machine learning results cannot be trusted, because supervised machine learning expects the data that is used to develop the models to be discretised, centred, and scaled [191].
- The *requirements* expected of a model that assists in the automated detection of human identity deception on SMPs can be fulfilled when the steps are implemented through the prototype. For example, the proposed model expects to detect identity deception from humans only. Therefore the step that cleans the data and removes non-human data can ensure that this requirement is met.
- The prototype allows for a range of *experimentation* and testing of hypotheses [110]. For the current research, additional features found in the fields of bots and psychology are proposed to increase the success of a model that can assist in the automated detection of human identity deception detection on SMPs. By executing various machine learning experiments, the importance of these features can be tested rather than when using attributes available from SMPs alone.
- The prototype proposes further *collaboration* with the industry and/or fellow academics [139] [110]. By providing an initial model to assist in the detection of human identity deception detection on SMPs, the industry and/or academia can use this knowledge to either validate or improve on the current process.

Once the above objectives are known, it is possible to define the components of the prototype to meet such objectives.

6.3 The components of the prototype

To describe the prototype, the researcher investigated various modelling languages. One such modelling language is the UML, a visual modelling language for systems [235]. It helps to define a prototype during the design phase, instead of during development. This approach not only describes the prototype at the beginning of development, but also minimises the risk of it not complying with the objectives and only finding this out at the end of the development. Other modelling languages were considered, for example, the Business Process Modelling Notation (BPMN) [97], Systems Modelling Language (SysML) [119], and Entity Relationship Diagrams (ERDs). BPMN models the actions between components as opposed to actions within the components themselves, SysML is an extension to UML for system management over a life cycle, and ERDs are most often used to describe entities and relationships in data. A UML diagram, on the other hand, allows for modelling the components of a prototype, including the relationships or interfaces between and within these components [116] [203]. For this reason, the researcher chose UML above the other modelling languages to design the prototype for the research at hand.

Figure 6.1 shows the UML component diagram of a prototype that assists in the automated detection of human identity deception on SMPs. The figure shows that the prototype consists of three main components:

- Prepare – This component is responsible for the gathering, cleaning and preparation of SMP data that describes the identity of a human.
- Discover – This component is responsible for developing various supervised machine learning models through experimentation. During this process the researcher could discover the most accurate model to assist in the identification of human identity deception on SMPs.
- Detect – This component proposes a so-called ‘Identity Deception Detection Model’ also referred to as the IDDM model, based on the results discovered through experimentation.

In addition, external data is received from Twitter [310], the Google Face Application

Program Interface (API) [135], an external names database [221], and the OpenStreetMap API [234].

The prototype is responsible for implementing the design steps identified in the previous chapter. A component is denoted by the ‘<<component>>’ label and the \square symbol [116]. A component, which is a reusable piece of functionality [116], can be composed of various sub-components and data. The sub-components are similarly denoted by the \square symbol but labelled as ‘<<executable>>’ to indicate that the sub-component executes a specific task within the functionality expected of the main component. The data is labelled as ‘<<document>>’ and is used to share information between components. The relationship or interfaces between the main components is indicated by a ‘lollipop’ and ‘socket’ symbol [116]. The lollipop, usually shown as a small circle ($\circ-$), shows that the component exposes an interface. The socket, usually shown as a half circle (\lrcorner), shows that a component consumes an interface of another component [116]. Interfaces to other external components are denoted by the ‘<<external interface>>’ label.

Figure 6.2 shows the correlation between the requirements expected of a model that assists in the detection of human identity deception on SMPs and the steps implemented via a prototype. The figure also confirms that human identity deception on SMPs is the cyber threat being studied in this research and lists the requirements of a model that can assist in the detection of human identity deception on SMPs. Next, the figure shows how these requirements are implemented by introducing three steps. The underlying steps are indicated with dashed lines. Lastly, Figure 6.2 shows how the components of the prototype propose to implement these steps.

Each of the three prototype components will be discussed in detail next.

6.3.1 Prepare

The first component of the prototype is concerned with preparing the data. To further explain this component, a UML sequence diagram [116] illustrates the messages and flow of data within the component itself. The ‘prepare’ component itself is denoted by the symbol. The arrows within the UML sequence diagram denote the flow of messages between the component and its sub-components. A solid arrow line indicates a request and a dashed arrow line indicates a response to a request. The vertical block lines in the sequence diagram indicate where a component’s messages begin and end.

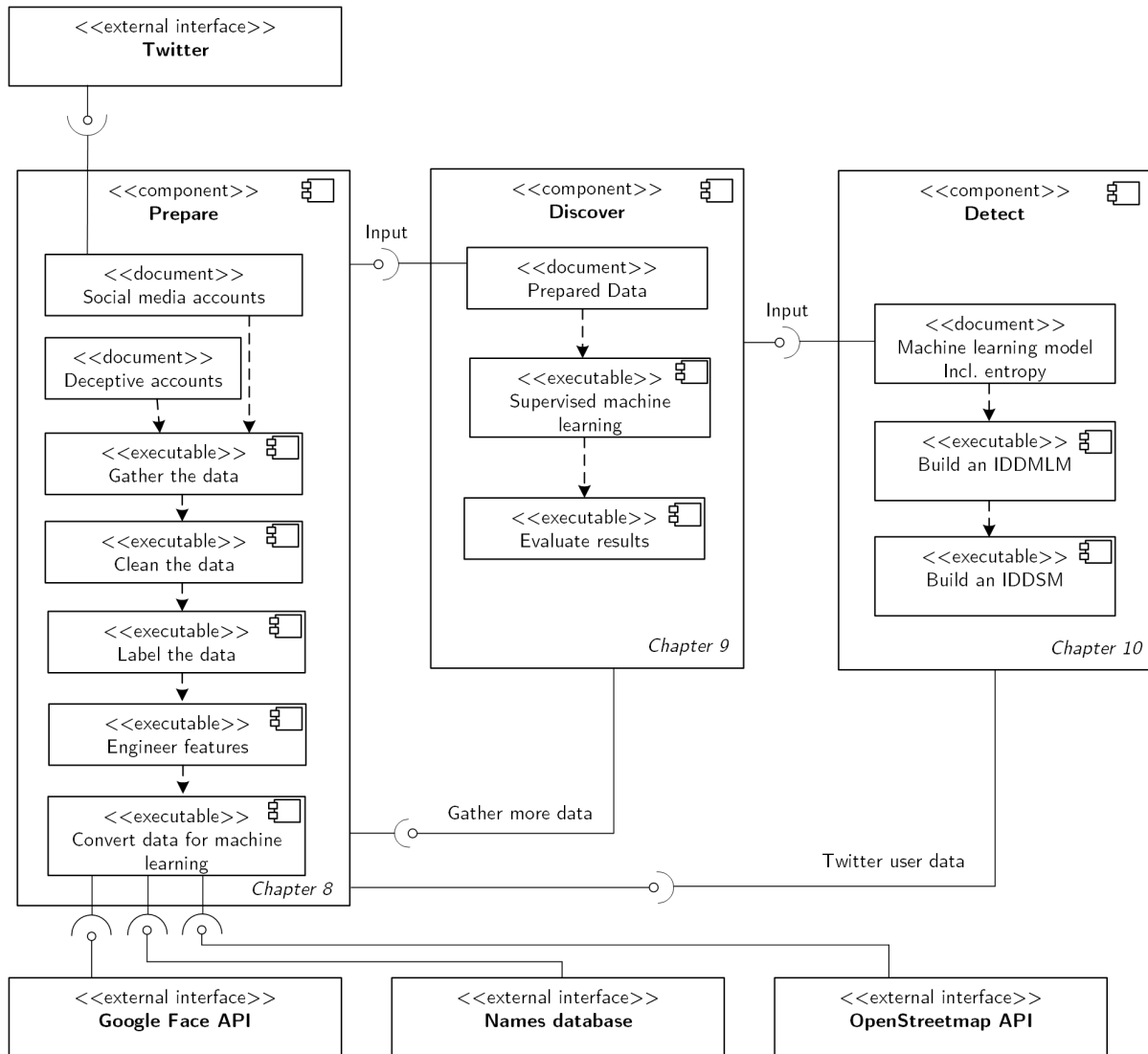


Figure 6.1: The prototype – UML component diagram

Figure 6.3 illustrates the sequence of messages of the component that prepares the data required by the prototype to assist in the automated detection of human identity deception by humans on SMPs. Data is gathered from Twitter and cleansed of potential bot accounts by using rules such as that the account has replied directly to a message of another, as described by Cresci et al. [80]. The resultant dataset is then combined with ‘deceptive’ accounts to create a labelled corpus of ‘trustworthy’ and ‘deceptive’ accounts. Additional engineered features are added to the corpus as determined from related research work stemming from bots and psychology. Lastly, supervised machine learning expects the SMP data to be discretised, centred, and scaled [191]. The data is therefore converted to meet these requirements.

Cyber threat	Identity deception by humans on SMPs		
Cyber-security requirements	<ul style="list-style-type: none"> Find those humans who are being deceptive about their identity for malicious purposes Ignore content posted by users on SMPs Identify identity deception by humans; non-humans accounts are disregarded Ignore attributes not contributing to human identity deception detection Focus on finding deception as opposed to finding the truth Use only attributes defining a user account that are available on SMPs Features should be engineered such that they complement the detection of identity deception 	<ul style="list-style-type: none"> Develop a supervised machine learning model Evaluate various machine learning models Make use of labelled data 	<ul style="list-style-type: none"> Ensure that the machine learning model results are reproducible Ensure that the machine learning model results are interpretable Given a machine learning model, the detection of identity deception by humans on SMPs should be automated
Research steps to meet the cyber-security requirements	Preparing the data <ul style="list-style-type: none"> Gather the data Clean the data Label the data Convert data for machine learning Engineer features 	Experimenting with the data <ul style="list-style-type: none"> Supervised machine learning Evaluating the machine learning results 	Developing a model for identity deception detection <ul style="list-style-type: none"> IDDMML IDDSM
Prototype to implement the research steps	Prepare	Discover	Detect

Figure 6.2: Correlation between the requirements, steps and a prototype proposing to assist in the detection of human identity deception on SMPs

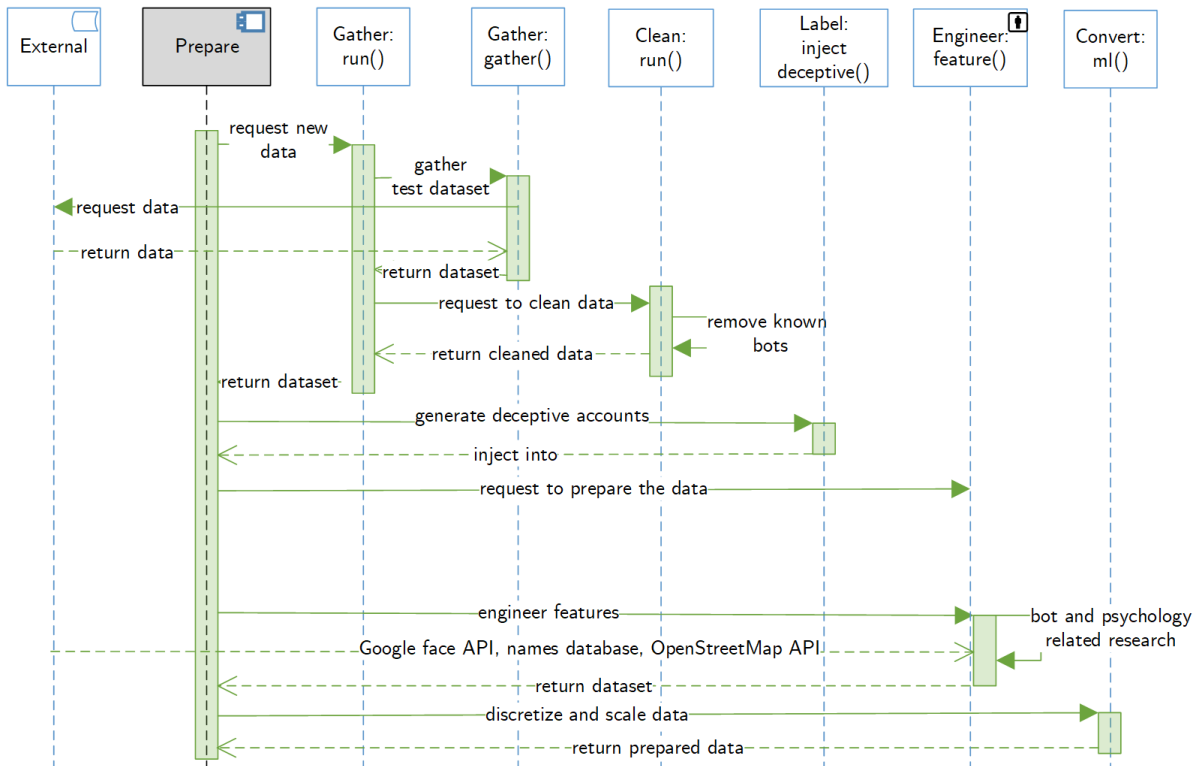


Figure 6.3: UML sequence diagram – ‘Prepare’ component

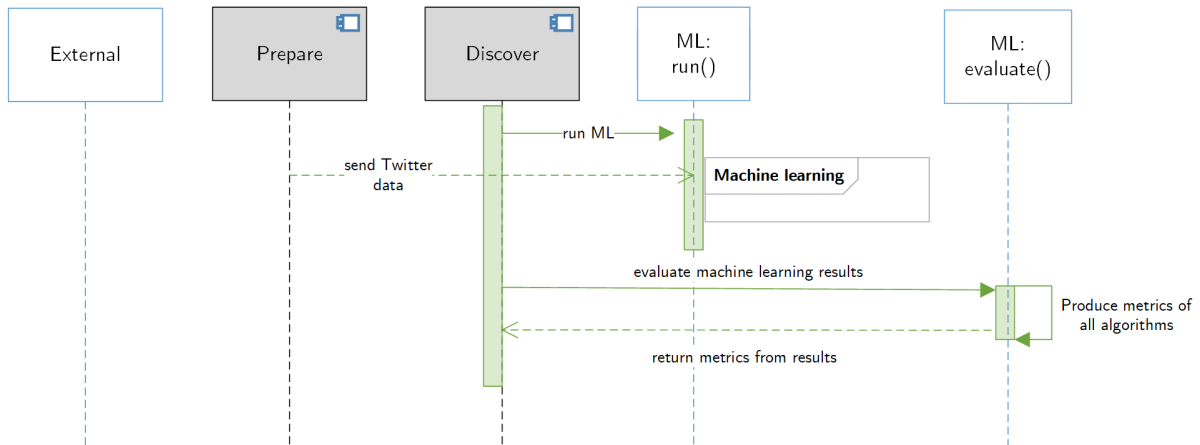


Figure 6.4: UML sequence diagram – ‘Discover’ component

6.3.2 Discover

Figure 6.4 illustrates the next component of the prototype. The ‘discover’ component proposes to develop various models with which to identify deceptive accounts. The result is a set of supervised machine learning models that could assist in the automated detection of human identity deception on SMPs – with differing accuracy. These supervised machine learning models are developed using different machine learning algorithms such as random forest, hyperparameters such as tree depth, attributes such as the number of friends, and engineered features such as the gender of the user. Each machine learning model includes a set of metrics such as F1 score, with which the models can be evaluated to determine how accurately each of them assists in detecting human identity deception on SMPs.

The development of a machine learning model that can assist in the automated detection of human identity deception on SMPs involves various messages. The process starts at splitting the prepared SMP data into a training and a test set. The training data set is used to create or develop a model, whereas the test data set is used to validate the developed model. Before using the training data set, a super sampling technique called Synthetic Minority Oversample Technique (SMOTE) is used to ensure that the training data set contains equal amounts of ‘trustworthy’ and ‘deceptive’ accounts. Many iterations are required in supervised machine learning to determine the best combination of algorithm and hyperparameters to assist in the detection of human identity deception on SMPs. The hyperparameters will depend on the supervised machine learning algorithm. Each algorithm has different hyperparameters that can be set. Lastly, the newly created models are tested against the test data set to confirm

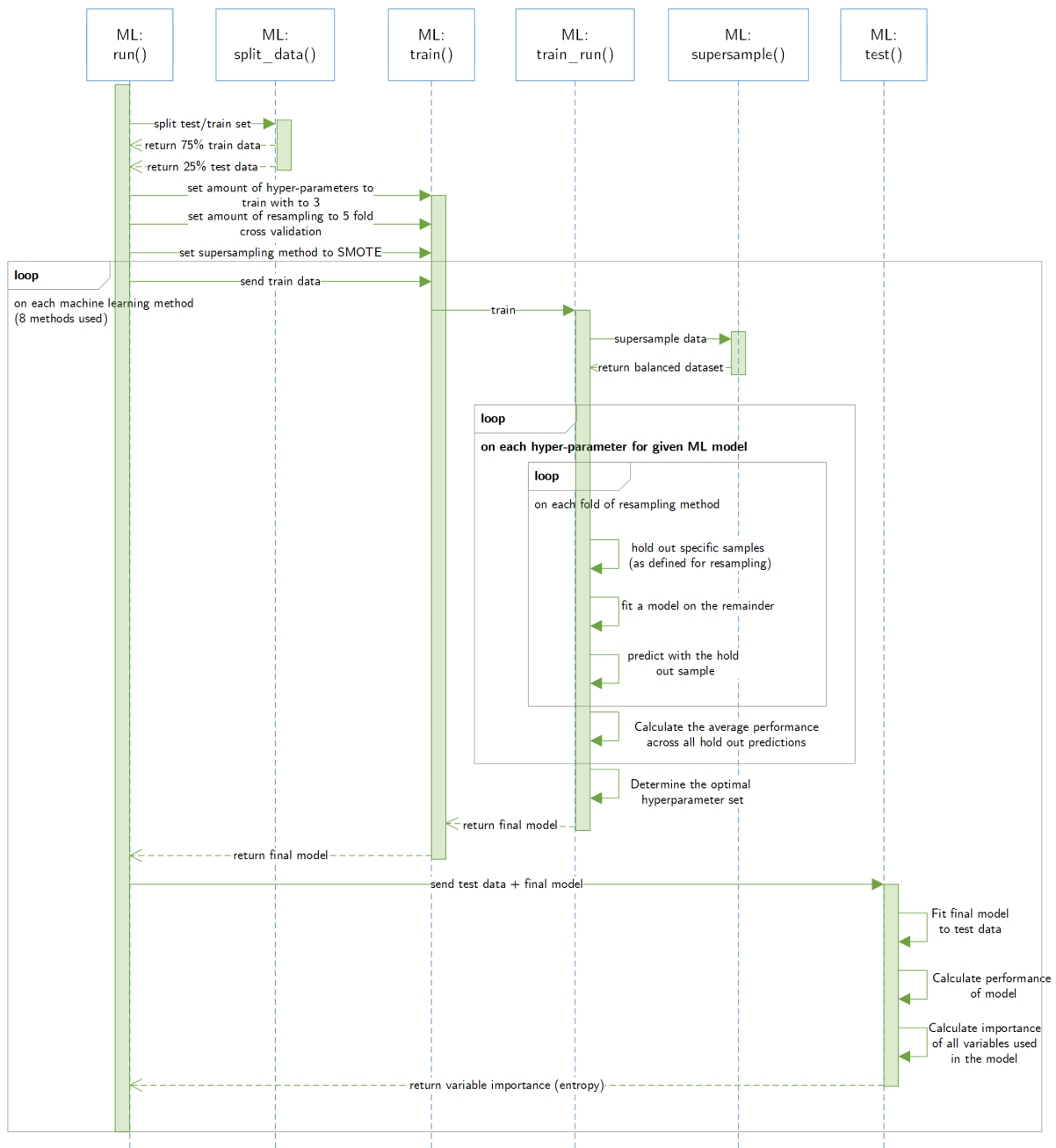


Figure 6.5: UML sequence diagram - Machine Learning within the ‘discover’ component

their performance. This flow of messages is illustrated in a separate sequence diagram in Figure 6.5.

6.3.3 Detect

The last component uses the best machine learning model results identified from the ‘discovery’ component, to propose a model with which to detect identity deception by

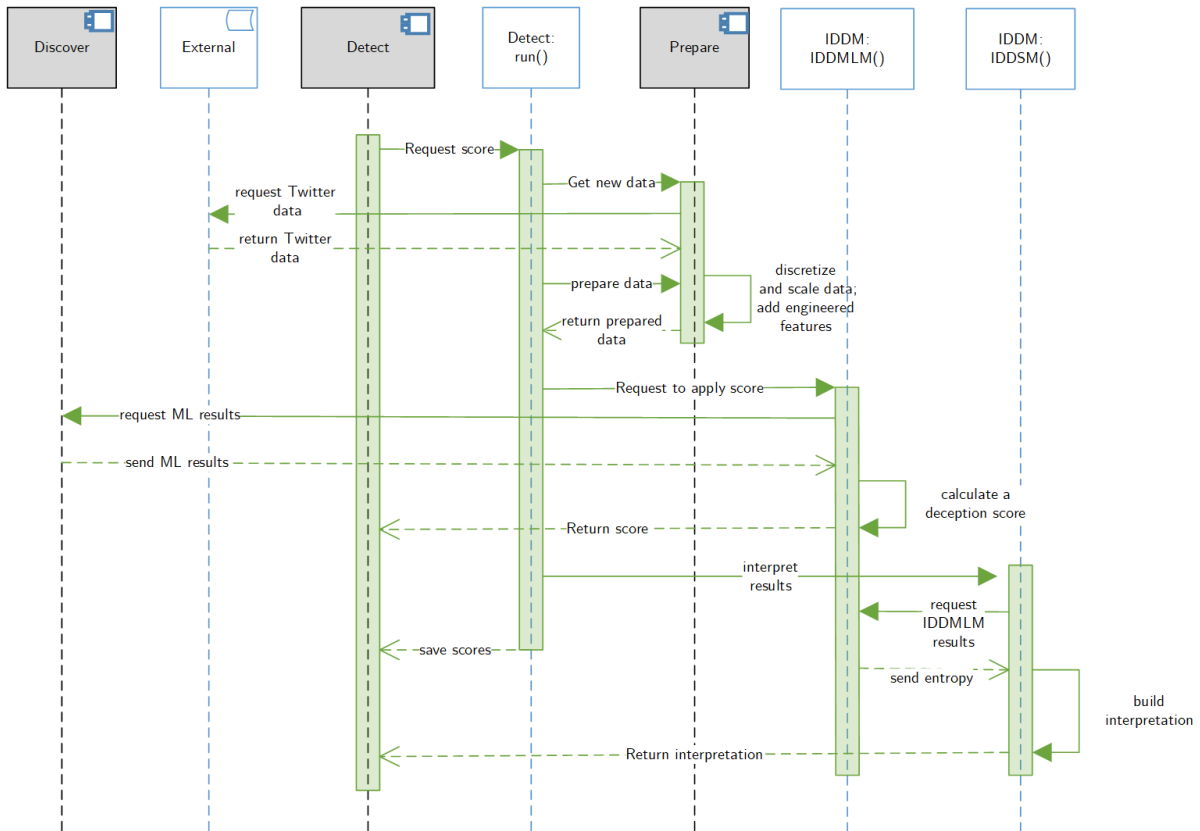


Figure 6.6: UML sequence diagram – ‘Detect’ component

humans on SMPs. The ‘detect’ component describes how identity deception by humans can be detected by supervised machine learning, but also how these results can be interpreted. To further explain this component, Figure 6.6 shows a UML sequence diagram [116] of the sequence of events within the component itself. The component accepts real-time data from Twitter and prepares the SMP data in the same way for supervised machine learning as in the previous ‘prepare’ component. This component then uses the best model, as per the F1 score and developed previously by the ‘discover’ component, to determine whether a human is deceptive or not. This detection is done by the Identity Deception Detection Machine Learning Model (IDDMMLM) component. Lastly, the result is explained by the Identity Deception Detection Score Model (IDDSM) component that develops an interpretive model to explain the previous result.

6.4 Conclusion

This chapter described the composition and functioning of a prototype. The objectives of the prototype showed that it will implement the steps required of a model that assists in the automated detection of human identity deception on SMPs. The prototype was described using UML notation.

A UML component diagram was used to illustrate the three main components of the prototype, namely those responsible for *preparing* the data, *discovering* a model towards human identity deception, and for *detecting* human identity deception in an automated fashion. Each component was depicted in an UML component diagram, together with its interfaces to external components and each other. Furthermore, UML sequence diagrams of each component showed the flow of messages and data within and between components. Twitter data was prepared by the ‘prepare’ component for supervised machine learning and to detect identity deception. The prepared data was used by the ‘discover’ component to develop a supervised machine learning model to detect identity deception. Lastly, current Twitter data was prepared and used by the ‘detect’ component to discover identity deception by means of the supervised machine learning model that had been developed before. With this understanding, a prototype can be implemented to assist in the automated detection of human identity deception on SMPs.

Besides the design of the prototype, a physical environment is required in which to implement the prototype. The next chapter describes the research environment and choices that were made for the development of the proposed prototype.

Chapter 7

The research environment

“If you do it right, it will last forever.” –Massimo Vignelli

7.1 Introduction

To successfully implement a prototype that assists in the automated detection of identity deception by humans on Social Media Platforms (SMPs), a specialised technical research environment is required. For example, one might consider the type of data required to detect deceptive identities. This data might include, among others, images of persons and their GPS locations, while the research environment should for example allow for the storage of such heterogenous content.

This chapter describes the research environment considered for a prototype that assists in the automated detection of identity deception by humans on SMPs. It is important to describe this research environment to not only ensure reproducibility in future research work, but also to understand the technical deliberations for implementing this prototype. Chapter 7 describes the infrastructure – the hardware, network, database and software – that was considered for the research environment to be able to implement the proposed prototype. It also provides an overview of the research environment infrastructure. Thereafter, a detailed discussion follows of each of the infrastructure’s components to emphasise their responsibility and necessity in the research environment.

7.2 The research environment

Related research posited various environments for the processing of big data in general and suggested distributed computing environments, like Hadoop [32] [36], and cloud computing environments, like Amazon Web Services [42]. Most of the environment choices were divided into problems relating to either data complexity or computational complexity [171]. Because identity deception detection is perceived as both a complex data problem and a computation problem, one wonders whether previous work like that of Agrawal et al. [193] and Xiaolong et al. [171], which focus on creating big data environments, can be used for the research at hand.

According to Agrawal et al. [193], heterogeneity, scale, timeliness, privacy, and human collaboration should be considered in a big data research environment, whereas Xiaolong et al. [171] mention the importance of clear requirements, the right data and an integrated solution to solve problems in a big data environment. Neither of them elaborated on what is technically required of a big data research environment. To assist in the automated detection of deceptive identities on SMPs, many machine learning iterations are, for example, proposed by the researcher in order to find the most accurate model for predicting this form of deception. The environment firstly allows for the timely discovery (as mentioned by both Agrawal et al. [193] and Xiaolong et al. [171]) of an accurate identity deception detection model over many iterations. Secondly, the experimental results must affirm the technical considerations for an infrastructure proposed to assist in the automated detection of human identity deception on SMPs. An overview of the infrastructure proposed by the researcher to this effect is illustrated in Figure 7.1. The infrastructure can be divided into hardware, database, network, and software components.

To support the use of the components within this infrastructure, the researcher executed various supervised machine learning experiments relevant to certain components. A subset of SMP user data was used to identify identity deception using SMP attributes only. The focus of these experiments was on the run times of the experiments rather than on the accuracy of the models. The intention was to understand how some of these components, proposed for the infrastructure, affected the research environment. The quicker a new model could be developed, the more time it would allow not only for more experimentation, but also for acting on the information about potential deceptive identities in the real world.

For these experiments, the researchers looked at work of Delgado et al. [112], who

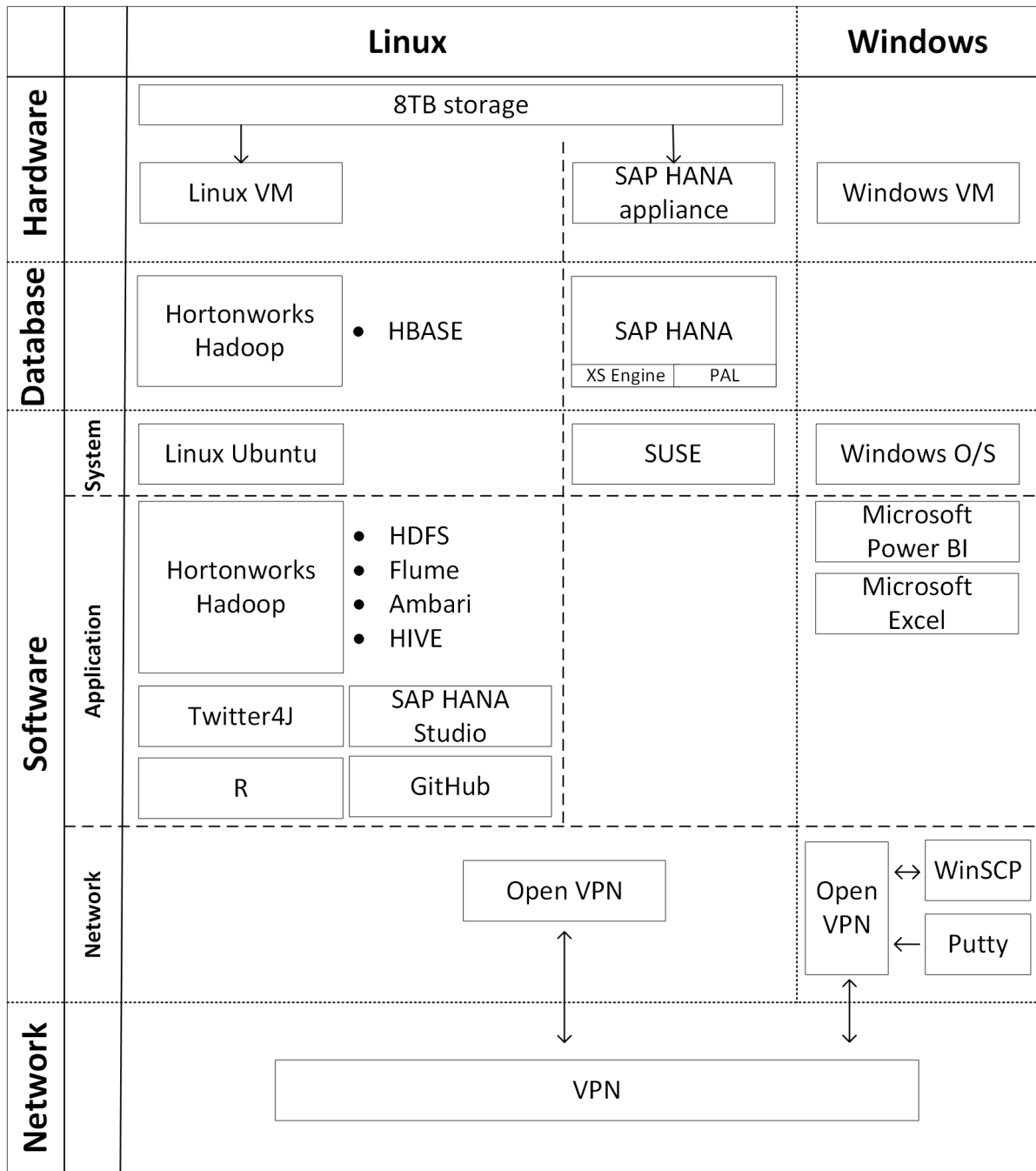


Figure 7.1: A proposed research environment infrastructure overview

tested 179 different classification algorithms from 17 different families and found that the random forest machine learning algorithm still outperforms most other algorithms. Therefore, the random forest algorithm was used for these supervised machine learning experiments. Cross validation was used to cater for potential bias in the data [16]. During cross validation, data is split equally across the indicated number of folds. Machine learning models are then developed over several iterations, with each iteration using a different fold for validation. The splitting of data can also be repeated to create

different versions of the folds. By recording the average performance over all developed models, more realistic model performance metrics are produced. Besides cross validation, the random forest algorithm itself accepts a value indicating the number of entries that should be used at each point in the tree before a decision can be made. This is also known as the hyperparameters of a machine learning algorithm. The choice of hyperparameter could well influence the time it takes to develop a machine learning model.

Each of the components of the infrastructure, and the considerations as to why these components were chosen for a research environment, are discussed in detail next. Where relevant, the results from the supervised machine learning experiments are included and discussed.

7.2.1 Hardware

The researcher mostly used hardware from the Future Service-Oriented Computing (Future SOC) lab for this research [120]. This lab is part of the Hasso Plattner Institute (HPI) based in Potsdam, Germany. The HPI Future SOC lab hosts an annual conference focused on operations within the context of cloud computing. The researcher participated at this conference as a speaker in 2015 [34] as well as in 2017 [313]. Furthermore, the HPI Future SOC lab provides an opportunity for researchers to make use of the hardware hosted by the lab for research purposes. Researchers are required to submit a proposal for use of the hardware on an ongoing 6-monthly basis, together with an agreement to submit a technical research progress report and research poster after each 6-month period. The initial research proposal in which the researcher requested ongoing access to hardware for research into identity deception detection by humans on SMPs, was submitted in 2014. An example of a proposal, technical research progress report and research poster are included in Appendix D.

The HPI Future SOC research lab provided access to the following hardware, based on the research proposal:

- *Access to an SAP HANA appliance server with 2TB RAM, 32 CPUs / 100 cores*
SAP HANA is both a hardware appliance and a database. With a hardware appliance, all hardware is dedicated and configured for a specific purpose [60]. In this instance, the hardware is configured such that the SAP HANA database runs optimally. Dedicated hardware for the SAP HANA database has great advantages.

Data is loaded in memory and therefore returns results much faster than in the case of traditional relational databases that read data from disk [266]. For this research, this means that data exploration tasks could firstly be executed with greater speed as all data required to detect deceptive humans with is available in memory. Secondly, this improved speed allowed for more experimental iterations to develop various identity deception detection models for a given time period. Access to the SAP HANA appliance server was shared across different research projects. The reason for this was that the SAP HANA appliance server is costly and only one or two instances of this appliance are at a given time available to all researchers.

- *A Linux Virtual Machine (VM) with 64GB RAM, 4 CPUs / 8 cores*

The RAM and CPU are important for this research project because of its influence on the accuracy and speed required to develop a machine learning model. Data is loaded in memory during machine learning. The more RAM, the more data can be used to develop the models for the research to identify identity deception by humans on SMPs. More data allows for more accurate identity deception detection models [99]. On the other hand, the more CPUs, the more machine learning models can be developed in parallel [272]. This increased speed means that more experiments can be performed within a given time period to find a model towards identity deception detection by humans on SMPs.

- *8TB of storage space shared between the server and Linux VM*

The storage is required to store the huge volumes of data gathered from Twitter, as well as information about the developed machine learning models after each experiment. The corpus gathered for the research in hand consisted of 200GB of SMP data made up by 606 914 240 tweets from 223 796 Twitter users. However, additional space was required for experimentation, the storage of the engineered features (given related work in bots and psychology with regard to deception), and database operational log files. Furthermore, the average machine learning model's information takes up about 300MB of hard disk space. For the purposes of the current research, a minimum of eight machine learning models were built for each of the three experiments with 30 repetitions per model. This equates to 720 models, which consumes a total of 216GB hard drive space. The researcher's personal laptop only has 500MB of storage and therefore this research would not have been possible locally. In the end, 3.7TB was used for the research at hand as illustrated in Figure 7.2.

Filesystem	Size	Used	Avail	Use%	Mounted on
udev	32G	0	32G	0%	/dev
tmpfs	6.3G	113M	6.2G	2%	/run
/dev/mapper/vm--20162015--vg-root	134G	5.0G	122G	4%	/
tmpfs	32G	0	32G	0%	/dev/shm
tmpfs	5.0M	0	5.0M	0%	/run/lock
tmpfs	32G	0	32G	0%	/sys/fs/cgroup
/dev/vda1	472M	105M	343M	24%	/boot
192.168.42.50:/data_SAF/exchange	7.9T	3.7T	4.2T	47%	/hanamnt/SAF
nas3-1.san.fsoc.hpi.uni-potsdam.de:/home	4.0T	3.0T	972G	76%	/home
tmpfs	6.3G	0	6.3G	0%	/run/user/5659

Figure 7.2: HPI Future SOC lab storage used in total

To understand more about the effect of CPU cores on machine learning model development time, the researcher performed a machine learning experiment on the Linux VM. Table 7.1 shows the results of executing the same random forest machine learning experiment with one and 6 CPU cores respectively. The query time is the time it took to retrieve data from the database and the total time includes the query time and the model development time. The results from the experiment showed that more CPU cores improve model development time and therefore the Linux VM machine provided by the HPI Future SOC lab, with its eight available cores, fitted the researcher’s requirements for being able to develop machine learning models in a timely manner.

In addition to the HPI Future SOC lab hardware, a personal laptop was used for the following purposes:

- To connect to the Future SOC lab network
- To manually generate examples of deceptive SMP users
- To visualise the results from the Identity Deception Detection Model (IDDM)

The laptop had 16GB RAM, 500MB of storage, and 1 CPU / 4 cores. This hardware configuration was found to be sufficient for the purposes of this research.

7.2.2 Database

Two databases were used during the research, namely HBASE [278] and SAP HANA [266]. The former was primarily used to store the original gathered Twitter data. HBASE is an extension of the Hadoop framework [18] and allows for quick data reads/writes, since HBASE allows small amounts of data to be read from random locations in a large dataset [267]. This is important when exploring huge volumes of

Table 7.1: Comparing the CPU core performance

O/S	CPU cores	Number of SMP users	Number of Folds	Number of Repeats	Number of hyperparameters	Query Time (seconds)	Total Time (seconds)
Linux	1	155K+	5	0	3	29	17 603
Linux	6	155K+	5	0	3	23	6 791

data. Firstly, data irrelevant to the research could be discarded at this point already. Secondly, it would take time to read volumes of data sequentially. Thirdly, HBASE was distributed over many low-cost servers [156]. It was much more convenient to store and explore volumes of data at a low cost before moving only the SMP data necessary for the research to another costlier database solution like SAP HANA. In SAP HANA, for example, only content pertaining to minors was stored. This demographic feature of users is an example of a group of users targeted for identity deception by other humans on SMPs.

SAP HANA was used in addition to HBASE firstly due to its capability of storing data in memory. Machine learning algorithms have to retrieve and query the data. As the researcher was aiming to perform many iterations of machine learning experiments, the SAP HANA database would decrease the experimental run times and indirectly allow for more experiments. The second reason for using SAP HANA was its column store capabilities. This allowed the reduction of the space required to store the SMP data, as column-oriented databases would not store the data of the empty SMP attributes. A column-oriented database requires fewer disk reads, which increases query performance when data is sparse. The last reason for using SAP HANA was that additional components with specific capabilities were integrated in the SAP HANA database. SAP HANA's Predictive Analytics Library (PAL) has machine learning capabilities and the SAP HANA Extended Application Services (XS) Engine is an application server that allows data to be ingested into the SAP HANA database. It was found during the research that the PAL library did not match the capabilities of R in terms of richness of supervised machine learning algorithms. Therefore, the PAL library was not used in the final model proposed to assist in the automated detection of human identity deception on SMPs. However, the XS Engine was used to ingest data into SAP HANA.

Even through the SAP HANA database was shared with other research projects, it allows for multi tenancy [266], which means that each researcher has an own dedicated area within the database. This allowed the researcher to run various experiments and store the required data towards finding a model that can assist in the automated detection of human identity deception on SMPs to run without interference from other research

projects.

7.2.3 Network

Communication with the HPI Future SOC lab domain was made possible via a Virtual Private Network (VPN). The researcher was provided with a unique username and password. The username is restricted to resources on the domain to which access has been granted. Once the VPN connection is established for the domain, hardware devices in the lab are accessible as if part of the local network. A few years ago, it was not possible to conduct this type of research from a developing country (South Africa), but using such advanced infrastructure in a developed country (Germany) made it possible. For example, in 2017 the average Internet speed in South Africa was 3.4Mbps, while in Germany it was 10.2Mbps [282]. Considering that the bandwidth in Germany was three times as much as in South Africa, it allowed for the SMP data required for this research to be gathered in a timely manner.

Another valuable advantage of using the infrastructure in Germany was the fact that the infrastructure was always available. Experiments could be run remotely over days without any risk of data loss. As South Africa experienced electricity blackouts from time to time, research work would have been lost when a laptop or server lost internet connectivity.

7.2.4 Software

Various software was used for this research and can be divided into system, application, and network software. Each of these types is discussed next.

7.2.4.1 System software

System software provides a platform for other software to run from [289]. An example of system software is operating systems. For this research, three different operating systems were used as each was complementary to the hardware they applied to.

The SAP HANA appliance server has SUSE Linux installed as an operating system – a choice that was controlled by the HPI Future SOC lab. An additional Linux VM was provided by the lab to host any application software required for the research. The HPI Future SOC lab required that no other software be installed on the SAP HANA

Table 7.2: Comparing operating system performance

O/S	CPU cores	Number of SMP users	Number of Folds	Number of Repeats	Number of hyperparameters	Query Time (seconds)	Total Time (seconds)
Linux	1	50K	5	0	3	6	1 983
Windows	1	50K	5	0	3	630	2 123

appliance. This Linux VM was installed with Ubuntu Linux [58] due to its compatibility with the SAP HANA appliance server running SUSE Linux and the fact that it required no licencing fees [58].

Lastly, a personal laptop was used to connect to the HPI Future SOC lab. This laptop had the Windows operating system [225] installed in a VM. The reason why Ubuntu Linux [58] was not considered as an operating system for the personal laptop is that some of the application software used for this research, like Microsoft Power BI [224], only operates on the Windows operating system. The researcher did however do a machine learning experiment to understand the difference between running the experiments on the Windows VM and the Linux VM. Table 7.2 shows the results of this experiment. It is clear from the query time, in seconds, that the experiment with the Windows VM took much longer. This would make one think that the Linux VM was faster, but the result was deceptive. The query time was significantly more for the Windows VM as the data had to travel from the Internet to reach the Windows VM, whereas the Linux VM was on the same Local Area Network (LAN) as the SAP HANA database. The Windows VM was faster in developing machine learning models, but due to the query time, the overall Linux VM's performance was the best. Therefore, for this research environment, the choice was made to run all machine learning experiments on the Linux VM that had been provided by the HPI Future SOC lab.

7.2.4.2 Application software

Application software is designed to perform a function or task [63]. To identify deceptive humans on SMPs, a variety of tasks are required. These include gathering Twitter data; exploring the SMP attributes; labelling SMP accounts; developing various supervised machine learning models; visualising results; and lastly, saving the research in a code repository. The choice of application software made in terms of these tasks is discussed next.

7.2.4.2.1 Gathering Twitter data

For this research, both streamed and historic Twitter data was gathered, using the Hadoop platform that comprises various application

software. This platform was preferred, as Hadoop is known for handling large volumes of heterogenous data [272], like social media data, in this instance. All application software services running on the Hadoop platform, like Apache Flume [294] and Apache Hive [278], are managed via Ambari [23]. Ambari, for example, provides a means to start and stop software services such as Apache Flume’s agents [294].

The gathered Twitter data contained the initial corpus of tweets from SMP users originating from a targeted demographic, in this case minors. The streamed Twitter data was gathered using Apache Flume [294] and Twitter4J [310]. Flume is application software that can start multiple streaming services that can gather data from multiple sources, with very little configuration. For this research, only data from one source, namely Twitter, was gathered. The researcher integrated Twitter4J [310], being a Java Application Program Interface (API), into Flume, which was also written in Java. This integration enabled Flume to be configured to gather streaming data from Twitter as a source. Flume expects APIs to be integrated to know how to treat the data coming from specialised sources, like social media. The additional benefit of using Twitter4J was that the API could also request historic data – in this case, the last 3 200 tweets of a particular SMP user and additional account information not contained in the tweets. Flume has the ability to store the Twitter data in the Hadoop Distributed File System (HDFS) and directly thereafter, to insert selective data into HBASE [278] from HDFS. This saved time for this research, as no further intervention was required to read the data from HDFS and insert it into HBASE.

7.2.4.2.2 Exploring the SMP attributes

The gathered Twitter data consisted of the SMP attributes that described the identity of the SMP user. The SMP attributes stored in HDFS were kept in their original format. For this research, it meant that the JavaScript Object Notation (JSON) content, which is heterogenous in nature and received from Twitter, were stored ‘as is’. Application software called Apache Hive [278] was used to explore the SMP data directly from HDFS. However, this form of data exploration is slow, as data is retrieved sequentially during exploration. The decision was therefore made to insert the SMP attributes that are of interest to this research into HBASE. HBASE allows for quick read/writes of data at random locations. Apache Hive [278] could also be used to explore the data in HBASE. Hence it was used to understand which attributes were not only available, but also required, to be able to develop a model that can assist in the automated detection of human identity detection on SMPs.

Once data exploration was completed, the data was ingested by SAP HANA through another Flume task. In this case, HBASE was the source of the data, where previously Twitter had been the source. The researcher wrote her own Java API for Flume to know how to send data from HBASE to SAP HANA's XS Engine. Once data was loaded into SAP HANA, the researcher used application software known as SAP HANA Studio [151] to query the data on the database server. Other mechanisms of querying the data does exist, like Python plugins. SAP HANA Studio, however, offers a user-friendly graphical interface that makes it easy for anyone without knowledge of Structured Query Language (SQL) (the query language used to query the data) to explore the data.

7.2.4.2.3 Labelling SMP accounts

Once the data relevant to the research at hand was ingested into SAP HANA, additional examples of deceptive accounts had to be generated. Supervised machine learning expects a labelled set of data [190], in this case 'deceptive' and 'trustworthy'. All data retrieved from Twitter was deemed trustworthy as Halevy et al. [145] found that only 5% of people tell 40% of all lies. This means that with the inclusion of 10% fabricated deceptive accounts, most lies should be catered for by the generated deceptive accounts.

The deceptive accounts were created using two random human data generator APIs from the internet [25] [178]. Further manual intervention was required to complete the remaining attributes which the APIs could not do. Microsoft Excel [223] was used for this purpose, as the SAP HANA database has built-in capabilities to import Microsoft Excel data.

7.2.4.2.4 Developing supervised machine learning models

Once the data was prepared for supervised machine learning, R language application software [164] was used to develop models to assist with identity deception detection by humans on SMPs. R language application software has many packages such as Caret [191] and ggplot2 [328] for machine learning and data visualisations. The Caret package was preferred for the following reasons:

- The package caters for 237 different machine learning algorithms as at June 2018.
- Since the Caret package provides a uniform interface to all packages, the researcher was not required to know each individual machine learning algorithm's methods for model development. This saved much time during the setup of the experiments.

- The scikit-learn package [79] found in Python was not as mature as the R Caret package at the time the experiments were being done. There was, for example, no uniform interface to all machine learning algorithms in 2014. However, the scikit-learn package is updated constantly and is a good alternative for future experiments or prototypes proposing to find models that can assist in the automated detection of human identity deception on SMPs.

In addition to the machine learning algorithm used, supervised machine learning requires several other input parameters like the number of folds, repeats, and hyperparameters. These input parameters can affect model development times. For this reason, the researcher performed additional machine learning experiments to understand how these parameters would influence the supervised machine learning models that had been developed by using the random forest algorithm. The results for the number of hyperparameters used in the machine learning model development are shown in Table 7.3, the results for the number of folds appear in Table 7.4, and the results for the number of repeats are shown in Table 7.5.

Table 7.3: Comparing hyperparameter performance

O/S	CPU cores	Number of SMP users	Number of Folds	Number of Repeats	Number of hyperparameters	Query Time (seconds)	Total Time (seconds)
Linux	6	155K+	5	1	3	23	11 695
Linux	6	155K+	5	1	5	23	16 873
Linux	6	155K+	5	1	10	23	39 072

Table 7.4: Comparing resampling fold performance

O/S	CPU cores	Number of SMP users	Number of Folds	Number of Repeats	Number of hyperparameters	Query Time (seconds)	Total Time (seconds)
Linux	6	155K+	5	3	3	23	25 323
Linux	6	155K+	10	3	3	23	54 154

Table 7.5: Comparing resampling repeat performance

O/S	CPU cores	Number of SMP users	Number of Folds	Number of Repeats	Number of hyperparameters	Query Time (seconds)	Total Time (seconds)
Linux	6	155K+	5	1	3	23	11 695
Linux	6	155K+	5	3	3	23	25 323
Linux	6	155K+	5	5	3	23	39 796

The above results show that the number of folds, repeats and hyperparameters have a direct linear impact on the total development time of the models. The higher the number, the longer the model development takes to complete. This is something that the researcher took into consideration for this research. Firstly, there was a limited time period in which the research could be conducted. To maximise the use of this time, a

trade-off had to be made between the number of experiments the researcher wanted to perform and the number of input parameters feasible within the given time period. In the end the researcher chose to use ten folds and three repeats for cross validation as suggested by similar research [16] [187]. In terms of the hyperparameters, the researcher chose three for this research. The choice of hyperparameter fell outside the scope of this research and therefore the researcher did not need to experiment with this input parameter. She accepted the three-best default hyperparameters as suggested by the Caret package [191].

7.2.4.2.5 Visualising results

Microsoft PowerBI [224] was used in the research to visualise the results obtained from the model proposed to assist in the automated detection of human identity deception on SMPs. Microsoft PowerBI runs on the Windows operating system only and has a built-in connector to the SAP HANA database. It also has a user-friendly graphical user interface to build a dashboard for displaying results. This step was not required and did not influence the results of this research. Therefore, this software was seen as optional.

7.2.4.2.6 Code repository

All code used for this research was backed up to a GitHub [252] repository. Since the step was not required and did not influence the results of this research, code repository software was considered optional.

7.2.4.3 Network software

For connection to the HPI Future SOC lab, the lab uses OpenVPN [347]. OpenVPN is open source and uses SSL to encrypt connections, but it does require additional software [280]. This software was installed on the researcher's Windows VM to connect to the lab in Potsdam, Germany. For connecting, transferring of data, and configuration of the VM instance, PuTTY [291] and WinSCP [291] were used. Firstly, PuTTY allowed the researcher to remotely gain control of the Linux VM to install any necessary packages and setup. The HPI Future SOC lab team is only responsible for the hardware, database, and network. It was up to the individual researchers to set up their own experiments with whatever application software they would require. The HPI Future SOC lab was available to provide assistance in terms of system and network software where required. Secondly, WinSCP [291] is application software to

easily transfer files between a Windows and Linux environment. Other application software does exist, like CoreFTrue Positive (TP) [77], and there is no particular benefit in using WinSCP over another, besides for a user-friendly graphical user interface and the fact that WinSCP is well supported.

7.3 Findings from the technical research environment considerations

After using the proposed infrastructure to implement a prototype that can assist in the automated detection of human identity deception on SMPs, the researcher found the following to simplify the architecture for future research:

- A Windows VM was used to connect to the remote environment with only Microsoft PowerBI [224] and Microsoft Excel [223] running on the VM used for data visualisation and the generation of deceptive example SMP accounts. These applications were not essential towards building an identity deception detection model. The researcher therefore proposes to have the infrastructure all on the same system software as it would help with maintenance and compatibility of software.
- Microsoft PowerBI [224] and GitHub [252] were optional to achieve the research goal. Both helped the researcher but could be seen as auxiliary software.
- The SAP HANA database provided the researcher with an environment that was very powerful to query the data and perform the experiments. However, various Not only SQL (NoSQL) databases like HBase, MongoDB and Cassandra [152] have since matured and could be regarded as more cost-effective alternatives.
- Many programming languages are currently available for developing machine learning models. There is no preference in the architecture for the chosen machine learning libraries used for developing a model.
- The researcher found the applications on the Hadoop platform to be beneficial to the research. Flume was very efficient at retrieving streaming data and the raw data, stored in HDFS, was accessible via Hive. The researcher used Hortonwork's distribution of the Hadoop platform although other distributions exist (e.g. Cloudera). The research did not require a specific distribution of Hadoop and therefore any could be included in a future architecture.

- Kafka and Nifi [9] could be alternatives to Flume in a future architecture, as these applications also provide real-time streaming capabilities.

Given this, a proposed simplified infrastructure for future research is depicted in Figure 7.3.

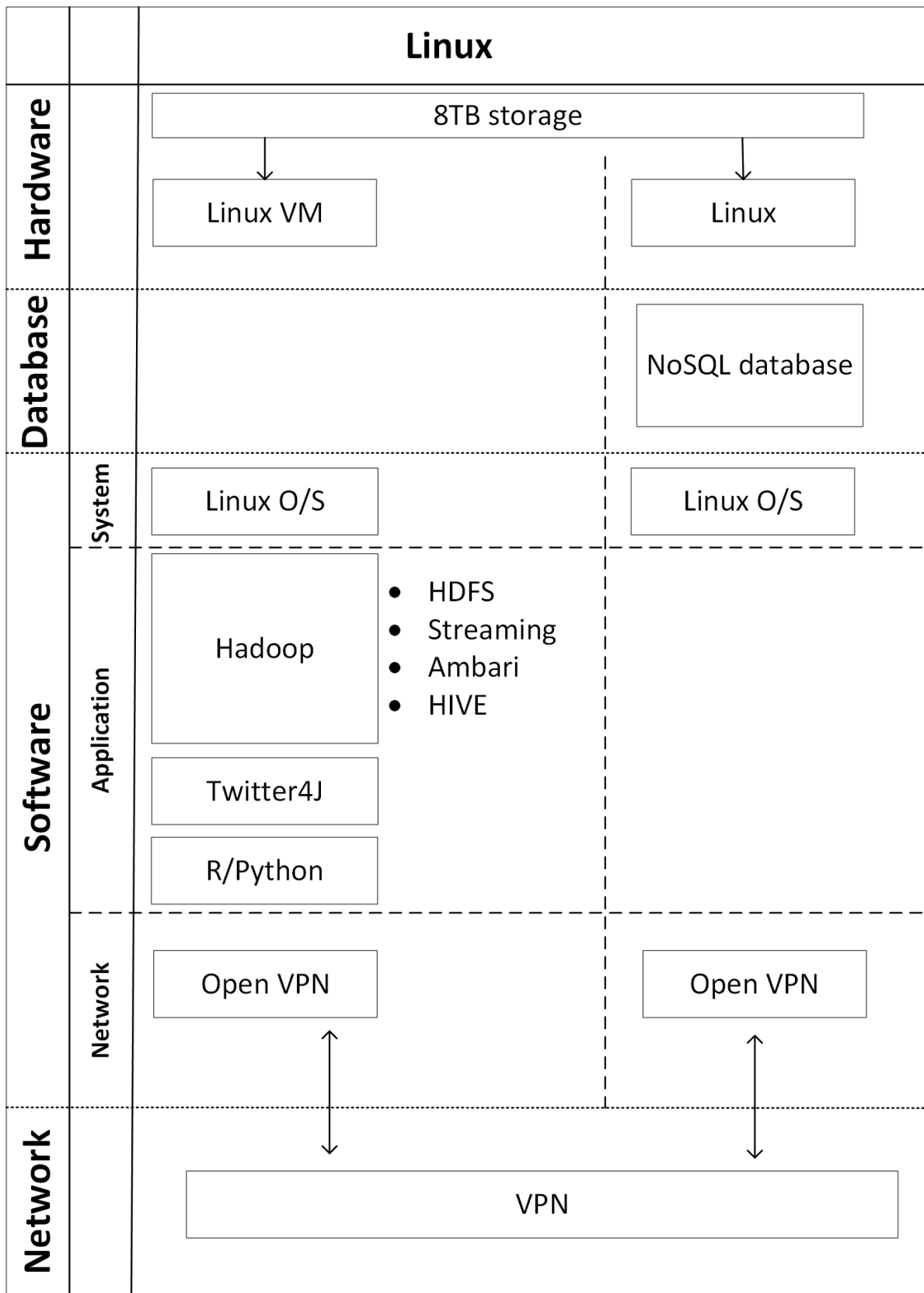


Figure 7.3: A proposed future research environment infrastructure

7.4 Conclusion

Chapter 7 described the infrastructure required for a prototype to propose a model that can assist in the automated detection of human identity deception on SMPs. The following main components were indicated: hardware, database, software, and network. The hardware was mostly provided by the HPI Future SOC lab and was accessible through an OpenVPN network. The resources in the lab were essential to solve for the research problem, as they provided the necessary storage to gather streaming and historic data from Twitter data and then to store the data required for detecting identity deception in the SAP HANA database. Furthermore, different software was proposed in the infrastructure with which to gather the SMP data, explore the SMP attributes, label SMP accounts, and perform machine learning experiments. Although the Hadoop platform was instrumental in gathering and exploring the data, many options of language application software could be used to perform various machine learning experiments. For this research, the researcher proposed the R language as it was deemed most mature at the time of research in terms of having a uniform interface to over 200 machine learning algorithms.

The next chapters will discuss the results of the implemented prototype, based on the proposed infrastructure.

Part III

The implementation of a prototype
to develop and validate a model for
assisting with the automated
detection of identity deception

Chapter 8

Prototype: Prepare

“You can have data without information, but you cannot have information without data.”- Daniel Keys Moran

8.1 Introduction

Various requirements are expected to be met by a model that assists in the automated detection of human identity deception on Social Media Platforms (SMPs). One of these requirements is, for example, to only use data pertaining to humans, as opposed to bots. The full list of requirements for identity deception detection by humans on SMPs appeared in Chapter 3. The researcher furthermore showed how these requirements can be met by implementing a prototype consisting of three main components – Prepare, Discover and Detect – in a specialised technical research environment. Chapter 8 is part of a three-part series describing the results after implementing each of these prototype components. This chapter specifically discusses the results after implementing the ‘Prepare’ component. Figure 8.1 shows the relationship between the ‘Prepare’ component and the other components by means of a high-level prototype overview.

The ‘Prepare’ component is sub-divided into sub-components as shown in Figure 8.2. Each sub-component has a specific task. The chapter firstly describes the results from gathering data from an SMP, in this case Twitter. Secondly, it presents results from cleaning the data. Thirdly, data from known deceptive accounts is appended to the existing gathered corpus to generate a labelled corpus of known ‘trustworthy’ and ‘deceptive’ accounts. This also includes additional features engineered from knowledge

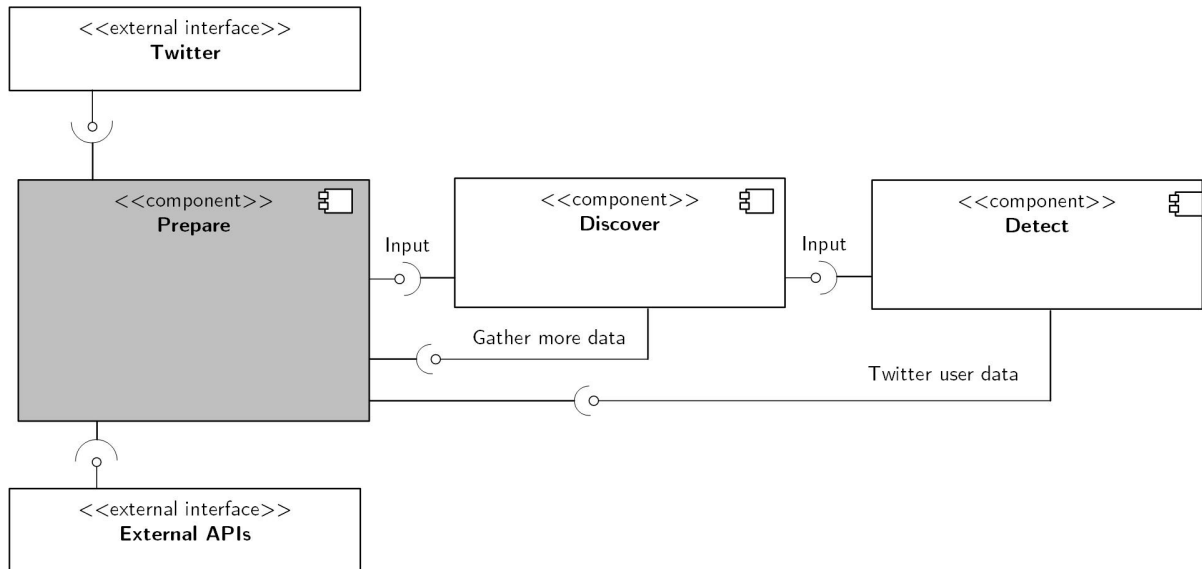


Figure 8.1: High-level overview of the prototype: 'prepare' component

of related research fields such as bot detection and psychology. Lastly, the results from preparing the data for supervised machine learning are presented. Throughout the discussion of these results, it will be evident how the research in hand met the various requirements expected of a model that assists in the automated detection of human identity deception.

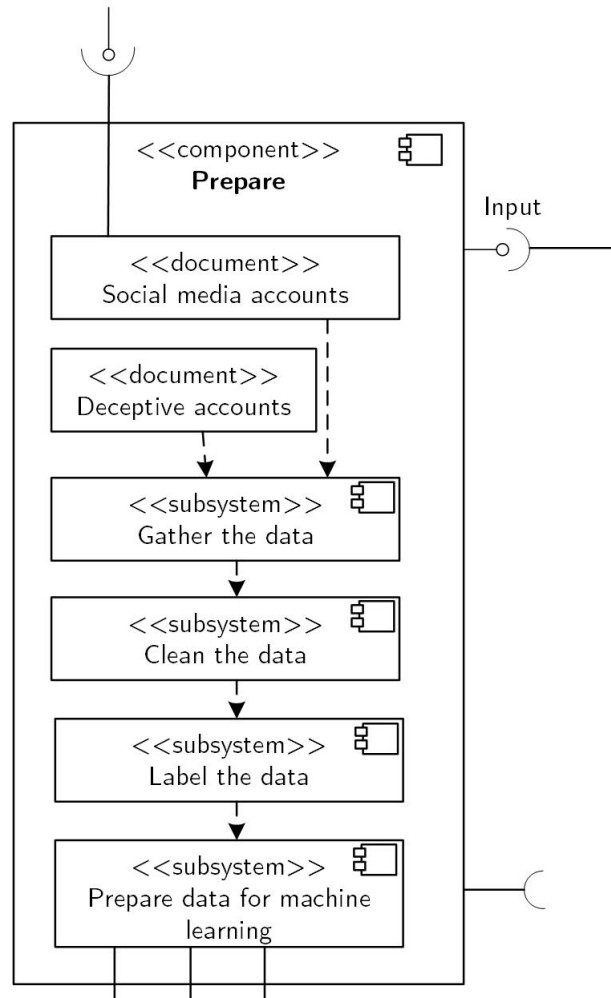


Figure 8.2: The ‘Prepare’ component

8.2 Gather the data

Requirements of a model that assists in the automated detection of human identity deception include, but are not limited to the following:

- The model should detect humans who are deceptive about their identity for malicious purposes.
- The model should use attributes available on SMPs.

Twitter is an example of an SMP platform where human identity deception occurs. The attributes available for a user with an account on Twitter describe the user’s identity, account, behaviour, relationships, and content. Gathering data from the attributes pertaining to the user’s identity and their account on an SMP meets the first requirement. Gathering SMP data from Twitter, as an example of an SMP, also

```

"created_at": "Fri Jan 23 23:57:36 +0000 2015",
"id": 558775589612437504,
"id_str": "558775589612437504",
"text": "Thanks, polar vortex: Attendance dips at major Chicago museums in 2014 http://t.co/jQ1LEurk9s http://t.co/3bss2nGemx",
"source": "\u003ca href='\"http://twitter.com\"' rel='\"nofollow\"'\u003eTwitter Web Client\u003c/a\u003e",
"truncated": false,
"in_reply_to_status_id": null,
"in_reply_to_status_id_str": null,
"in_reply_to_user_id": null,
"in_reply_to_user_id_str": null,
"in_reply_to_screen_name": null,
"user": {
  "id": 7313362,
  "id_str": "7313362",
  "name": "Chicago Tribune",
  "screen_name": "chicagotribune",
  "location": "Chicago, IL",
  "url": "http://www.chicagotribune.com/",
  "description": "Chicago Tribune news, features and so much more live from our newsroom. A part of your life since 1847.",
  "protected": false,
  "verified": true,
  "followers_count": 321548,
  "friends_count": 523,
  "listed_count": 8074,
  "favourites_count": 34,
  "statuses_count": 47367,
  "created_at": "Sat Jul 07 14:10:07 +0000 2007",
  "utc_offset": -21600,
  "time_zone": "Central Time (US & Canada)",
  "geo_enabled": false,
  "lang": "en",

```

Figure 8.3: An example of gathered Twitter data

satisfies the second requirement.

Twitter data is regarded as semi-structured data, as the text or posts from users do not contain a predefined consistent format (unstructured). However, a description exists of all the attributes available on Twitter (structured) [310]. Figure 8.3 contains an example of some gathered Twitter data.

For this research, Twitter data was gathered for a period starting from 2016 for all tweets containing the words ‘school’ and ‘homework’. This allowed the researcher to obtain a corpus of a targeted user demographic, in this case minors. In addition, the tweets for these accounts’ friends and followers were gathered to create a larger corpus still within the same demographic. Humans are believed to be friends and they follow others similar to them [75]. Initially it was found that data was gathered at a rate of 500 000 tweets per day. Optimising the code in February 2016 allowed for an average data-gathering rate of 3 million tweets per day. From March to December, data was only gathered on an *ad hoc* basis. Final optimisations to the code in December 2016, using threads running in parallel, allowed for 75 million tweets to be gathered per day on average. The parallel threads took cognisance of the rate limits imposed by the Twitter Application Program Interface (API) and were paused programmatically when Twitter’s API limits were reached. The mining of tweets over the indicated time period is shown in Figure 8.4. The results show how the rate at which data was gathered improved by optimising code (in February 2016) and running code in parallel (in December 2016).

A total of 606 914 240 tweets were gathered for this research from Twitter. This equates to 200GB of SMP data consisting of 223 796 Twitter users. Of the 223 796

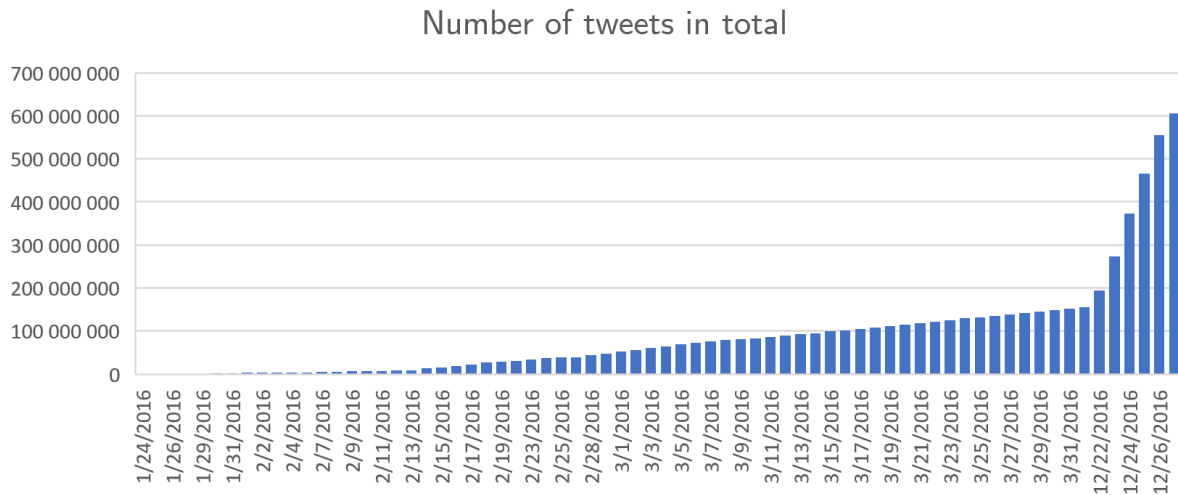


Figure 8.4: Tweets gathered from Twitter

Twitter users, 61 977 were from the original gathered corpus of tweets containing the words ‘school’ and ‘homework’. The remaining 161 819 users were gathered using the relationship information about the original Twitter user’s friends and followers. To further understand the gathered Twitter users, Exploratory Data Analysis (EDA) was performed on the corpus. The results from this EDA are discussed next.

8.2.1 Results from exploratory data analysis

The corpus of 606 million tweets was analysed and revealed the following:

- Most tweets were posted in 2016. Figure 8.5a shows the exponential increase of tweets on a yearly basis. Interestingly, every year the number of tweets was almost double that of the previous year. This pattern shows how the volume of data produced is exponentially growing, which correlates with existing literature on the topic of data volume growth [181].
- Most of the gathered Twitter data came from America. Figure 8.5b shows the number of tweets per time zone. The number of American- (or rather English-) speaking countries was high, since the two words used to obtain the initial corpus – ‘school’ and ‘homework’ – were English.
- Most posts were made during the week. Figure 8.5c shows the number of tweets posted per day of the week. Saturdays and Sundays showed fewer tweets than weekdays. This is expected behaviour on an SMP platform like Twitter [188].
- The tweets were of varied sentiment. Figure 8.5d shows how most sentiments were

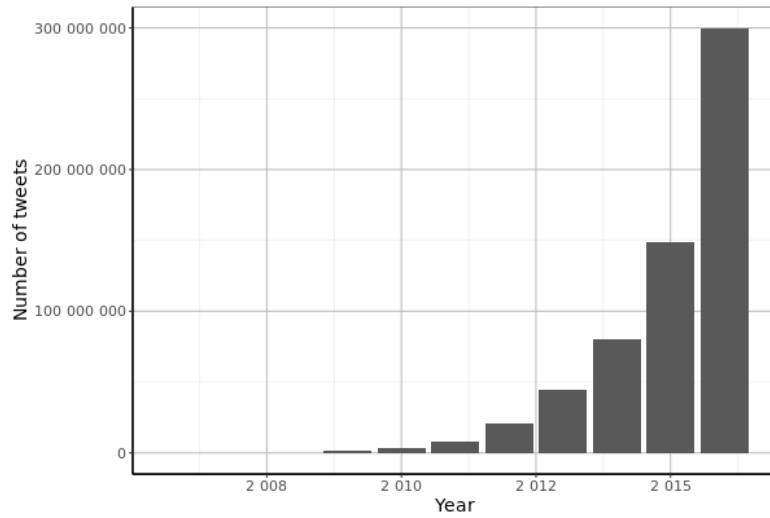
represented in the corpus and suggests that the tweets covered different emotions and (according to the researcher) also different personality types.

- Most tweets originated from users using ten or fewer different platforms to tweet from. Figure 8.5e shows the number of tweets from users per platform. These platforms were not related to a physical device. For example, if tweets were sent by a specific user from a news and a movie application on the same mobile device, this would count as two different platforms in the Twitter API.

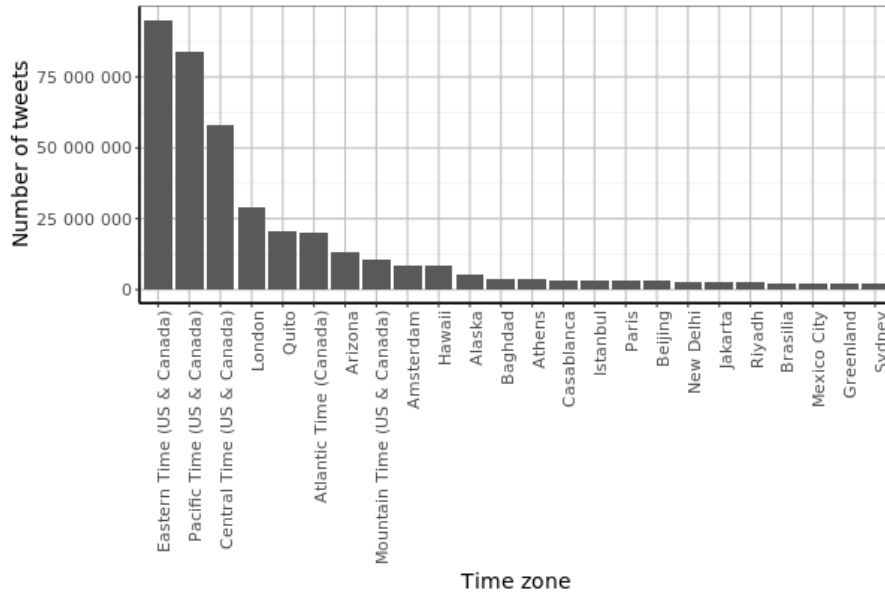
During the EDA of Twitter tweets, normal expected patterns were detected in the gathered Twitter data – i.e. most people were English speaking; more tweets were produced every year; and more tweets were produced during the week. These patterns furthermore suggested that the threat of identity deception was becoming increasingly difficult to detect in the vast amount of SMP data and that automated detection methods were warranted. Although more tweets were produced during the week, it would be dangerous to assume that identity deception attacks by humans are more prevalent on weekdays. The researcher believes that identity deception by humans can occur at any time. An outcome of this EDA was that no patterns could be seen in tweets that could indicate potential identity deception. All data patterns were expected. The results suggest that different methods are required to detect identity deception.

Besides the Twitter tweets, also known as the content, those attributes describing the user's identity and their account were also explored. The EDA results showed the following:

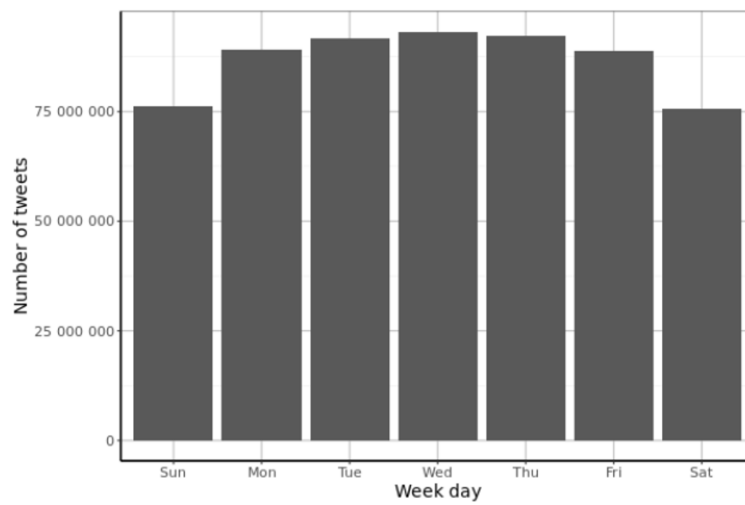
- A user's account name and screen name show very little correlation. The account name is the name they register with on the platform, whereas the screen name is the name that will be displayed to other users. Figure 8.6a shows how screen name lengths are typically shorter than account names. These upper limits are however imposed by the API, where screen names are capped at 15 characters and account names are capped at 20 characters [310]. The researcher did however notice a number of outliers in terms of the account names. A possible explanation for this is that these API restrictions were not always enforced as stated by the API [310].
- Users have fewer friends than followers. Figure 8.6b shows how the number of friends per account is much smaller than the number of followers. This is expected, as many users follow people of interest with whom they are not necessarily friends. An example would be fans of a musician. The fans are not



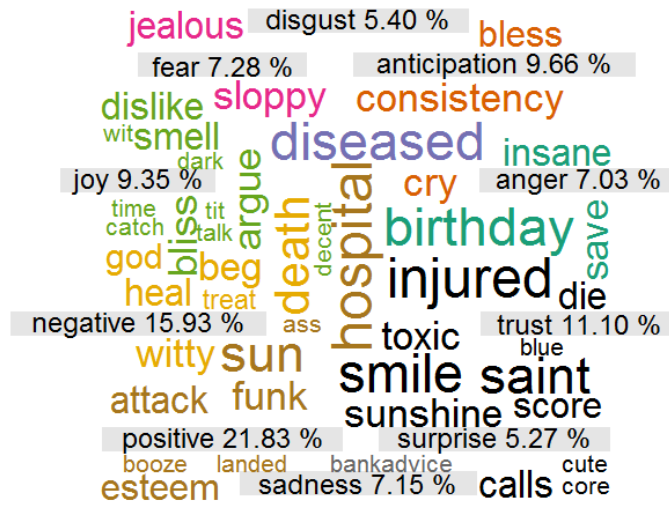
(a) Total number of tweets per year



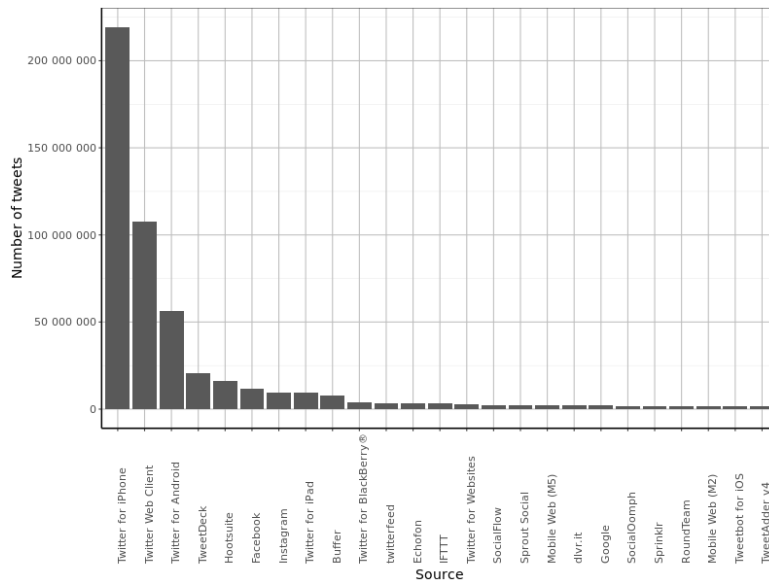
(b) Number of tweets for the top 25 time zones



(c) Number of tweets per week day



(d) Tweet sentiment



(e) Number of tweets for the top 25 platforms

Figure 8.5: Exploration of Twitter content data

friends with the musician, but they want to follow him to know what he is doing and when a new album may potentially be released.

- More users define a location than geo-enable their account. Figure 8.6c shows that the majority of users have a location specified in a free text attribute provided by Twitter. This is in contrast with Figure 8.6d, which shows that users do not always have their geo-location enabled. Once enabled, the geo-location from where every single tweet was sent, is stored. Many users who may not want to be tracked in such a way, choose to disable this feature. However, others do not seem to mind

stating a location. The danger is that the user could lie, seeing that the location field is a free format text attribute. This implies that the location field could be a good indicator of deception if it were possible to compare with another attribute indicating where the user really is.

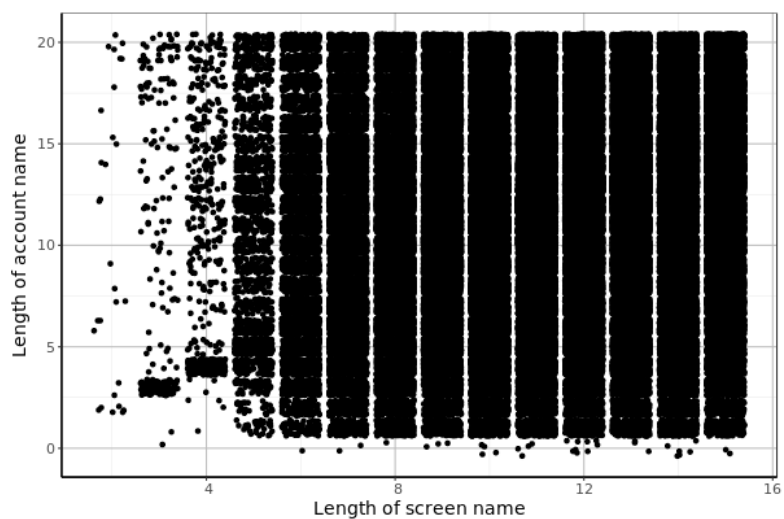
- Account name and screen name lengths follow distinct patterns. Figure 8.6e shows the number of users per account name length, whereas Figure 8.6f shows the number of users per screen name length. Both distributions show distinctly different patterns. Due to these distinct patterns, these attributes can also be regarded as good candidates for potential deceptiveness. If a particular account does not follow this pattern, it could point to potential identity deception.

8.2.2 Findings that emerged from exploratory data analysis

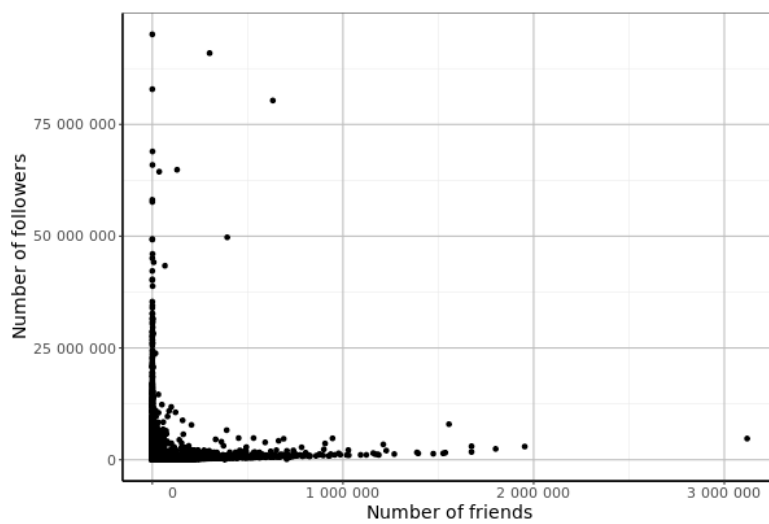
The EDA performed on the gathered corpus highlighted the following:

- The tweets gathered for this research from Twitter presented normal patterns expected from SMP content. This made human identity deception detection difficult from looking at the content as it could be comparable to ‘finding a needle in a haystack’. For the purposes of this research, the content would thus be removed from the corpus.
- Certain attributes did not contribute to the detection of human identity deception, for example, the geo-location of many users was not enabled on Twitter. For the purposes of this research, these attributes were therefore identified and removed from the corpus.
- Knowing the distributions of attributes pertaining to the user’s identity or account could potentially be used to detect outliers. These outliers could be indicative of identity deception.
- An automated model would be required to detect human identity deception due to the voluminous nature of the data.

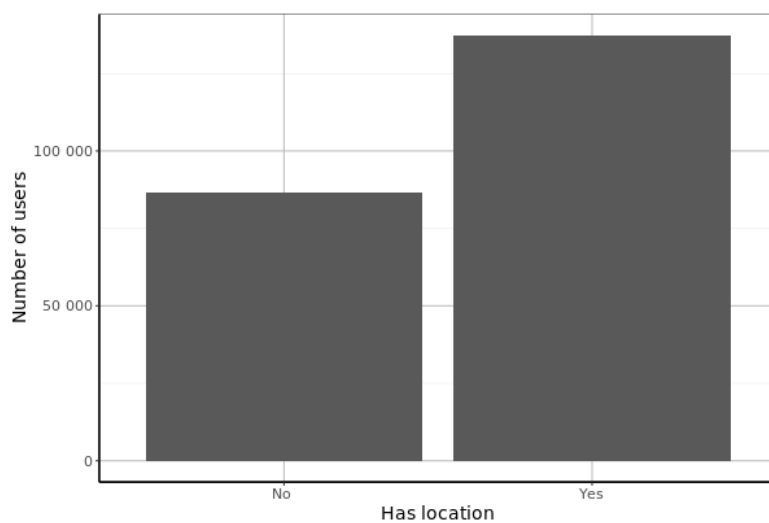
By transforming the data, these points can be addressed. The next sections will discuss the various sub-components defined as part of the ‘prepare’ component to this effect.



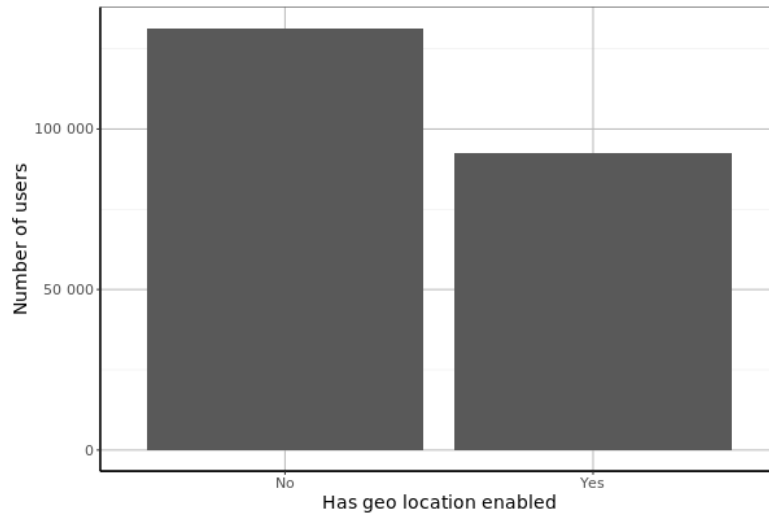
(a) User account name vs screen name



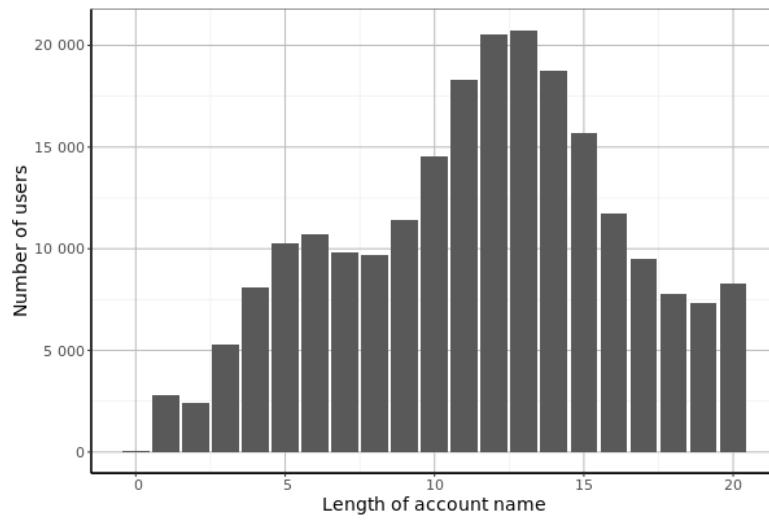
(b) Friends vs followers



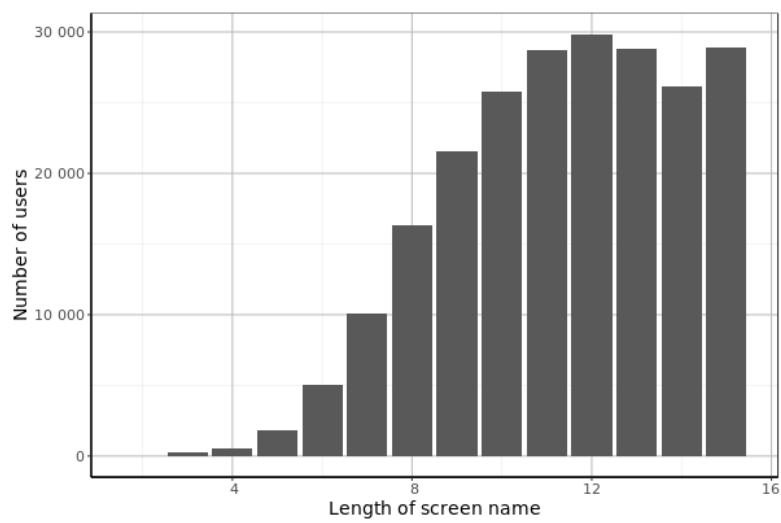
(c) User accounts with a location



(d) User accounts with a geo-location



(e) User account name length



(f) User account screen name length

Figure 8.6: Exploration of Twitter user data

8.3 Clean the data

Requirements of a model that assists in the automated detection of human identity deception include, but are not limited to the following:

- The model should ignore content by users on SMPs.
- The model should disregard non-human accounts.
- The model should ignore attributes that do not contribute to human identity deception detection.

These requirements can be addressed by cleaning the data. Data cleaning allows for the corpus to contain only the specific data required by a model that can assist in the automated detection of human identity deception on SMPs.

Data cleaning can either mean to get rid of users (complete rows or data entries) or columns (attributes) as illustrated in Figure 8.7. Given this, the removal of non-human accounts will be achieved by getting rid of data entries, whereas irrelevant attributes such as content will be removed from the corpus. The results of cleaning the data for the purposes of this research are discussed next.

8.3.1 Disregard non-human accounts

Past research proposes a range of strategies for detecting non-human accounts or bots. For the prototype, the researcher looked at work specifically presented by Cresci et al. [80]. Their work compared various bot detection studies and resulted in a list of rules that could detect bots. For the purposes of data cleaning, the researcher picked the top three rules presented in their work as they demonstrated an accuracy of over 60%. According to their rules, human accounts present at least one of the following traits:

- They have more than 30 followers.
- They have more than 50 tweets.
- At least one other user is mentioned in their tweets.

By applying these rules to the corpus gathered for this research, 53 091 out of 224 796 user accounts could be classified as non-human.

ID	NAME	SCREENNAME	CREATED	ORIGINAL_PROFILE_IMAGE	PROFILE_IMAGE
390	Marco Duran	tinyfrog537	3/6/2016		https://randomuser.me/api/portraits/men/99.jpg
391	Zachary Thompson	silvertiger573	12/12/2011		https://randomuser.me/api/portraits/men/77.jpg
392	Jake Myers	redostrich621	10/12/2013		https://randomuser.me/api/portraits/men/37.jpg
393	Minea Jokela	greenwolf661	2/19/2012		https://randomuser.me/api/portraits/women/3...
394	Elmer Lowe	bluekoala436	11/12/2005		https://randomuser.me/api/portraits/men/14.jpg
395	Onni Lepisto	organicswan...	6/26/2010		https://randomuser.me/api/portraits/men/65.jpg
396	Samuel Guillot	lazyfrog645	2/20/2003		https://randomuser.me/api/portraits/men/74.jpg
397	Lukas Brunet	whitebear759	6/4/2009		https://randomuser.me/api/portraits/men/32.jpg
398	MaTlia Andre	lazygoose477	4/16/2007		https://randomuser.me/api/portraits/women/3...
399	Andy Lee	redfish603	8/23/2014		https://randomuser.me/api/portraits/men/91.jpg
400	Sarah Richards	bigtiger639	10/25/2004		https://randomuser.me/api/portraits/women/8...
401	Peppi Hatala	greenpanda...	8/27/2011		https://randomuser.me/api/portraits/women/0...
402	Lutz Baier	tinypeacock...	1/15/2010		https://randomuser.me/api/portraits/men/68.jpg
403	Moshe Roovers	silverladybu...	1/20/2013		https://randomuser.me/api/portraits/men/2.jpg
404	Thea Thomsen	yellowladyb...	4/29/2015		https://randomuser.me/api/portraits/women/1...
405	Yasemin Tahincio...	goldenleopa...	3/18/2014		https://randomuser.me/api/portraits/women/4...
406	Lillian Boyd	bigfrog696	5/12/2013		https://randomuser.me/api/portraits/women/2...
407	Phyllis Larson	orangebutte...	2/23/2016		https://randomuser.me/api/portraits/women/1...
408	Bill Curtis	organictiger...	10/2/2011		https://randomuser.me/api/portraits/men/97.jpg
409	Clarissa Ritter	whitepeacoc...	6/1/2008		https://randomuser.me/api/portraits/women/8...
410	Gabriel Ambrose	smallkoala799	4/10/2008		https://randomuser.me/api/portraits/men/50.jpg
411	Auzust Jensen	tinvhird990	1/28/2004		https://randomuser.me/api/portraits/men/78.jpg

Figure 8.7: Data cleaning: removing entries vs attributes

In addition to removing all non-human accounts, Twitter can also verify accounts as being trustworthy. It is up to the account holder to request this and verification is usually done in the case of celebrities [310]. 17 188 user accounts were found to have been verified by Twitter and were removed for the purposes of this research. Figure 8.8 shows the distribution of the remaining and original corpus, split according to the year when the account was created. It was interesting to note the rise in cleaned accounts since 2010. This was expected, as bot accounts and celebrities have become more prevalent on SMPs over the last few years [233]. Figure 8.8 also shows that bot accounts have been active on Twitter since the SMP's inception in 2006.

8.3.2 Remove irrelevant attributes

Supervised machine learning algorithms expect data to be available as they cannot infer missing content. The best a machine learning algorithm can do in such scenarios is to assign a default value to missing attribute values or to ignore these attributes completely [337]. Missing values introduce bias into the models. For example, consider user accounts where the location attribute is empty for all trustworthy accounts, but deceptive accounts have a location. Machine learning algorithms could use this fact to infer that all trustworthy accounts have no location and that this fact distinguishes the trustworthy accounts from the deceptive accounts. Such a deduction would be

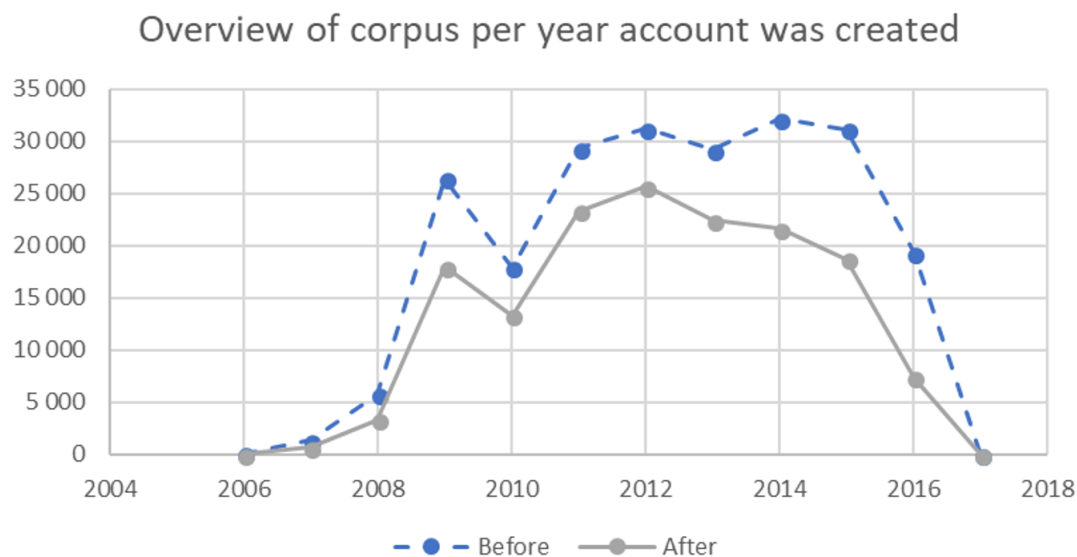


Figure 8.8: Corpus before and after data cleaning

incorrect. Therefore, the attributes containing missing values that could lead to incorrect human identity deception detection should be identified and removed.

Unique values are also a problem for machine learning algorithms. If there is only one distinct value for a specific attribute, that attribute does not contain enough information about the class the machine learning model is trying to predict and therefore it can be omitted. If the attribute values are completely unique, the attributes introduce variance as the algorithm might use the specific individual values to learn whether an account is deceptive or not – rather than to generalise a rule with which to detect the deception.

Lastly, the EDA performed on both the Twitter tweets and Twitter user data implied that attributes, like friends and followers, are correlated. Correlation should be considered for this research, as the prototype makes use of supervised machine learning to build a model that can detect human identity deception on SMPs. For supervised machine learning, correlated attributes can be omitted when there is a strong correlation between them [349]. The reason for this is that one attribute is good enough to learn the data distribution and additional correlated attributes would only increase the time the machine learning algorithm requires to build a human identity deception detection model.

The results after cleaning the data from missing, unique and correlated values are discussed next.

8.3.2.1 Removing missing and unique attribute values

Table 8.1 shows the number of missing and unique values found in the identity attributes gathered for the Twitter accounts. Some attributes (indicated in grey in the table) were removed from the corpus, based on the following decisions:

- For the results pertaining to missing values in the corpus, the following attributes could be removed as they were found to be missing and would introduce bias: `LATITUDE` and `LONGITUDE`. However, the researcher decided to keep these values, as it emerged from the field of psychology that location shows promise in indicating deception and the researcher wanted to have these values available for further exploration or feature engineering later.
- For the results presented on unique values in the corpus, the following attributes were removed as they were found to be the same for most attributes and could introduce bias: `CREATED`, `GEO_ENABLED`, `AND IS_DEFAULT_PROFILE`, `IS_DEFAULT_PROFILE_IMAGE`, AND `IS_BACKGROUND_IMAGE_USED`.
- For the results presented on unique values in the corpus, the following attributes were removed as they were found to be distinct and could introduce variance: `ID`, `NAME`, `SCREENNAME`, `ORIGINAL_PROFILE_IMAGE`, `PROFILE_IMAGE`, AND `DESCRIPTION`. However, the researcher decided to keep `PROFILE_IMAGE`, `NAME`, and `SCREENNAME`, as the research field of psychology showed that image and name show promise in indicating deception and the researcher wanted to have these values available for further exploration or feature engineering later.

8.3.2.2 Removing highly correlated attributes

In terms of correlation, Pearson's correlation coefficient was used as it has been used before in similar identity deception research [160] [7]. This coefficient produces a result between -1 and 1, with -1 showing that two attributes have a strong negative relationship and 1 showing that they have a strong positive relationship. An example of a strong positive relationship would be that if a person has an original profile image other than the default, the same person also tends to have a current profile image other than the default. Figure 8.9 shows the correlations in the form of a matrix for all attributes available to describe the identity of the user or their account on Twitter. The bigger the circle in the figure, the stronger the relationship. For this research, only correlations

Table 8.1: Missing and unique values in the attributes of Twitter accounts

Twitter attribute name	Number of missing values	Number of unique values
Identifier (ID)	-	154 417
NAME	-	154 417
SCREENNAME	-	154 417
CREATED	-	1
ORIGINAL_PROFILE_IMAGE	5 410	148 549
PROFILE_IMAGE	-	153 949
BACKGROUND_IMAGE	6 175	54 539
DESCRIPTION	-	154 417
LOCATION	1	48 473
LANGUAGE	-	56
FRIENDS_COUNT	-	16 411
FOLLOWERS_COUNT	-	30 077
STATUS_COUNT	-	38 130
LISTED_COUNT	5 410	2 584
TIMEZONE	60 169	231
UTC_OFFSET	5 410	34
GEO_ENABLED	-	2
LATITUDE	153 690	804
LONGITUDE	153 690	804
IS_DEFAULT_PROFILE	-	2
IS_DEFAULT_PROFILE_IMAGE	5 410	3
IS_BACKGROUND_IMAGE_USED	5 410	3
PROFILE_TEXT_COLOR	5 410	8 152
PROFILE_BG_COLOR	5 410	13 496

with a coefficient of 1 and -1 were considered for removal, as these attributes do not add any value to a model that proposes to detect human identity deception on SMPs.

Based on the correlation matrix in Figure 8.9, a relationship is present between ORIGINAL_PROFILE_IMAGE and PROFILE_IMAGE. As mentioned earlier, once a user has chosen a profile image other than the default, they tend to not go back to the original default image provided by Twitter. A negative correlation is furthermore highlighted between LATITUDE and LONGITUDE. This is not generally expected in the real world, but rather an effect of the corpus that had been gathered for this research. The EDA showed that the corpus mainly consisted of English-speaking users and that most were from English-speaking countries like America. This limited the corpus to only a subset of geo-locations. These correlations were however not strong and therefore the matrix identified no additional attributes to be removed.

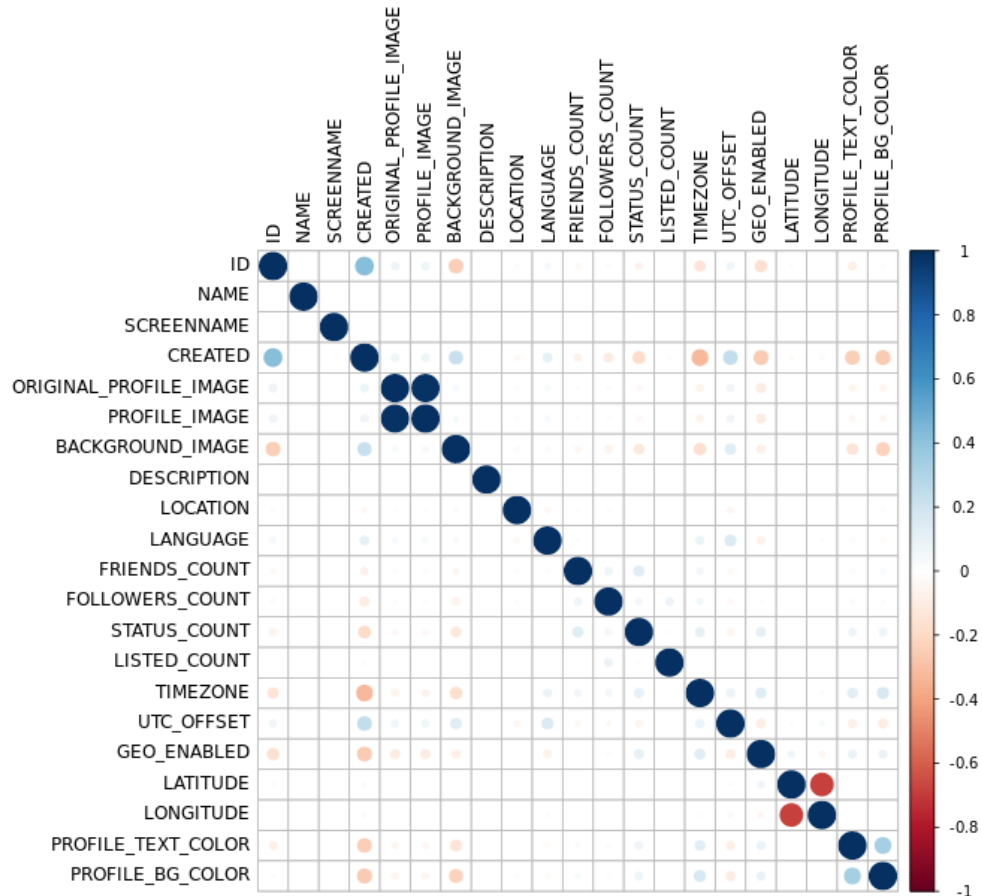


Figure 8.9: The correlation between attributes found on Twitter describing a user’s account or identity

The correlation matrix is nonetheless a method that can be used to validate that the cleaned data does not introduce new correlated relationships into the corpus. Figure 8.10 shows a view of all remaining attributes and their correlation with each other. From this new correlation matrix, it is clear that the remaining attributes show very little correlation. The highest correlation is shown between FRIENDS_COUNT and FOLLOWER_COUNT, with a value of 0.275. This is not a strong enough correlation (value of 1.0) to warrant their removal from the corpus.

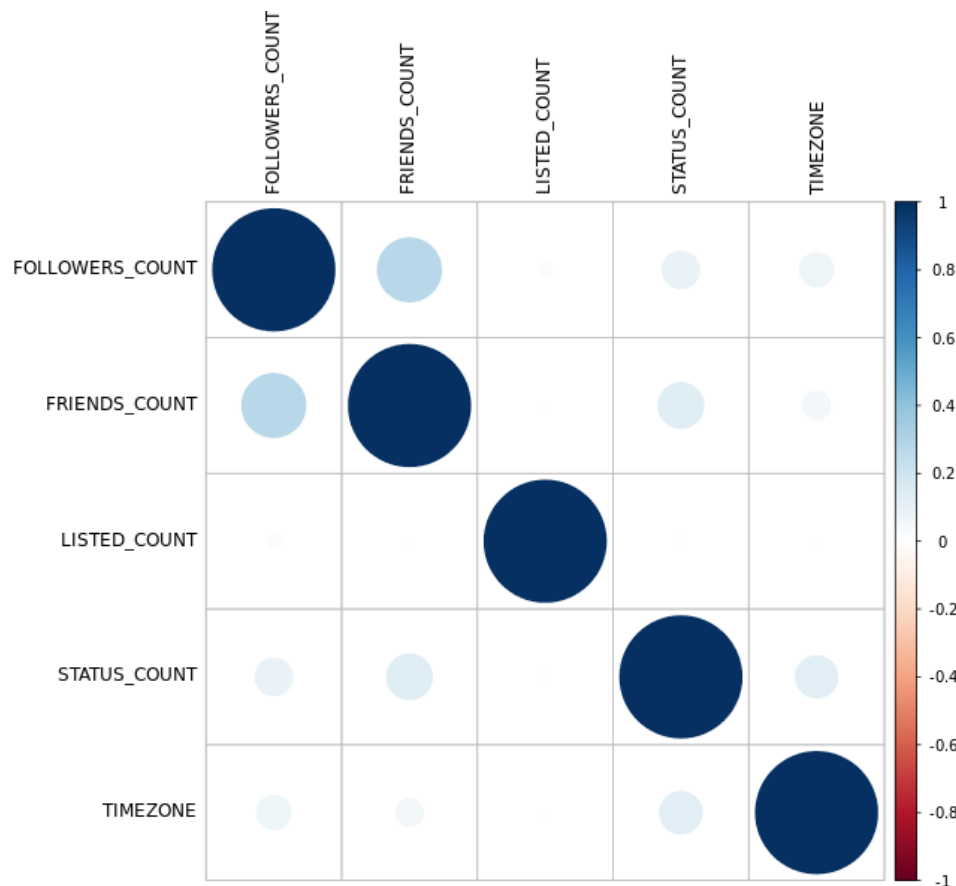


Figure 8.10: The correlation between attributes after data cleaning

8.4 Label the data

After it has been cleaned, the data should be labelled, as supervised machine learning algorithms expect labelled data [190]. For the current research, it was assumed that most data in Twitter is trustworthy, seeing that research has confirmed that most people do not lie [145]. Therefore all 154 417 accounts were labelled as being trustworthy. For deceptive accounts, additional accounts were appended to the original corpus. Requirements of a model that assists in the automated detection of human identity deception include but are not limited to the following:

- Attributes and engineered features should be indicative of deception rather than the truth.
- Features should be engineered such that they complement the detection of identity deception.

Keeping these requirements in mind, the generation of new deceptive accounts is

discussed next. This will be followed by the addition of new engineered features to complement the detection of identity deception, based on the knowledge from past research work on bots and psychology.

8.4.1 Generating deceptive accounts

To generate or append deceptive examples to the corpus, the researcher considered the following options:

- Labelling the Twitter data by using a mechanism such as crowd sourcing (i.e. paying people to manually perform some repetitive tasks) [350]. For this research, for example, people could have been given the original corpus and asked to manually inspect each account for deceptiveness.
- Finding a labelled corpus in addition to gathering the data from Twitter. In this case, the researcher could have approached other researchers for their data solving for a similar problem or used an example of such a deceptive dataset that is available to the public on the Internet.
- Generating deceptive accounts manually and appending them to the corpus. In this instance, tools exist to generate deceptive accounts. These tools could be used to generate user accounts with similar attributes expected from the gathered Twitter corpus.

Crowd sourcing proved to be a problem due to three reasons: The volume of data involved; the fact that people’s decisions on whether an account was deceptive (or not) was based on their own subjective perception; and the fact that there was no guarantee that deceptive accounts actually existed in the original corpus gathered. In terms of existing labelled data, none could be found on the Internet or in past research that either reflected the same attributes as Twitter or that demonstrated examples of human identity deception on SMPs. For these reasons, the researcher chose to generate deceptive accounts manually.

8.4.1.1 Approach adopted to generate example accounts

Two APIs were used to generate deceptive example SMP accounts and create a complete deceptive dataset. These were the ‘Generate data’ API [178], shown in Figure 8.11, and ‘Random User generator’ API [25], shown in Figure 8.12. Both these APIs can be used

```

Generated 100 of 100 results
cancel
1 NAME|EMAIL|STREET|LAT_LON
2 Palmer|felis.orci@Ouisami.ca|389-2409 Aliquet St.|40.97977, 32.62631
3 Martina|mauris.rhonus@est.edu|177-2754 Eu Street|-6.09447, 160.54747
4 Heather|tristique.senectus@Donecat.edu|Ap #186-8882 Nunc Street|-21.60494, -173.25437
5 Gray|risus.varius@infelis.com|Ap #613-952 Nam Ave|-57.22051, -152.3441
6 Dorothy|ornare@ultriciesadipiscingenim.net|P.O. Box 994, 8143 Donec Av.|-69.4152, -89.65437
7 Melissa|sem.magna@ut.net|881-4665 At, Rd.|-84.35562, -103.01921
8 Brett|mauris.Integer@Curabituremassa.com|433-957 Velit Avenue|28.90513, -131.82476
9 Trevor|Nunc@egestasadui.net|325-1758 Ridiculus Avenue|-41.15132, 81.32097
10 Jarrod|Nunc@magnanecquam.net|288-1879 Tincidunt St.|51.71359, 132.85227
11 Luke|velit.eu@scelerisqueui.edu|2947 Turpis. St.|30.00012, 10.63656
12 Zelenia|libero@morbitristique.edu|634-4516 Magna Rd.|-30.816, 170.94652
13 Victor|cursus@Fuscudolor.net|810-5095 Suspendisse St.|-26.35336, -65.29268
14 Audra|nibh.Quisque@urnaUt.net|659-9952 Ligula Street|42.61115, -82.93037
15 Neville|scelerisque@imperdietnonvestibulum.co.uk|5869 Quisque St.|-65.29304, -14.94128
16 Teegan|ut.pellentesque.eget@lorem.co.uk|Ap #766-5121 Justo Av.|-45.03725, 155.21443
17 Selma|mi@consecteturcursuset.edu|483-4237 Dui. Road|-21.26487, -162.3893
18 Rose|Duis.dignissim@scelerisque.edu|P.O. Box 508, 5529 Nibh Av.|-70.79174, 132.14487
19 Luke|pretium.et.rutrum@egestasAliquamfringilla.co.uk|4954 Cursus Avenue|-19.19459, 87.15455
20 Halla|dictum.placerat.augue@magnaLorem.ca|7141 Elit, Rd.|26.32187, -58.23277
21 Jasmine|neque.vitae@aliquetnecimperdiet.ca|P.O. Box 787, 635 Dictum Rd.|-14.99761, -88.27216
22 Dante|suscipit@risusMorbimetus.net|992-4548 Magna. Ave|-66.9571, -43.33596
23 Kuame|eleifend.nunc.risus@non.edu|Ap #385-7967 Metus. Av.|23.08507, -108.952
24 Kylee|Nulla.semper@estNunclaoret.co.uk|4292 Mi Rd.|74.97628, -76.09619
25 Shelby|pellentesque.a.facilisis@nislQuisque.co.uk|Ap #123-9708 Ullamcorper, Ave|-21.59307, 90.84362
26 Fiona|mollis.non@sapien.org|2190 Quam. Ave|13.22594, -98.61665
27 Tatiana|in@luctusaliquet.net|P.O. Box 336, 2686 Arcu Avenue|4.85274, 139.49421
28 Nolan|in.consectetur.ipsam@ametluctus.edu|Ap #968-5224 A Ave|41.85476, 58.72218
29 Indira|Ut@nibhAliquam.com|5158 Ullamcorper. Av.|15.24975, 117.2609
30 Rufus|sem.magna@ut.net|881-4665 At, Rd.|-84.35562, -103.01921

```

Figure 8.11: Example results for the ‘Generate data’ API

for free from the Internet, but each has certain limitations. The ‘Generate data’ API limits the generation of records to 100 at a time and is unable to provide examples of fake profile images. On the other hand, the ‘Random User generator’ API allows for a maximum of 5 000 accounts at a time but is unable to provide geo-locations for these fake generated users.

The data from both APIs was combined to create a deceptive example dataset of user accounts. Table 8.2 shows how the attributes of each SMP were generated from the APIs to create a deceptive example dataset. It is also worthy to note that certain attributes were generated without any API, because it was not available in either API. The technique that was used to generate these manual values is indicated in Table 8.2 in brackets.

Lastly, Table 8.3 shows an example of one such deceptive generated account. This account is perceived as deceptive due to the following reasons:

- The NAME and SCREENNAME are unrelated.
- The LATITUDE and LONGITUDE intersect somewhere over Somalia and the

```

{
  "results": [
    {
      "gender": "male",
      "name": {
        "title": "mr",
        "first": "romain",
        "last": "hoogmoed"
      },
      "location": {
        "street": "1861 jan pieterszoon coenstraat",
        "city": "maasdriel",
        "state": "zeeland",
        "postcode": "69217"
      },
      "email": "romain.hoogmoed@example.com",
      "login": {
        "username": "lazyduck408",
        "password": "jokers",
        "salt": "UGtRFz4N",
        "md5": "6d83a8c084731ee73eb5f9398b923183",
        "sha1": "cb21097d8c430f2716538e365447910d90476f6e",
        "sha256": "5a9b09c86195b8d8b01ee219d7d9794e2abb6641a2351850c49c309f1fc204a0"
      },
      "dob": "1983-07-14 07:29:45",
      "registered": "2010-09-24 02:10:42",
      "phone": "(656)-976-4980",
      "cell": "(065)-247-9303",
      "id": {
        "name": "BSN",
        "value": "04242023"
      },
      "picture": {
        "large": "https://randomuser.me/api/portraits/men/83.jpg",
        "medium": "https://randomuser.me/api/portraits/med/men/83.jpg",
        "thumbnail": "https://randomuser.me/api/portraits/thumb/men/83.jpg"
      },
      "nat": "NL"
    }
  ],
  "info": {
    "seed": "2da87e9305069f1d",
    "results": 1,
    "page": 1,
    "version": "1.1"
  }
}

```

Figure 8.12: Example results for the ‘Random User generator’ API

location is not Neerijnen, Netherlands, as suggested by the LOCATION attribute.

- Neerijnen’s UTC offset is actually 7 200 and not -28 800 as suggested.
- The PROFILE IMAGE is not one that represents Chris Fuller.

Table 8.2: Origin of appended deceptive dataset attributes

Attribute name	Where was the data generated from?
CREATED	Manual (use a random date from the population)
ID	Manual (use a random number from the population)
NAME	Random User generator API
SCREENNAME	Random User generator API
PROFILE IMAGE	Random User generator API
LOCATION	Generate data API
LANGUAGE	Manual (constant 'English')
FRIENDS COUNT	Manual (use a random number from the population)
FOLLOWERS COUNT	Manual (use a random number from the population)
STATUS COUNT	Manual (use a random number from the population)
TIMEZONE	Manual (use a random time zone from the population)
UTC OFFSET	Manual (use a random UTC offset from the population)
LATITUDE	Generate data API
LONGITUDE	Generate data API

Table 8.3: Example of a generated deceptive account

Attribute name	Attribute value
NAME	Chris Fuller
SCREENNAME	Purpledog887
CREATED	1/13/2008
PROFILE IMAGE	https://randomuser.me/api/portraits/men/99.jpg
LOCATION	Neerijnen
LANGUAGE	en
FRIENDS COUNT	140
FOLLOWERS COUNT	2445
STATUS COUNT	10 670
LISTED COUNT	40
TIMEZONE	Eastern Time (US & Canada)
UTC OFFSET	-28 800
LATITUDE	9.65812
LONGITUDE	-50.20248

Table 8.4: Testing the validity of the introduced deceptive SMP accounts

Attribute name	Mann-Whitney U test (p-Value)	Chi-Squared test (p-Value)
PROFILE IMAGE	0.05	0.09
LANGUAGE	0.90	1.00
FRIENDS COUNT	1.00	1.00
FOLLOWERS COUNT	0.90	1.00
STATUS COUNT	1.00	1.00
LISTED COUNT	1.00	1.00
TIMEZONE	0.90	1.00
UTC OFFSET	0.80	0.20

8.4.1.2 Confirming the validity of the generated example accounts

To ensure that the appended deceptive accounts would not introduce bias by themselves, two statistical tests were performed, namely the Mann-Whitney U and Chi-Squared (goodness-of-fit) tests. Both tests indicate whether one can accept the hypothesis that the sample – in this case the generated deceptive SMP accounts – is representative of the population of gathered trustworthy SMP accounts. Results from the tests are shown in Table 8.4. For this research, a 95% significance level was chosen, as this is common in various similar research studies [112] [26]. All results showed a p-value, or significance, of more than 0.05. Therefore, the hypothesis is accepted that the sample data introduced was representative of the population.

To further illustrate that the sample data was representative of the population, distribution graphs were created comparing all attributes added to the population. These comparisons are shown in Figure 8.13. The distributions drawn for each of the attributes per deceptive and trustworthy class show that they are similar. Therefore, the data for the trustworthy and deceptive SMP accounts was drawn from the same population and the appended deceptive example accounts are thus representative of the population.

8.4.1.3 Testing the correlation of the labelled corpus

As a final test, a correlation matrix was created to show the correlation between attributes within the corpus, given the addition of 15 000 deceptive user accounts. Figure 8.14 shows the correlation matrix with the addition of the CLASS attribute

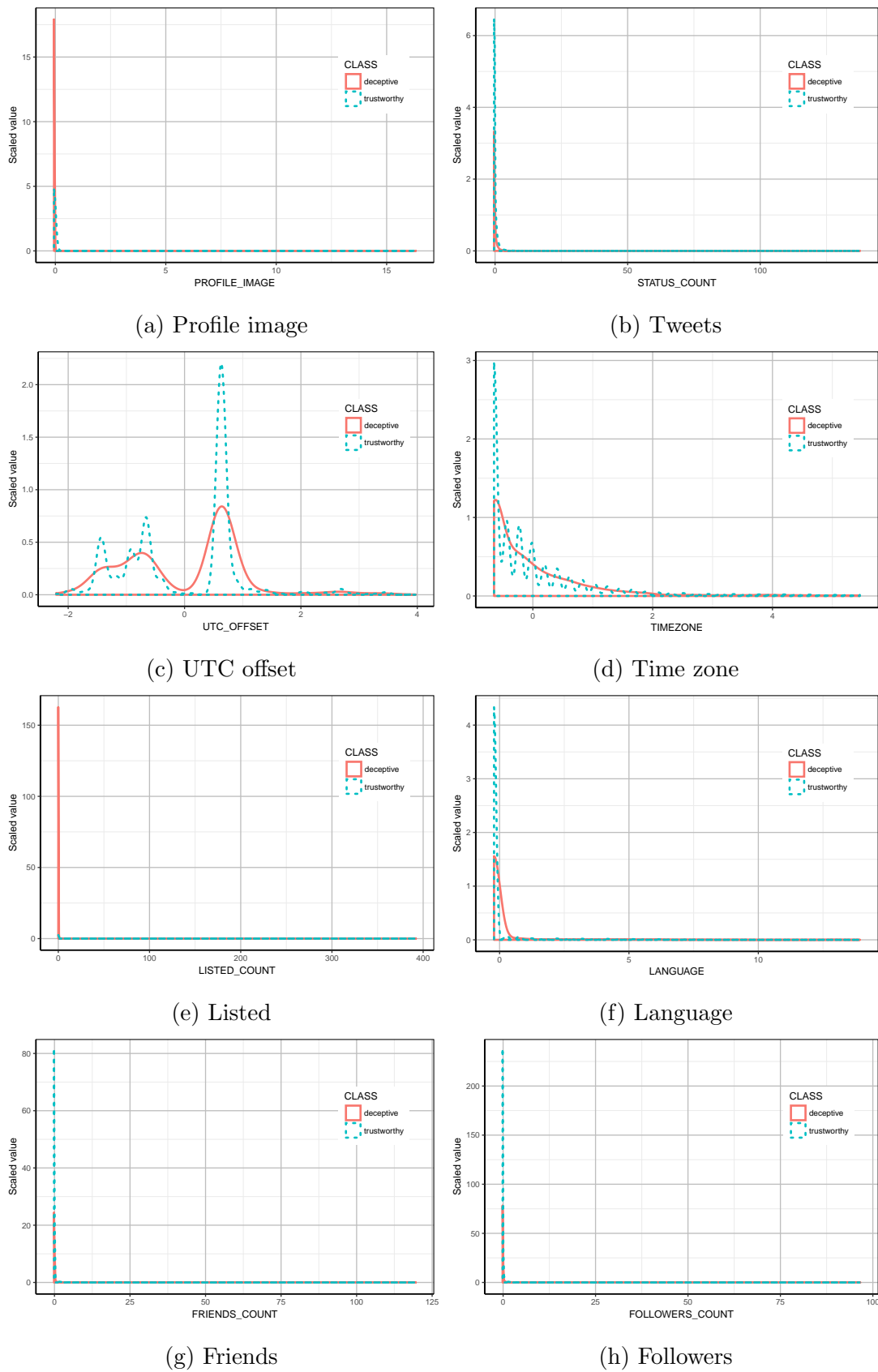


Figure 8.13: Comparing the appended accounts to the original Twitter user account corpus



Figure 8.14: The correlation between attributes after data labelling

containing the label of the user account. For this research, the label indicated whether the user account was deceptive or trustworthy. The correlation matrix shows no changes in the correlation between attributes from the cleaned corpus results presented in Figure 8.10.

8.5 Engineer features

Besides labelling the corpus by adding example generated deceptive accounts, new engineered features that were indicative of deception were introduced to the corpus. Chapter 4 described those features indicative of deception, based on related research work from bots and psychology. These additional features were engineered for each of the user accounts – deceptive and trustworthy – in the corpus. To confirm that these engineered features did not introduce unexpected correlations, the correlation matrix was used. Figure 8.15 shows the result of the new set of attributes and engineered features in the corpus. `COMPARE_AGE` and `COMPARE_GENDER` indicate a correlation of 0.511 and `HAS_PROFILE` shows a correlation of -0.914 with `DUP_PROFILE`. Since neither of these attribute sets were yet indicative of a strong

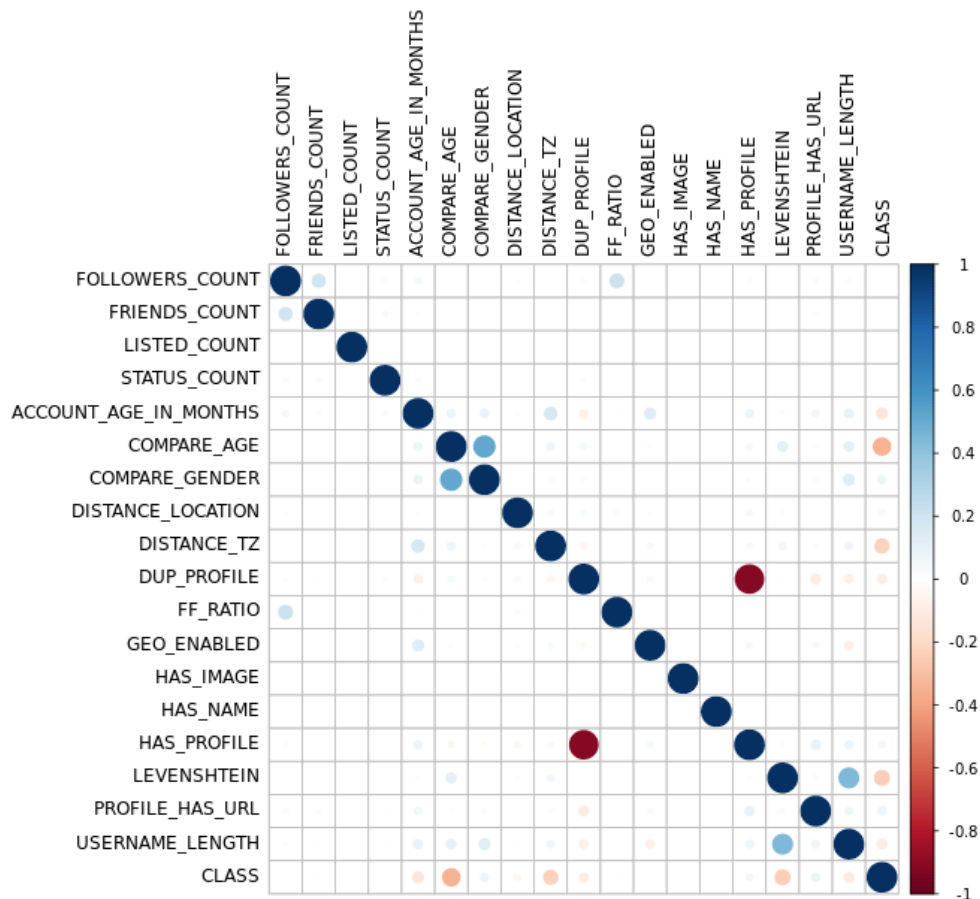


Figure 8.15: The correlation between attributes after feature engineering

relationship, they were both kept for the purposes of the research.

8.6 Prepare the data for machine learning

At this point, the data is in a state where it can be used to assist in developing a model for the automated detection of identity deception by humans on SMPs. However, supervised machine learning in itself has requirements for the data it uses as input to build such an identity deception detection model. Further engineered features can be added to fulfil the requirement that the data used for the model should complement the detection of human identity deception.

Statistically it has been shown that grouping continuous values in smaller bins has a better outcome for predictive modelling [124]. This technique is also referred to as binning or discretising the data [124]. Besides discretising, it is suggested that values be centred and scaled for supervised machine learning [233]. Centring and scaling the

values ensure that bias is not introduced into the predictive models. An example of bias introduced would be where one attribute contains values much larger than another so that it would be incorrectly regarded as more important. The results from these proposals are discussed next.

8.6.1 Discretisation of the attributes

With discretisation, data was binned to ensure that any continuous variable had at most 30 unique values. To achieve this result, different strategies were followed for different data types:

- For text, the top 29 values within an attribute were assigned their own bins and the remainder were added to the last, or 30th, bin. Figure 8.16 shows the results from binning LOCATION, excluding the bin that contains the remainder. It is clear that many of these locations are from traditional English-speaking towns, as was expected. There is also a location called ‘Worldwide’, which was not expected. This is indicative of a free-format type attribute, like LOCATION, where users can insert any value they want without any validation from the SMP.
- For numeric values, the values were divided by 500 and rounded to the nearest integer. Figure 8.17, which shows the results from binning FOLLOWERS COUNT, suggests that very few SMP users have a lot of friends. This fact was also indicated in earlier EDA.

8.6.2 Centring and scaling of the attributes

Centring ensured that each attribute in the dataset had a mean of 0. Scaling, on the other hand, ensured that each attribute in the dataset had a Standard Deviation (STD) of 1. Table 8.5 shows the final results, confirming that all attributes had been centred and scaled.

8.6.3 Testing the correlation of the prepared corpus

Finally, after the data had been discretised, centred, and scaled, a correlation matrix was created to confirm that new correlations had not been introduced during this process. The coefficients were for example the same as after the engineered features had been introduced to the labelled dataset. This result is illustrated in Figure 8.18. Therefore,

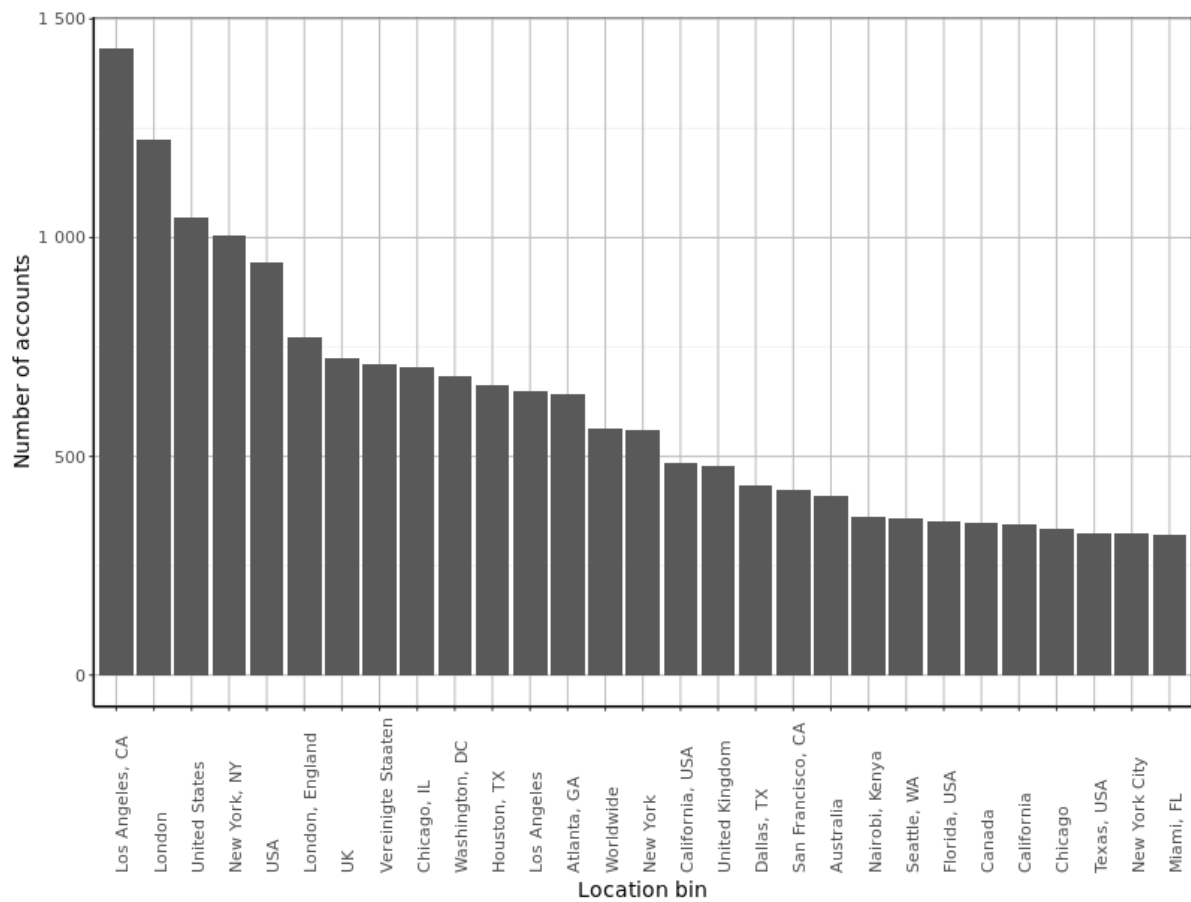


Figure 8.16: Results from binning the location

Table 8.5: The mean and standard deviation of all attributes

Twitter identity attribute	Min	Max	Mean	Median	Standard Deviation
PROFILE_IMAGE	-0.061	16.320	0	-0.061	1
LOCATION	-0.323	7.128	0	-0.323	1
LANGUAGE	-0.214	13.920	0	-0.214	1
FRIENDS_COUNT	-0.159	119.600	0	-0.140	1
FOLLOWERS_COUNT	-0.137	96.590	0	-0.133	1
STATUS_COUNT	-0.383	137.600	0	-0.288	1
LISTED_COUNT	-0.005	392.100	0	-0.005	1
TIMEZONE	-0.648	5.458	0	-0.438	1
UTC_OFFSET	-2.207	3.975	0	0.626	1
LATITUDE	-24.240	20.160	0	-0.023	1
LONGITUDE	-20.050	16.780	0	0.037	1

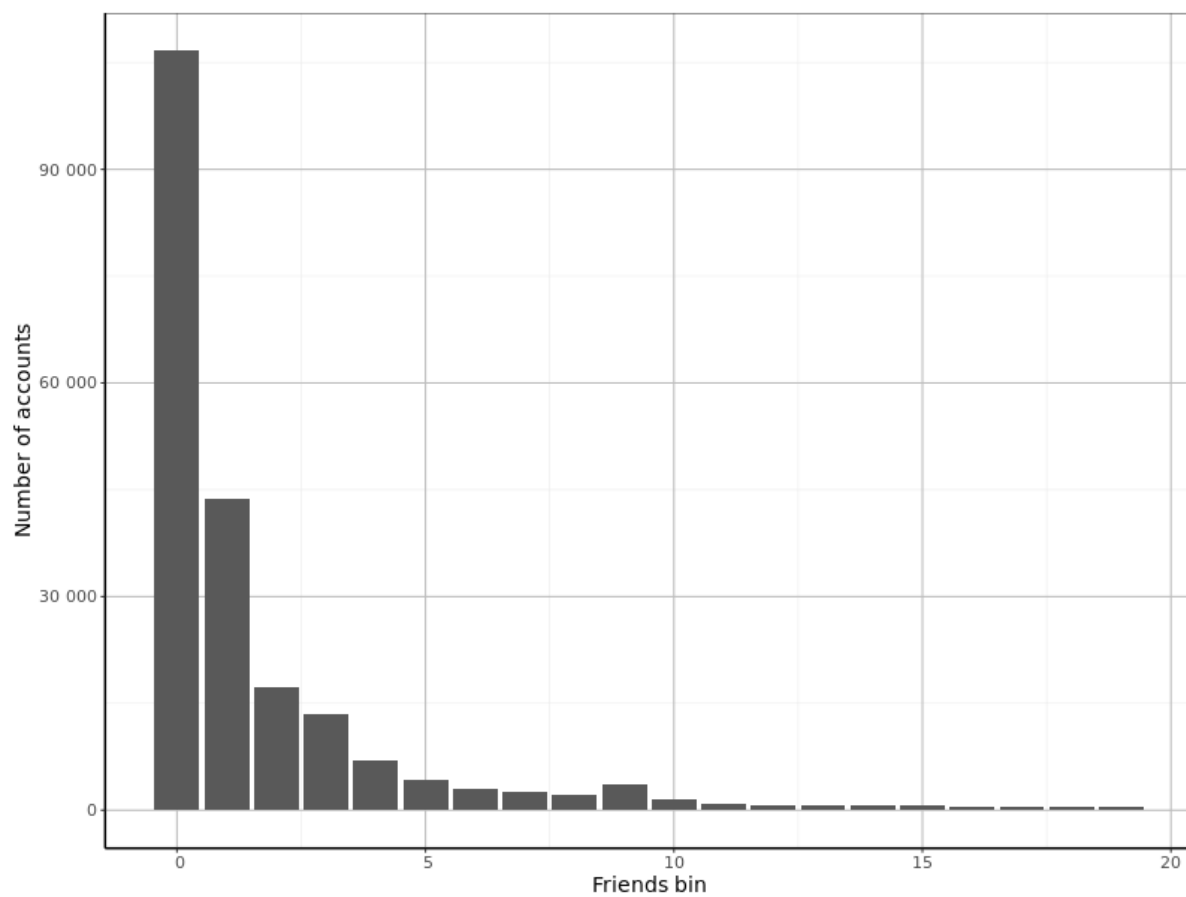


Figure 8.17: Results from binning the FRIENDS_COUNT attribute

the researcher concluded that the correlations of the attributes and features had remained intact.

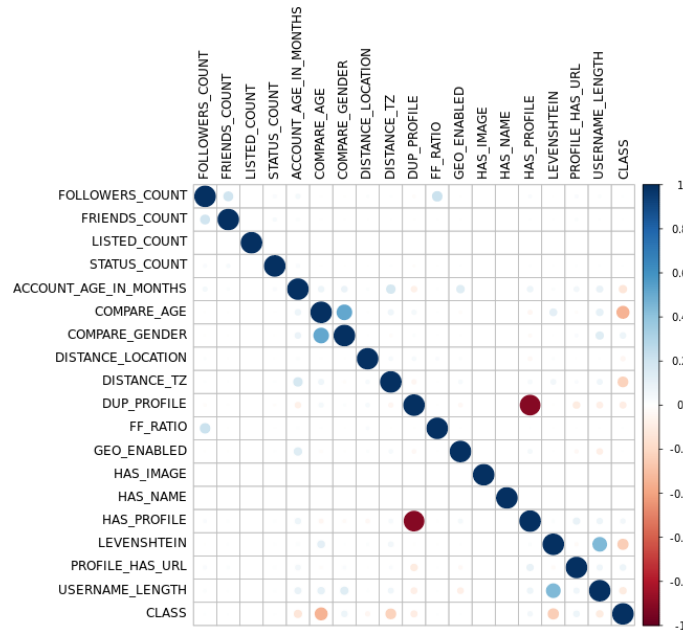


Figure 8.18: The correlation between attributes after machine learning preparation

8.7 The ‘prepare’ component as a state transition diagram

With the ‘prepare’ component, data evolves from one state to another. This process is shown by means of a state transition diagram that describes the behaviour of a system, in this case the preparation of data [51]. Figure 8.19 shows how the gathered data was firstly cleaned by removing unnecessary data that did not meet the requirements of this research. Next, generated accounts were added to change the data to a labelled state. Finally, engineered features known to be indicative of deception in other research fields were added and the data was prepared to be in a state acceptable for supervised machine learning.

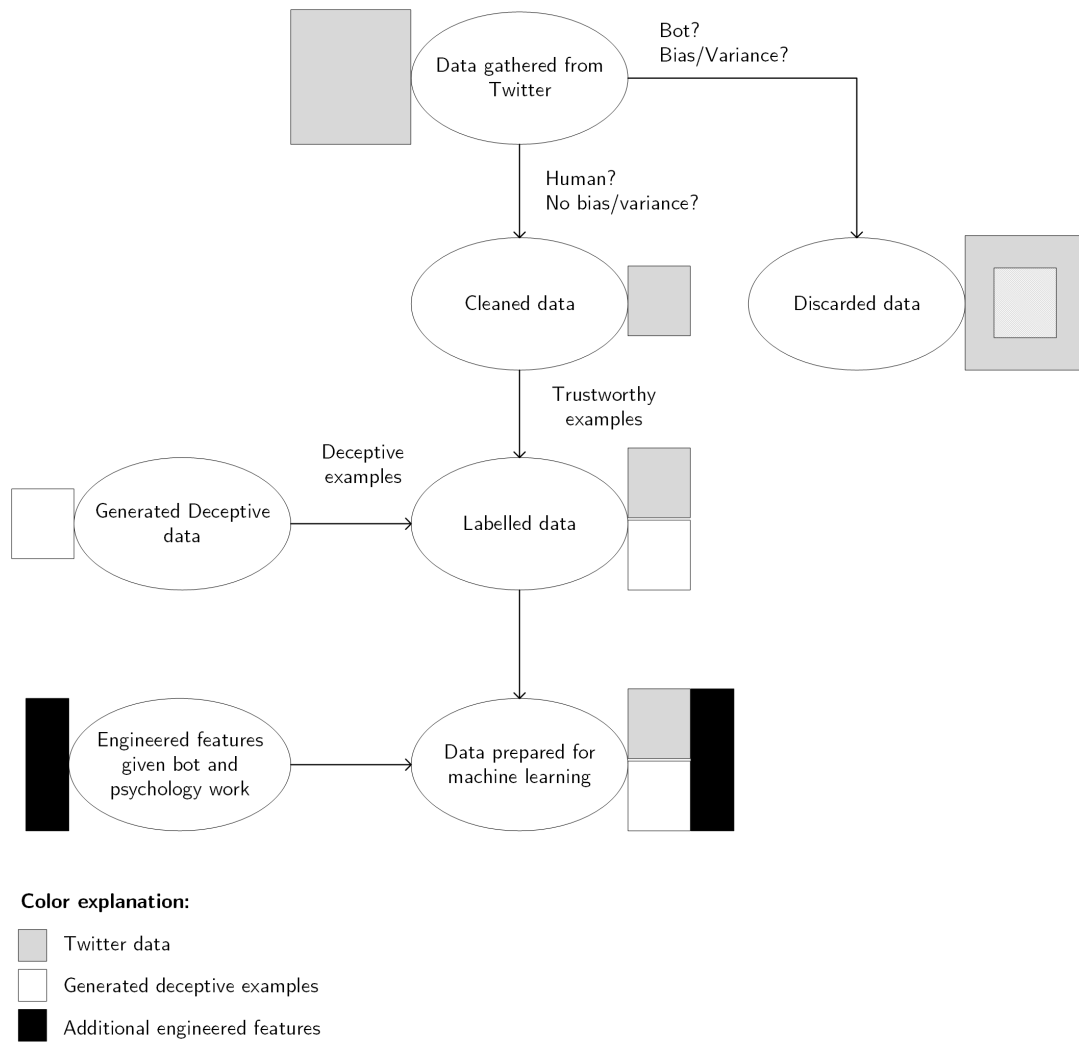


Figure 8.19: Data preparation as a state transition diagram

8.8 Conclusion

To develop a model that can assist in the automated detection of human identity deception on SMPs, data is required. This data should be gathered from an SMP, be cleared of any non-human accounts, contain examples of deceptive and trustworthy accounts, contain engineered features indicative of identity deception, and be acceptable for use by supervised machine learning algorithms. All of these requirements were discussed in this chapter. It was shown how these requirements are implemented through a prototype component, namely the ‘prepare’ component.

SMP data cannot be introduced ‘as is’ when developing a machine learning model that can detect human identity deception. This is because SMP data is full of examples of missing and incomplete values. With the ‘prepare’ component, the gathered SMP data

was transformed from one state to the next and this process was illustrated in a state transition diagram. The SMP data was enriched by adding generated user accounts in an effort to create a labelled corpus of trustworthy and deceptive accounts. Engineered features that were indicative of deception were added, and great care was taken to ensure that bias and variance were not introduced by incorporating these additions. Lastly, the data was prepared for machine learning. A correlation matrix was used during each state of the data to ensure that no additional correlations were introduced by the corpus.

Chapter 9 introduces the ‘discover’ component of the prototype that is proposed to assist in the automated detection of human identity deception on SMPs. This component will develop various machine learning models by using the prepared data in various experiments.

Chapter 9

Prototype: Discover

“You can fool some of the people all the time, and all of the people some of the time, but you cannot fool all of the people all the time.” - Abraham Lincoln

9.1 Introduction

This research proposes an identity deception detection model, implemented through a prototype, to assist in the automated detection of human identity deception on Social Media Platforms (SMPs). The proposed prototype consists of three main components – prepare, discover and detect. The previous chapter discussed the first main component that gathers and prepares the data for supervised machine learning. An initial exploration of the data showed that many expected trends could be found in the data, such as users tweeting more during the week than over weekends [188]. During exploration of the data it was however found that many attributes were incomplete, such as the background image chosen for the user’s profile. Certain attributes were also found to introduce variance or bias because of their content. For example, since the user identifier (ID) is unique, it would introduce variance and a machine learning model would be built by incorrectly associating specific IDs with deceptiveness.

With the ‘prepare’ component, the data was cleaned from all attributes that were incomplete and that introduced high variance or high bias. In addition, the SMP data was cleaned from non-human accounts, as the final model should assist in the automated detection of human identity deception on SMPs only. As the last phase of preparation, the data was labelled, additional engineered features were introduced to

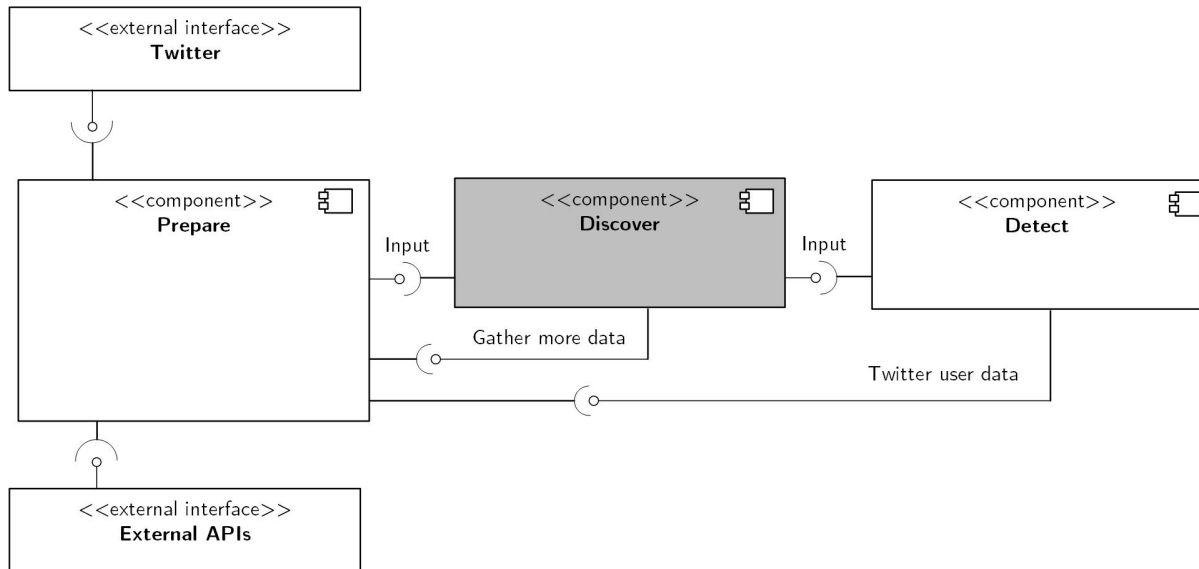


Figure 9.1: High-level overview of the prototype: ‘Discover’ component

complement the detection of human identity deception, and the data was converted for consumption by supervised machine learning algorithms.

This chapter discusses the ‘discover’ component (see Figure 9.1), which uses the prepared SMP data as input in the application of various experiments with supervised machine learning algorithms. The outcome is a model that can assist in the automated detection of human identity deception on SMPs.

To create this human identity deception detection model, the ‘discover’ component was divided into two sub-components as illustrated in Figure 9.2. Firstly, supervised machine learning experiments were performed to build models that can detect human identity deception on SMPs. Next, the results obtained from the experiments were evaluated as some models can perform better than others. The goal of the ‘discover’ component was to present the best model for detecting human identity deception on SMPs.

Chapter 9 firstly discusses in detail each of the experiments executed for this research to show how each experiment builds on the results from the previous. Secondly, the parameters used across all supervised machine learning experiments are presented to show that the same parameters were consistently applied across all experiments. Thirdly, the results of each experiment are presented, in other words the attributes and engineered features that were used in the relevant experiment, the results from the machine learning models developed, and lastly the importance, or entropy, of each attribute or feature. The chapter concludes with a critical evaluation of the results obtained from these experiments.

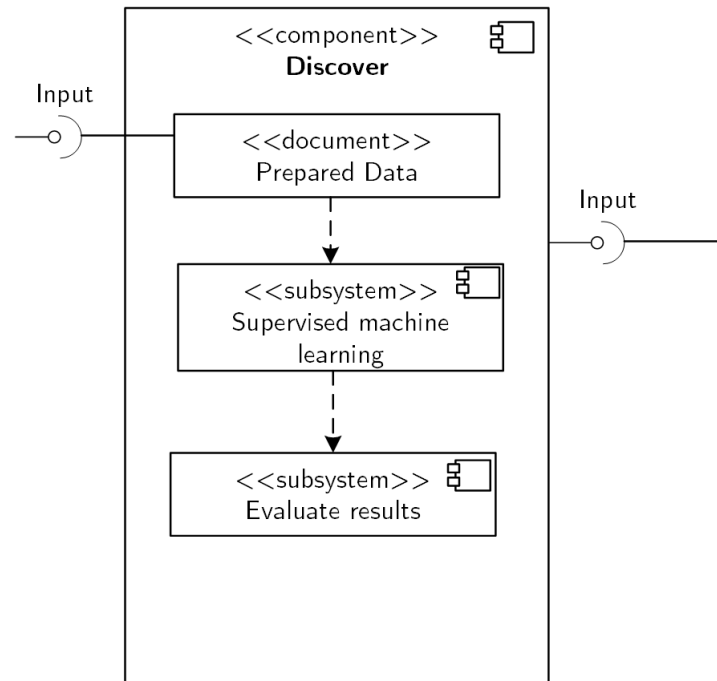


Figure 9.2: The ‘discover’ component

9.2 The experiments to detect identity deception on SMPs

To build a supervised machine learning model that can assist in the automated detection of human identity deception on SMPs, various experiments were executed. The nature of the research problem at hand, namely finding a supervised machine learning model that can detect humans who lie about who they are, warranted a range of iterative experiments dealing with the following challenges:

- It was initially not known which attributes and engineered features could make a positive contribution to the detection of deceptive humans on SMPs. By experimenting with various inputs, the importance of these attributes and features could be determined so as to improve subsequent experiments.
- Since different machine learning algorithms make use of different techniques, the problem of finding deceptive identities on SMPs could potentially be solved by using more than one machine learning algorithm. The concept of having to find the most accurate model among different machine learning algorithms is known as the “no free lunch” theorem [112]. For the current research problem, the user accounts were labelled as trustworthy or deceptive. Therefore, the set of available machine learning algorithms could be limited to those solving supervised machine

learning classification problems.

- Each machine learning algorithm had different hyperparameters that affected the algorithm. An example was the depth to which a decision tree should be constructed, given the SMP user account data as input. If the tree depth was too shallow, the model would be too general (underfit), and if the tree depth was too deep, the model would be too specific (overfit). Therefore, various hyperparameter values had to be tested to find a balance between underfitting and overfitting in an effort to find deceptive humans that had not been detected before.
- The initial SMP data provided to the supervised machine learning algorithms had to be changed, as it was considered that the more data provided to the supervised machine learning process, the better [144]. More data would however slow down the time required to develop a model. At some point, enough sample data, also known as training data, had to be available to create an accurate model, thus making additional data an unnecessary overhead to the model development process. A balance had to be found to determine the correct amount of data to achieve the most accurate results.

The concept of iterative experimentation, as used in this research, is illustrated in Figure 9.3. In an ideal iterative experimentation scenario, the input and what was learnt from previous experiments propose to improve the results of the next experiment.

The researcher proposed the following experiments to build a supervised machine learning model that could assist in the automated detection of human identity deception on SMPs:

- The first experiment used the attributes ‘as is’, as found on SMPs, to detect humans lying about their identity.
- The second experiment used additional engineered features that were known to be successful in related work to detect bots, with the intention to improve the results obtained from the previous experiment.
- The third experiment also used additional engineered features, but those that were known to identify deceptive humans in the field of psychology, to improve the results achieved from the previous experiments.

Even though more experiments could have been performed, these three were chosen

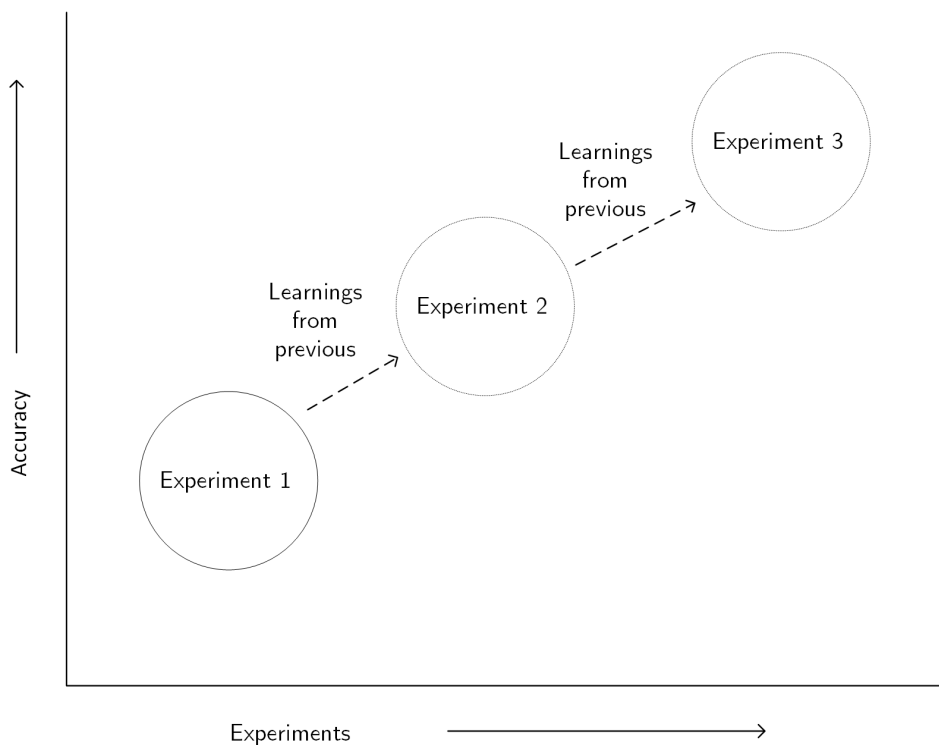


Figure 9.3: The intention with iterative experimentation

as they illustrated the benefit of related work in bots and psychology to solve for the research problem of detecting humans who lie about their identity on SMPs.

For each experiment, various machine learning models were developed by using a combination of machine learning algorithms, hyperparameters, and sample data. By keeping two of these three elements consistent across all experiments, it was possible to infer that any changes in the results of the experiments were due to the third element. The next section describes the machine learning algorithms and hyperparameters that were consistently used across all experiments to ensure comparability between the experiments. This meant that any changes in determining whether a human was deceptive could, from here on, be attributed to changes in the data alone.

9.3 Supervised machine learning

For this research, various machine learning algorithms were used to create models that were compared to determine the best model to assist in the automated detection of human identity deception on SMPs. The eight machine learning algorithms identified for this research are shown in Table 9.1. The table indicates which R Caret

Table 9.1: Machine learning algorithms used to develop a model for each experiment

Machine learning algorithm description	R Caret library name [191]	Algorithm Family
Adaptive boosting	Adaboost	Boosting
Bayesian generalised linear	bayesglm	Linear
J48 library from Weka	J48	Tree
K Nearest Means	kknn	Clustering
Neural Network	nnet	Neural Network
Random Forest	rf	Tree
Recursive partitioning tree	rpart	Tree
SVM with Radial Basis Function Kernel	svmRadial	SVM

library [191] implements the machine learning algorithm, as well as which algorithm family (as described in Chapter 5) the machine learning algorithm belongs to.

Each model was developed using 10-fold, 3-repeat cross-validation resampling. This in itself means that each model was developed 30 times using data samples drawn from the training data set. Furthermore, up to three hyperparameters were tested for each algorithm. These hyperparameters values were the default provided by the R Caret [191] package and are shown in Table 9.2. A description of each of these hyperparameters was provided in Chapter 5. Table 9.2 extends the description with a comma-separated list of hyperparameter values that had been evaluated during each experiment.

Various metrics were produced per machine learning model to evaluate its performance. These metrics included *inter alia* Accuracy, Kappa, F1 score and the cost (i.e. the time it took to develop each model). In addition to the F1 score, which was the final metric used to determine the model performance results, the PR-AUC and ROC-AUC metrics were also produced. The researcher gave preference to PR-AUC, because it is not affected by data imbalances and does not consider true negatives – in this case, finding the users who are not deceptive.

Based on the supervised machine learning algorithms and hyperparameters, the following results were produced per experiment:

- The input data was used by each supervised machine learning algorithm to develop a model that can detect humans lying about their identity on SMPs. The outcome included machine learning metrics, like the F1 score.
- The above process was repeated 30 times to produce 30 machine learning models per supervised machine learning algorithm. The 30 results were used to

Table 9.2: Machine learning algorithm hyperparameters used across all experiments

Machine learning algorithm description	R Caret library name [191]	Hyperparameter name	Hyperparameter values used in experiments
Adaptive boosting	Adaboost	nIter (#Trees)	50,100,150
		method (Method)	Adaboost.M1, Real adaboost
Bayesian generalized linear	bayesglm	-	
J48 library from Weka	J48	C (Confidence Threshold)	0.01, 0.255, 0.5
		M (Minimum Instances Per Leaf)	1,2,3
K Nearest Means	kkn	kmax (Max. #Neighbors)	5,7,9
		distance (Distance)	2
		kernel (Kernel)	Optimal
Neural Network	nnet	size (#Hidden Units)	1,3,5
		decay (Weight Decay)	0, 0.01, 0.00001
Random Forest	rf	mtry (#Randomly Selected Predictors)	2,3,5
Recursive partitioning tree	rpart	cp (Complexity Parameter)	0.0296, 0.0318, 0.0677
SVM with Radial Basis Function Kernel	svmRadial	C (Cost)	1

demonstrate the confidence in the results of a particular machine learning algorithm.

- The Mann-Whitney U test was performed, using the results obtained from the 30 machine learning models, per machine learning algorithm, to determine whether one machine learning algorithm outperformed the others. All 30 F1 score results of each machine learning algorithm's model were compared with the 30 results of another machine learning algorithm's model. For every comparison, a Mann-Whitney U test was performed at the 95% level of significance. If the first model significantly outperformed the second model, a win was recorded. If no statistical difference was observed, a draw was recorded. If the second model outperformed the first model, a loss was recorded for the first model [288].
- A supervised machine learning model was developed, per machine learning algorithm, by using various sizes of input data to explain the influence of the size of input data on the results. For example, the researcher attempted to establish whether a random forest algorithm would detect identity deception by humans on SMPs equally good when using 10 000 user accounts as when using 100 000 user accounts.
- Finally, the entropy for each attribute or engineered feature used in an experiment was calculated to determine its contribution towards detecting deceptive humans

on SMPs. The entropy was measured as a value between 0 and 100, with 100 contributing most. The entropy results helped to determine which features would be used in the next experiment, as the researcher omitted those features that did not contribute to detecting deception from later experiments.

The results produced for each of the three experiments are discussed in detail next.

9.4 Results from the supervised machine learning experiments

Three experiments were executed for this research. As mentioned earlier, each experiment had the intent to improve on the results of the previous experiment by applying knowledge from the previous experiment as well as adding features that could help further in detecting deceptive humans on SMPs. The results of this iterative experimental process are discussed next.

9.4.1 Experiment 1 – Using attributes from social media platforms ‘as is’

For the first experiment, the prepared SMP data was taken ‘as is’. This means that no extra engineered features were added. The prepared SMP data was used to develop the eight mentioned supervised machine learning algorithms. With this particular experiment, the intention was to understand whether the attributes normally found in SMP data alone could assist in the automated detection of identity deception by humans on SMPs.

9.4.1.1 Supervised machine learning input data

The prepared dataset that was used in Experiment 1 is represented as boxplots in Figure 9.4. The figure shows that the mean and distributions across all attributes per class were equal. This is expected for supervised machine learning algorithms to be able to develop a model. Furthermore, the appended deceptive user accounts were similar in terms of their identity attributes found on Twitter. Supervised machine learning algorithms would thus not favour one class over another.

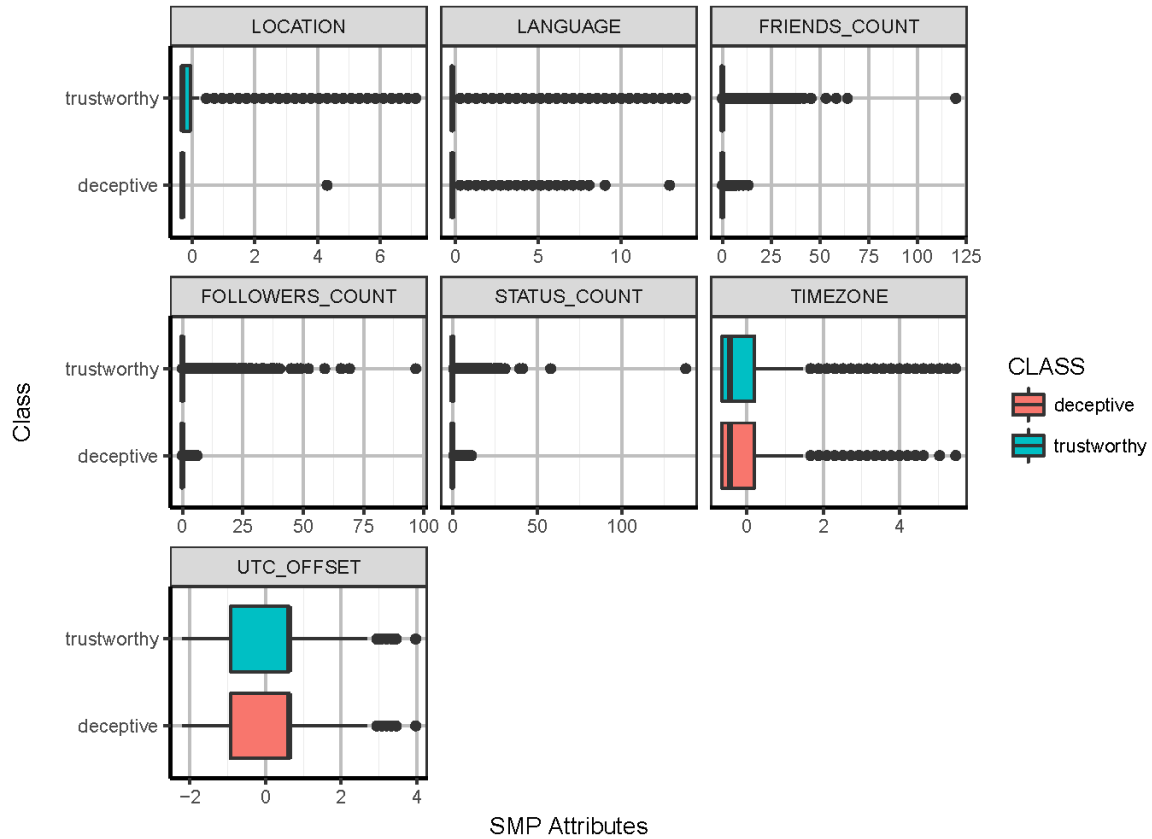


Figure 9.4: Experiment 1 - Distribution of input data

9.4.1.2 Supervised machine learning model results

The results obtained from Experiment 1 are presented in Table 9.3. The best result was achieved by the rf model with an F1 score of 32.83%. The combined Area Under Curve (AUC) results nevertheless confirm that this result was not optimal (see Figure 9.5). With ROC-AUC, all models are shown to have performed well in their predictions of identity deception, whereas with the PR-AUC only half of the models were accurately detecting about 25% of the deceptive humans.

To ensure consistency of the results, the same experiment was repeated to produce 30 results. The F1 scores for all 30 results, per supervised machine learning model, are presented in Table 9.4. A full list of results for all metrics can be found in Appendix F. From these results the following was determined (as indicated at the end of the table):

- The average F1 score varied less than 3% from the initial presented machine learning results presented in Table 9.4. This implies that the results can be consistently reproduced.

Table 9.3: Experiment 1 - Machine learning results

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	15.33	-0.12	8.79	91.41	16.04	53.67	11.11	71.659
rf	79.31	23.13	23.03	57.15	32.83	76.11	28.57	98.164
J48	77.92	19.24	20.56	52.21	29.50	71.69	24.82	115.409
bayesglm	65.38	2.26	10.09	36.80	15.83	52.73	9.54	4.542
kknn	71.32	12.68	15.99	52.69	24.53	68.36	18.55	61.368
Adaboost	78.86	19.96	21.20	51.12	29.97	71.13	32.02	883.437
rpart	66.25	11.02	14.77	58.96	23.62	63.05	13.55	4.115
nnet	66.14	13.96	16.26	68.11	26.25	74.31	33.19	39.153

- The Mann-Whitney U test results showed that the rf model significantly outperformed the other models and scored a win over all other models.

Lastly, Experiment 1 was executed with different-sized datasets, which made it possible to determine whether enough data had been used to develop the models. Table 9.5 shows the results of running this experiment on a dataset consisting of 16 000, 25 000, 115 000 user accounts, which includes the 15 000 appended deceptive accounts. These results were compared to the prepared corpus of 169 417 user accounts and showed that smaller datasets produced better accuracy. Upon further investigation, this was found to be misleading, due to the introduction of bias. Less data makes it harder for a machine learning algorithm to generalise and it therefore uses only what it knows. As shown in Figure 9.6, the results converged at the final dataset, similar to what one would see in the elbow method used for clustering [37]. The researcher concluded that the size of the dataset used to develop the machine learning model was sufficient.

9.4.1.3 Attribute or engineered feature entropy

Not only did the machine learning models present results to understand whether they could predict identity deception based on the data, but they also provided results, in the form of entropy, regarding those attributes or features that contributed most. This was important for future experiments. Attributes with high entropy could be the focus of subsequent experiments, which would use these attributes to engineer new features in the hope of improving the accuracy of the predictions. Table 9.6 presents the entropy results for each of the machine learning algorithms. Values are presented in a range from 0 to 100. A value of 0 indicates that the attribute contributed nothing, whereas a value of 100 indicates that the attribute contributed very much to the final outcome of developing the machine learning models. The results show that FRIENDS_COUNT, FOLLOWERS_COUNT, and STATUS_COUNT were important in developing the machine learning models.

Table 9.4: Experiment 1 - F1 scores over 30 repeats

	Machine learning algorithm (%)							
	svmLinear	rf	J48	bayesglm	kknn	Adaboost	rpart	nnet
1	16.04	32.83	29.50	15.83	24.53	29.97	23.62	26.25
2	16.26	32.58	30.07	13.84	23.93	29.03	23.73	26.80
3	9.30	32.98	29.54	14.01	24.30	28.65	23.38	26.24
4	6.51	32.91	29.92	14.60	24.73	30.39	23.02	25.79
5	14.53	32.33	29.39	14.90	23.49	29.82	18.29	26.45
6	14.18	31.93	28.82	15.11	24.01	28.90	20.48	26.50
7	15.55	32.07	29.25	14.99	24.60	29.52	22.85	25.99
8	14.64	32.13	28.45	16.08	23.20	29.91	22.10	27.63
9	14.49	32.91	29.53	14.80	23.65	29.26	23.49	27.28
10	15.76	33.09	0.43	17.18	23.79	28.74	23.14	28.22
11	16.17	31.97	28.78	13.80	24.32	30.05	23.26	27.12
12	14.29	31.79	27.79	14.46	23.72	29.53	23.20	27.13
13	15.09	31.77	27.16	15.33	22.62	29.24	23.05	27.90
14	13.95	31.74	27.28	15.26	23.18	29.40	23.49	27.59
15	15.00	32.21	29.43	14.69	24.00	29.84	22.63	26.07
16	14.98	33.20	28.62	15.69	23.98	29.68	25.15	28.05
17	13.21	32.12	27.64	13.28	23.41	28.48	18.47	26.90
18	13.21	32.12	27.64	13.28	23.41	28.48	18.47	26.90
19	14.47	31.86	30.14	14.19	23.97	31.39	24.38	27.81
20	15.47	31.76	29.33	15.21	24.73	29.19	24.02	24.61
21	5.37	32.57	30.25	15.20	24.29	30.23	25.71	26.19
22	14.49	32.79	27.97	15.70	23.62	29.95	23.61	27.22
23	8.14	32.18	29.88	15.34	24.09	29.05	25.15	26.71
24	16.06	33.50	30.28	16.91	25.33	31.09	24.01	27.21
25	16.27	32.35	27.17	12.32	24.59	29.27	22.56	26.25
26	14.01	32.74	29.10	15.22	22.96	30.12	23.88	23.27
27	4.58	32.72	29.67	14.42	23.89	30.12	23.20	27.12
28	15.97	32.42	27.47	13.60	24.17	28.67	25.16	25.70
29	7.07	32.63	28.54	14.31	24.16	30.77	22.26	28.29
30	16.12	32.08	29.32	17.48	23.62	29.06	22.08	25.77
Average F1 score (%)	13.37	32.41	27.95	14.90	23.94	29.59	22.93	26.70
Variance from original F1 score (%)	2.67	0.42	1.55	0.93	0.59	0.38	0.69	-0.45
Mann Whitney U (Win-Draw-Loss) *	0-1-6	7-0-0	5-0-2	0-1-6	3-0-4	6-0-1	2-0-5	4-0-3

* For example, (4-0-3) indicates that the nnet model significantly outperformed four other models. Furthermore, no draws and three losses were recorded.

Table 9.5: Experiment 1 - F1 scores for various-sized dataset results

	F1 scores per dataset size (%)			
	16 000	25 000	115 000	169 417
svmRadial	80.33	14.26	3.04	16.30
rf	83.55	73.74	40.72	33.78
J48	96.77	71.30	38.41	31.52
bayesglm	40.28	57.42	22.74	16.55
kknn	74.41	65.64	32.58	24.12
Adaboost	82.32	70.25	35.73	30.48
rpart	82.88	59.76	30.06	24.90
nnet	67.18	69.62	36.25	27.75

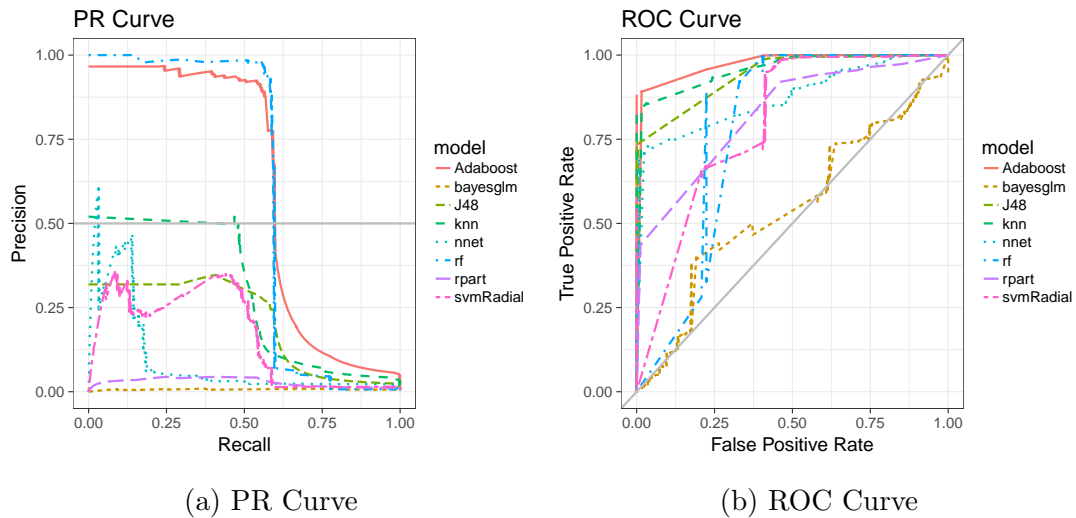


Figure 9.5: Experiment 1 - Combined AUC results

9.4.1.4 Summary of Experiment 1

In summary, the following was learnt from Experiment 1:

- The random forest machine learning algorithm produced the most accurate result and gave an F1 score of 32.83
- The Mann Whitney U test results confirmed that the random forest machine learning algorithm was superior to the other machine learning algorithms.
- The result, given the F1 score and PR-AUC, was however still not optimal, as only half of the machine learning models managed to predict about 25% of the deceptive accounts successfully. Although an optimal prediction value is subjective, the researcher was expecting a success rate of at least 50% or more. This would mean finding more deceptive accounts than getting it wrong.
- Results from 30 repeats of the experiment showed that the values were consistent with F1 scores, varying less than 3
- Entropy results showed that FRIENDS_COUNT, FOLLOWERS_COUNT, and STATUS_COUNT were the features that were most indicative of deception.

Given these results, it was clear that the first experiment failed at detecting human identity deception on SMPs. The results were consistent across 30 repeats. Interesting though was the entropy results, which revealed three attributes that also emerged from related work on the detection of bot accounts. This knowledge would be used in the next experiment.

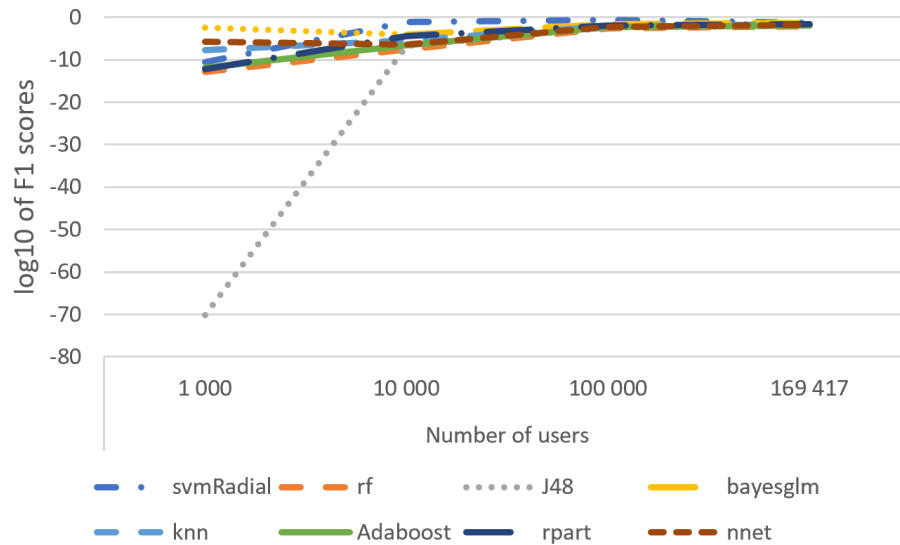


Figure 9.6: Experiment 1 - Various-sized dataset results

What Experiment 1 revealed, was that the metrics produced by machine learning models can be deceiving. Accuracy, for example, looked very good for the first experimental results, but due to the nature of data, this measure was misleading and F1 score was a more indicative metric. This same fact was also visible in the difference between the ROC-AUC and PR-AUC results, due to the skewness in data. Accuracy does not account for getting the prediction wrong. If you have nine trustworthy accounts and one deceptive account, for example, the accuracy will be 90%, even if the algorithm predicts an account to always be trustworthy.

9.4.2 Experiment 2 – Using bot detection rules

Experiment 2 enriched the dataset by adding features previously defined as being successful at detecting bots or non-human accounts on SMPs. These features that were used to identify non-human accounts, closely resembled the results found in the entropy of Experiment 1. The new features were used to develop the same eight supervised machine learning models with the intention to improve on the accuracy of the previous experiment.

9.4.2.1 Supervised machine learning input data

Related research that proposed to distinguish between deceptive bots or non-human accounts on SMPs identified engineered features to aid in such detection. Given the

Table 9.6: Experiment 1 - Entropy results

	Machine learning algorithm							
	svmRadial	rf	J48	bayesglm	kknn	Adaboost	rpart	nnet
FOLLOWERS_COUNT	61.932	100.000	61.932	61.932	61.932	61.932	54.271	99.703
FRIENDS_COUNT	64.116	81.959	64.116	64.116	64.116	64.116	88.617	50.751
LISTED_COUNT	5.627	0.000	5.627	5.627	5.627	5.627	0.000	28.432
STATUS_COUNT	40.725	78.489	40.725	40.725	40.725	40.725	62.809	11.468
TIMEZONE	54.148	40.175	54.148	54.148	54.148	54.148	29.133	2.825

entropy results from Experiment 1, the researcher proposed to introduce these additional engineered features as they could improve a model that detects identity deception by humans on SMPs. The final additional engineered features and attributes used in Experiment 2 are shown in Figure 9.7 as distribution boxplots. The figure shows that the mean and distributions across all attributes per class were equal (as was expected for supervised machine learning algorithms in order to be able to develop a model) and indicates that the appended deceptive user accounts were similar in terms of their identity attributes found on Twitter. Supervised machine learning algorithms would thus not favour one class over another.

9.4.2.2 Supervised machine learning model results

The dataset that contained engineered features used in earlier related work to detect bot accounts, was used as input to develop the machine learning models. The results from this experiment are presented in Table 9.7. The best result was achieved by the rf model, and an F1 score of 49.75% was obtained. The combined AUC results confirmed that this result was not optimal, as shown in Figure 9.8. With ROC-AUC, all models were shown to have performed well in their predictions of identity deception. The PR-AUC however indicated that only half of the models managed to accurately detect about 40-50% of the deceptive humans.

To ensure consistency of the results, the same experiment was repeated to produce 30 results. The F1 scores for all 30 results, per supervised machine learning model, are presented in Table 9.8. A full list of results for all metrics can be found in Appendix F. From these results, the following was determined (as indicated at the end of the table):

- The average F1 score varied less than 2% from the initial presented machine learning results presented in Table 9.8. This implied that the results would be consistently reproduced.

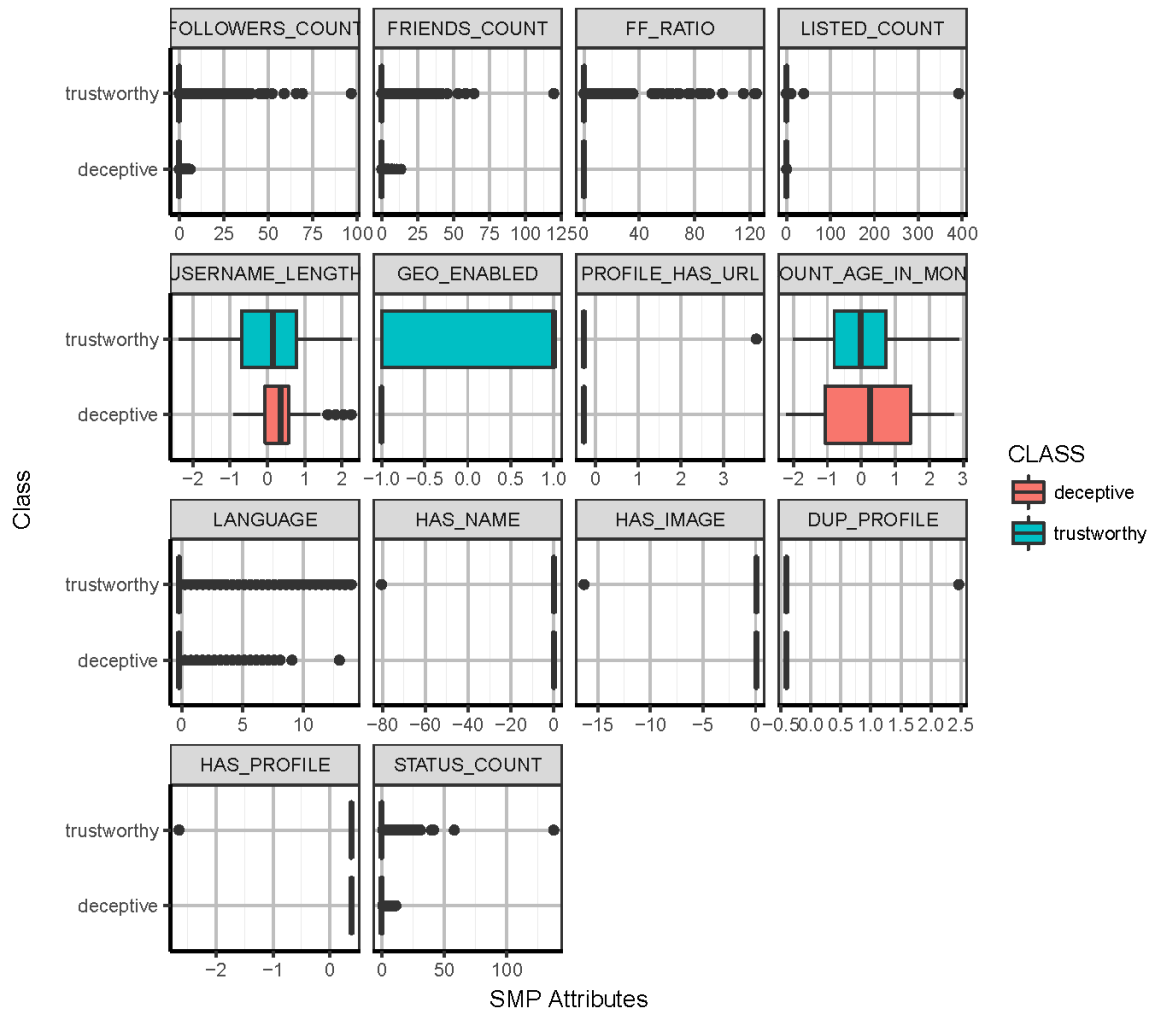


Figure 9.7: Experiment 2 - Distribution of input data

- The Mann-Whitney U test results showed that the rf model significantly outperformed the other models, with a win over all other models.

Lastly, the same experiment was performed with different-sized datasets. to defer whether enough data was used to develop the machine learning models. Table 9.9 shows the results from executing this experiment over a dataset consisting of 16 000, 25 000, 115 000 user accounts, which included the 15 000 appended deceptive accounts. These results were compared to the prepared corpus of 169 417 user accounts and showed that smaller datasets produced better accuracy. Upon further investigation, this finding was found to be misleading and due to the introduction of bias. The results converged, as shown in Figure 9.9, at the final dataset. Therefore, the researcher concluded that the size of the dataset used to develop the machine learning models was sufficient.

Table 9.7: Experiment 2 - Machine learning results

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	68.05	20.78	19.80	85.60	32.16	80.47	27.77	218.256
rf	87.11	43.16	37.98	72.11	49.75	90.24	49.90	131.81
J48	84.48	36.89	32.56	70.40	44.53	84.73	33.56	143.202
bayesglm	74.14	22.87	21.64	73.31	33.41	81.62	27.76	5.268
kknn	82.98	32.14	29.29	65.28	40.43	85.43	36.81	64.359
Adaboost	85.92	40.48	35.46	72.11	47.54	89.53	49.53	1278.777
rpart	68.82	21.09	20.03	84.35	32.37	77.21	21.70	4.066
nnet	82.48	32.70	29.23	68.99	41.07	87.03	39.87	54.614

9.4.2.3 Attribute or engineered feature entropy

The entropy of each engineered feature used in Experiment 2 is shown in Table 9.10. The following two engineered features showed the most entropy: DUP_PROFILE, HAS_NAME, and USERNAME_LENGTH. This was very similar to what is known from past research work in psychology and the fact that people tend to lie about their name and image.

9.4.2.4 Summary of Experiment 2

In summary, the following was learnt from Experiment 2:

- The random forest machine learning algorithm produced the most accurate result and gave an F1 score of 49.75
- The Mann Whitney U test results confirmed that the random forest machine learning algorithm was superior to the other machine learning algorithms.
- The result, given the F1 score and PR-AUC, is however still not optimal, as only half of the machine learning models could predict 40-50% of the deceptive accounts successfully. Although an optimal prediction value was subjective, the researcher expected a success rate of at least 50% or more. This would mean finding at least half of the deceptive accounts.
- Results from 30 repeats of Experiment 2 showed that the values were consistent with F1 scores and varied less than 2%.
- Entropy results showed that DUP_PROFILE, HAS_NAME, and USERNAME_LENGTH were the features that were most indicative of deception.

Based on these results, it was clear that the second experiment also failed to detect human identity deception on SMPs. The results were however better than those of the

Table 9.8: Experiment 2 - F1 scores over 30 repeats

	Machine learning algorithm (%)							
	svmLinear	rf	J48	bayesglm	kknn	Adaboost	rpart	nnet
1	32.16	49.75	44.53	33.41	40.43	47.54	32.37	41.07
2	32.26	48.23	44.03	33.15	40.81	47.96	35.83	42.00
3	32.21	50.60	43.92	32.77	40.13	47.21	34.97	42.70
4	32.10	48.74	44.77	33.18	39.83	46.90	35.52	40.77
5	32.37	48.33	44.02	32.88	39.49	47.74	31.38	39.80
6	32.10	48.90	43.55	33.25	40.04	46.81	35.68	40.95
7	31.75	48.17	45.00	32.47	40.25	47.51	30.31	40.02
8	32.33	47.51	44.21	32.70	39.93	46.75	35.56	39.36
9	32.75	48.01	45.60	33.35	40.57	47.35	35.52	42.52
10	32.60	50.58	45.66	33.13	40.92	48.45	38.46	41.97
11	32.32	48.85	44.24	33.15	41.02	48.03	32.48	42.34
12	32.63	48.76	45.86	33.19	40.88	47.32	35.67	40.31
13	32.12	48.81	45.69	32.83	40.23	47.00	30.68	41.89
14	31.84	46.83	44.60	32.56	39.67	47.07	35.05	41.19
15	32.52	48.79	42.11	33.34	40.00	46.88	35.94	41.79
16	32.54	49.73	45.38	32.92	40.42	47.04	35.44	39.64
17	32.36	49.24	45.23	32.92	40.58	47.45	35.14	41.73
18	32.34	51.01	45.67	33.02	41.05	47.63	30.92	43.46
19	32.11	48.43	44.35	33.10	40.08	47.06	30.26	42.52
20	32.54	48.43	44.49	33.13	40.24	47.58	37.46	41.05
21	32.31	49.23	43.53	32.96	41.57	47.47	35.42	40.52
22	32.70	47.85	45.42	33.38	41.43	47.96	31.25	43.69
23	32.28	49.45	45.49	33.39	40.14	47.60	35.52	43.04
24	32.23	48.83	44.39	33.30	40.99	47.08	35.42	41.59
25	32.16	49.62	43.36	33.28	41.57	47.21	31.01	39.83
26	32.25	49.65	45.19	33.31	41.22	47.23	35.56	44.28
0 27	32.38	49.24	44.33	33.33	40.32	46.97	38.40	41.34
28	32.24	50.87	43.15	33.25	40.60	47.98	35.42	39.58
29	32.53	49.29	44.29	33.41	40.98	47.12	31.06	42.36
30	32.37	50.32	45.81	33.65	40.67	47.10	30.50	42.60
Average F1 score (%)	32.31	49.07	44.59	33.12	40.54	47.37	34.14	41.53
Variance from original F1 score (%)	-0.15	0.68	-0.07	0.29	-0.10	0.18	-1.77	-0.46
Mann Whitney U (Win-Draw-Loss) *	0-0-7	7-0-0	5-0-2	1-1-5	3-0-4	6-0-1	1-1-5	4-0-3

* For example, (4-0-3) indicates that the nnet model significantly outperformed four other models. Furthermore, no draws and three losses were recorded.

Table 9.9: Experiment 2 - F1 scores for various-sized dataset results

	F1 scores per dataset size (%)			
	16 000	25 000	115 000	169 417
svmRadial	90.70	82.79	42.17	32.16
rf	94.42	86.15	57.65	49.75
J48	90.89	85.18	53.12	44.53
bayesglm	90.32	81.15	42.58	33.41
kknn	88.87	79.54	48.50	40.43
Adaboost	90.30	85.10	56.70	47.54
rpart	92.48	80.77	40.68	32.37
nnet	91.31	83.18	51.07	41.07

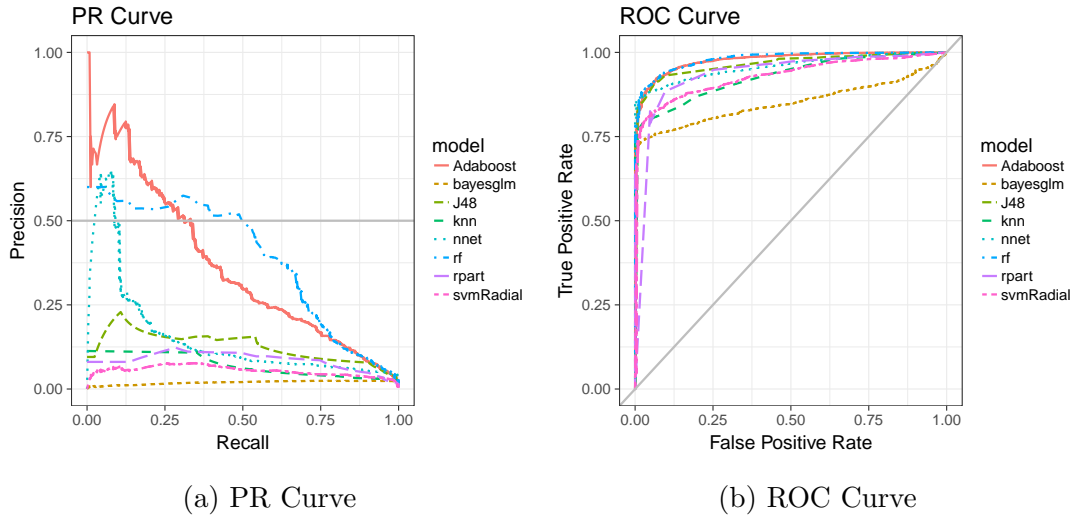


Figure 9.8: Experiment 2 - Combined AUC results

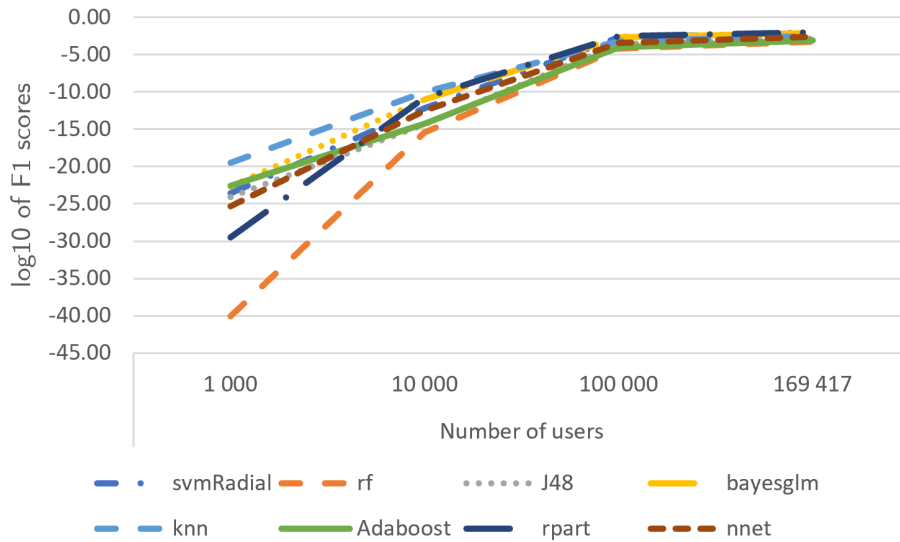


Figure 9.9: Experiment 2 - Various-sized dataset results

Table 9.10: Experiment 2 - Entropy results

	Machine learning algorithm							
	Adaboost	bayesglm	J48	knn	nnet	rpart	rf	svmRadial
ACCOUNT_AGE.IN.MONTHS	58.098	37.336	58.098	58.098	58.098	58.098	25.171	14.910
DUP_PROFILE	95.140	86.450	95.140	95.140	95.140	95.140	98.584	89.428
FF_RATIO	0.417	1.189	0.417	0.417	0.417	0.417	0.109	28.141
FOLLOWERS_COUNT	7.139	24.940	7.139	7.139	7.139	7.139	1.478	69.149
FRIENDS_COUNT	35.618	23.063	35.618	35.618	35.618	35.618	3.308	44.856
GEO_ENABLED	28.351	22.630	28.351	28.351	28.351	28.351	27.239	9.547
HAS_IMAGE	0.140	0.000	0.140	0.140	0.140	0.140	0.000	14.203
HAS_NAME	99.070	99.599	99.070	99.070	99.070	99.070	97.297	40.376
HAS_PROFILE	11.490	11.704	11.490	11.490	11.490	11.490	2.504	68.299
LISTED_COUNT	5.211	3.604	5.211	5.211	5.211	5.211	0.587	29.391
PROFILE.HAS.URL	14.186	9.923	14.186	14.186	14.186	14.186	7.724	51.166
STATUS_COUNT	18.259	22.385	18.259	18.259	18.259	18.259	0.876	10.328
USERNAME_LENGTH	39.276	60.644	39.276	39.276	39.276	39.276	59.205	42.320

first experiment, which implied that using knowledge from a related field solving for a similar problem could aid in the detection of humans lying about their identity on SMPs. The results were consistent across 30 repeats and therefore strengthened any results achieved. Entropy results indicated three new features that were more indicative of deception. These features related to a user's name and image. Originating from related work in psychology, these features matched those identified in Chapter 4 as being indicative of human identity deception.

Some lessons learnt from Experiment 2 showed that using the knowledge obtained from related research could improve the detection of human identity deception on SMPs. By only adding new features from related work in the detection of bots, the results were improved by 100

9.4.3 Experiment 3 – Using knowledge from psychology

The features created and used to develop the machine learning models in Experiment 2 improved the original model to assist in the detection of human identity deception on SMPs. The researcher performed a further experiment by adding features found in related work in the field of psychology to detect deceptive humans.

9.4.3.1 Supervised machine learning input data

In Chapter 4, additional features were identified towards identity deception based on related research work in the field of social sciences – more specifically psychology. The additional engineered features and attributes used in Experiment 3 are shown in Figure 9.10 as distribution boxplots. The figure shows that the mean and distributions across all attributes per class were equal, as was expected for supervised machine learning algorithms used for developing a model. Figure 9.10 furthermore indicates that the appended deceptive user accounts were similar in terms of their identity attributes found on Twitter – hence supervised machine learning algorithms would not favour one class over another.

9.4.3.2 Supervised machine learning model results

The dataset containing the engineered features based on psychology was used as input to develop the machine learning models to detect identity deception. The results from

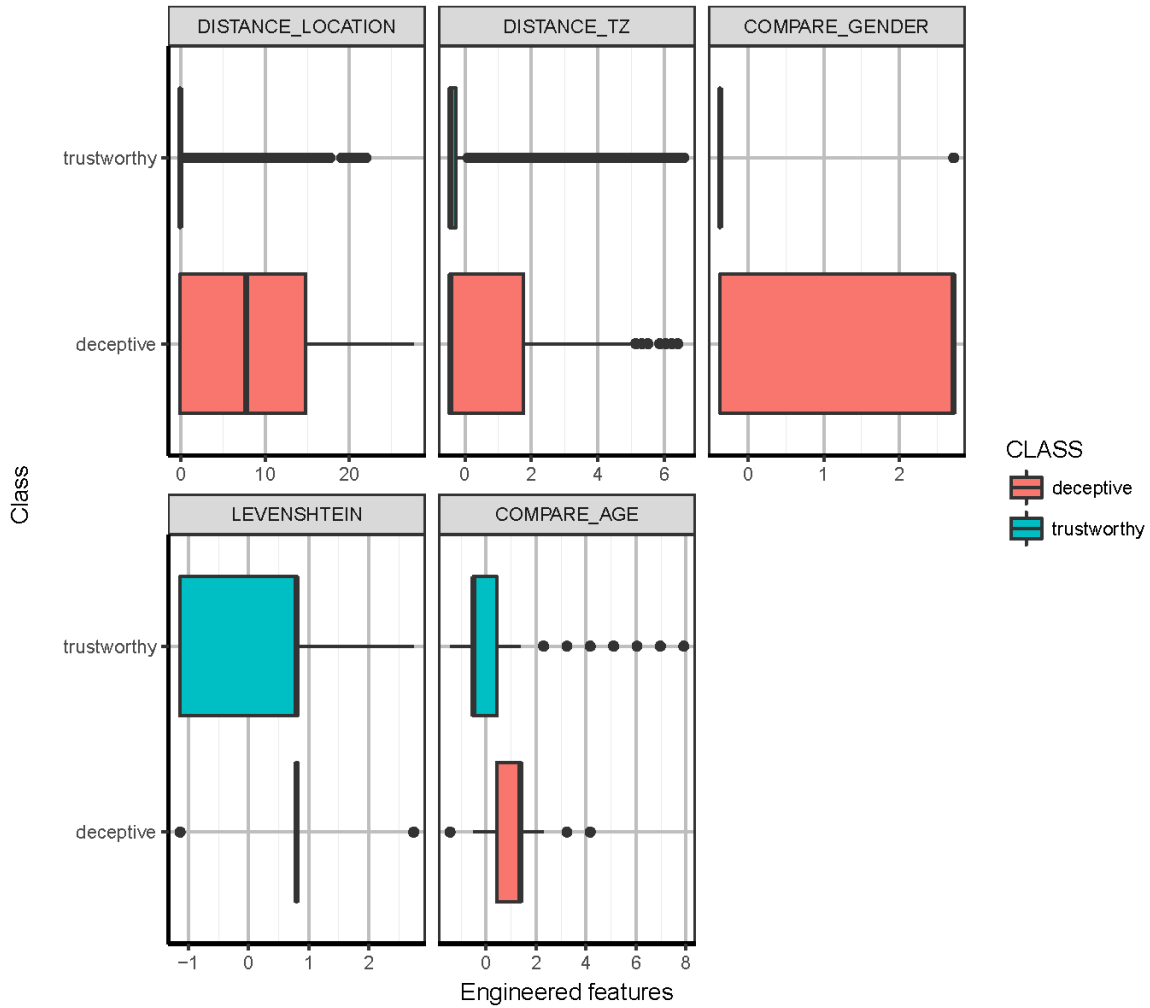


Figure 9.10: Experiment 3 - Distribution of input data

Experiment 3 are presented in Table 9.11. The best result was achieved by the rf model with an F1 score of 86.24%. The combined AUC results confirmed that this result was better than the previous experiments, as shown in Figure 9.11. With ROC-AUC, all models were shown to have performed well in their predictions of identity deception. Equally, the PR-AUC indicated that most models accurately detected 70-85% of the deceptive humans.

To ensure consistency of the results, the same experiment was repeated to produce 30 results. The F1 scores for all 30 results, per supervised machine learning model, are presented in Table 9.12. A full list of results for all metrics can be found in Appendix F. From these results, the following was determined (as indicated at the end of the table):

- The average F1 score varied less than 2% from the initial machine learning

Table 9.11: Experiment 3 - Machine learning results

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	92.20	62.15	53.67	86.67	66.29	96.01	76.80	45.198
rf	97.49	84.86	83.89	88.72	86.24	98.91	93.00	157.801
J48	95.79	76.75	70.66	89.71	79.05	95.58	64.94	178.649
bayesglm	92.07	61.65	53.20	86.40	65.85	96.00	77.01	6.932
kknn	94.18	69.79	61.96	88.64	72.93	97.01	81.32	80.933
Adaboost	97.03	83.01	77.94	92.61	84.65	99.01	93.70	2127.87
rpart	87.32	49.67	40.36	90.53	55.83	89.17	38.29	5.091
nnet	95.21	74.33	66.98	90.37	76.94	98.18	87.76	62.796

results presented in Table 9.12. This means that the results could be consistently reproduced.

- The Mann-Whitney U test results showed that the rf model significantly outperformed the other models and scored a win over all other models.

Lastly, the same experiment was performed with different-sized datasets to defer whether enough data was used to develop the machine learning models. Table 9.13 shows the results of running this experiment over a dataset consisting of 16 000, 25 000, 115 000 user accounts – which included the 15 000 appended deceptive accounts. The results were compared to the prepared corpus of 169 417 user accounts and showed that smaller datasets produced better accuracy. Upon further investigation, this was found to be misleading and due to the introduction of bias. The results converged, as shown in Figure 9.12, at the final dataset, which led the researcher to conclude that the size of the dataset used to develop the machine learning models was sufficient.

9.4.3.3 Attribute or engineered feature entropy

The entropy of each engineered feature used in Experiment 3 is shown in Table 9.14. Based on these results it seems that age and name are good indicators of deception.

9.4.3.4 Summary of Experiment 3

In summary, the following was learnt from Experiment 3:

- The random forest machine learning algorithm produced the most accurate result – an F1 score of 86.24
- The Mann Whitney U test results confirmed that the random forest machine learning algorithm was superior to the other machine learning algorithms.

Table 9.12: Experiment 3 - F1 scores over 30 repeats

	Machine learning algorithm (%)							
	svmLinear	rf	J48	bayesglm	kknn	Adaboost	rpart	nnet
1	66.29	86.24	79.05	65.85	72.93	84.65	55.83	76.94
2	64.90	84.62	80.03	64.49	72.46	83.03	54.86	72.88
3	65.27	84.89	79.51	64.97	73.12	83.01	55.63	70.56
4	64.39	85.49	79.71	64.04	72.41	83.45	55.85	77.44
5	65.42	85.43	76.66	64.93	72.22	83.21	55.77	75.54
6	65.54	85.53	79.41	65.50	72.06	83.23	56.61	77.10
7	64.48	84.50	76.36	64.17	73.43	82.86	55.31	72.84
8	64.69	84.81	76.78	64.51	72.86	83.85	55.85	77.04
9	66.12	84.82	77.07	66.07	72.68	83.40	55.91	74.75
10	65.80	85.53	79.03	65.53	73.59	84.18	55.41	76.71
11	64.78	85.03	77.91	64.72	73.01	83.64	55.59	74.81
12	65.00	84.87	78.36	64.95	74.18	83.70	55.37	77.37
13	65.02	85.14	76.24	64.90	73.97	83.83	55.27	75.51
14	65.48	85.08	77.80	65.40	73.75	84.29	55.91	77.39
15	64.87	85.30	74.05	64.92	72.56	83.24	55.94	73.49
16	65.42	84.26	78.30	64.78	73.04	83.05	55.22	72.40
17	64.60	85.48	78.06	64.55	73.26	84.46	56.08	75.44
18	64.37	85.19	76.00	64.22	72.39	82.95	55.70	76.80
19	64.41	85.13	77.69	64.51	72.79	83.14	55.42	76.72
20	64.64	84.80	75.21	64.21	72.69	82.90	55.80	74.24
21	63.80	84.20	76.65	63.84	71.66	83.61	55.74	76.77
22	64.66	85.55	81.21	64.52	73.73	84.61	55.44	78.66
23	65.21	85.70	77.90	64.93	74.13	83.50	55.58	74.26
24	65.00	85.82	80.90	64.66	73.13	83.17	55.65	74.92
25	66.00	84.86	79.59	65.56	72.85	84.08	53.28	76.28
26	65.77	83.95	77.63	65.39	71.68	82.65	55.29	75.75
27	65.55	85.67	79.52	65.02	74.29	84.58	55.37	78.59
28	63.89	85.56	80.39	63.67	72.03	83.89	55.44	76.48
29	65.24	85.82	79.06	65.44	73.53	83.81	55.95	79.06
30	64.83	85.48	80.20	64.99	72.93	83.48	55.20	75.73
Average F1 score (%)	13.37	32.41	27.95	14.90	23.94	29.59	22.93	26.70
Variance from original F1 score (%)	65.05	85.16	78.21	64.84	72.98	83.58	55.54	75.75
Mann Whitney U (Win-Draw-Loss) *	1-1-5	7-0-0	5-0-2	1-1-5	3-0-4	6-0-1	0-0-7	4-0-3

* For example, (4-0-3) indicates that the nnet model significantly outperformed four other models. Furthermore, no draws and three losses were recorded.

Table 9.13: Experiment 3 - F1 scores for various-sized dataset results

	F1 scores per dataset size (%)			
	16 000	25 000	115 000	169 417
svmRadial	95.14	91.40	73.11	66.29
rf	96.82	95.91	87.97	86.24
J48	93.62	94.88	83.25	79.05
bayesglm	94.68	91.48	72.81	65.85
kknn	92.73	92.56	79.63	72.93
Adaboost	96.36	96.53	87.22	84.65
rpart	94.42	91.20	64.99	55.83
nnet	94.07	94.30	82.75	76.94

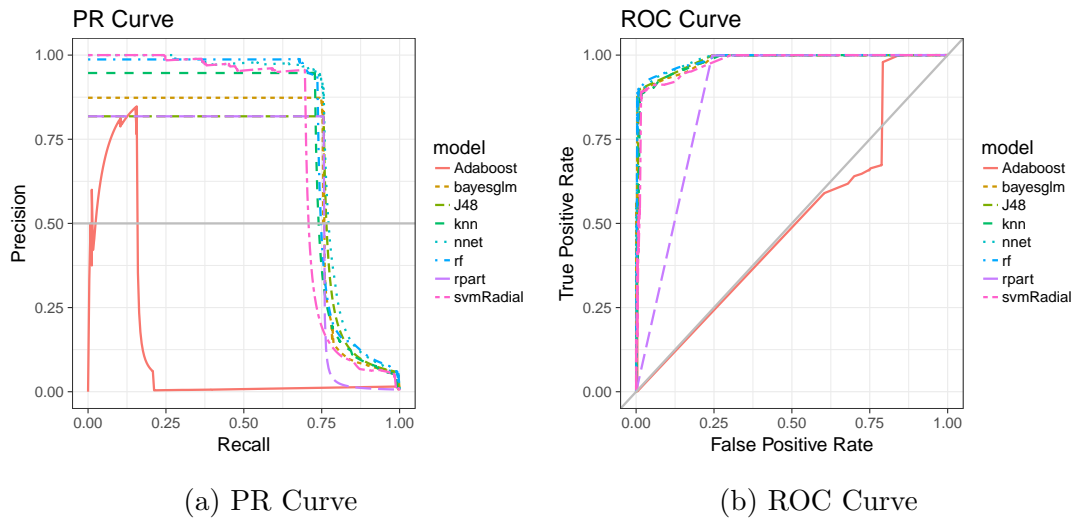


Figure 9.11: Experiment 3 - Combined AUC results

- The experiment, given the F1 score and PR-AUC, presented good results as almost all machine learning models managed to predict 70-85% of the deceptive accounts successfully. An optimal prediction value would be subjective, but the researcher expected a success rate of at least 50% or more. This would mean that at least half of the deceptive accounts would be detected.
- Results from 30 repeats of the experiment showed that the values were consistent with F1 scores and varied less than 2%.
- Entropy results showed that COMPARE_AGE, HAS_NAME, and LEVENSHTTEIN were the features that were most indicative of deception.

Based on these results, it was clear that the last of the three experiments performed the best. It was affirmed that using knowledge from a related field, solving for a similar problem, could aid in the detection of humans who lie about their identity on SMPs. By creating better features that are indicative of human deception, the accuracy of the developed models increased. A point could of course be reached where it would be difficult to further improve the accuracy of the model. In such scenarios other machine learning algorithms or hyperparameters can be experimented with for further improvement.

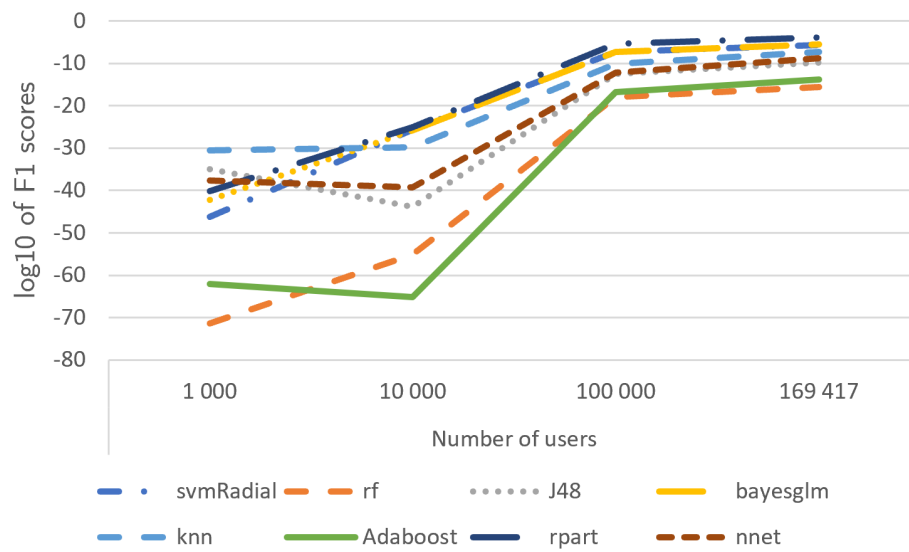


Figure 9.12: Experiment 3 - Various-sized dataset results

Table 9.14: Experiment 3 - Entropy results

	Machine learning algorithm							
	svmRadial	rf	J48	bayesglm	kkn	Adaboost	rpart	nnet
ACCOUNT_AGE_IN_MONTHS	29.580	12.607	29.580	29.580	29.580	29.580	0.000	11.433
COMPARE_AGE	97.517	97.068	97.517	97.517	97.517	97.517	96.667	66.295
COMPARE_GENDER	15.363	21.089	15.363	15.363	15.363	15.363	27.811	45.674
DISTANCE_LOCATION	0.894	0.802	0.894	0.894	0.894	0.894	0.948	28.166
DISTANCE_TZ	47.367	15.560	47.367	47.367	47.367	47.367	0.296	15.813
DUP_PROFILE	59.829	25.106	59.829	59.829	59.829	59.829	55.873	24.478
FF_RATIO	1.999	0.992	1.999	1.999	1.999	1.999	1.480	20.114
FOLLOWERS_COUNT	3.656	6.508	3.656	3.656	3.656	3.656	0.000	19.419
FRIENDS_COUNT	14.630	6.865	14.630	14.630	14.630	14.630	0.000	17.135
GEO_ENABLED	5.130	3.589	5.130	5.130	5.130	5.130	0.000	3.730
HAS_IMAGE	0.180	0.106	0.180	0.180	0.180	0.180	0.000	15.308
HAS_NAME	74.260	47.340	74.260	74.260	74.260	74.260	80.305	47.245
HAS_PROFILE	58.534	23.045	58.534	58.534	58.534	58.534	43.349	23.535
LEVENSHTEIN	77.507	55.244	77.507	77.507	77.507	77.507	90.679	60.824
LISTED_COUNT	3.125	2.695	3.125	3.125	3.125	3.125	3.333	20.443
PROFILE_HAS_URL	8.825	3.634	8.825	8.825	8.825	8.825	0.000	70.305
STATUS_COUNT	8.236	6.633	8.236	8.236	8.236	8.236	0.000	7.557
USERNAME_LENGTH	25.865	23.952	25.865	25.865	25.865	25.865	13.966	84.811

9.5 Comparing the results of Experiments 1, 2 and 3

Once all three experiments had been concluded, a comparison was made between the results (see Table 9.15), based on the F1 scores for the best model that emerged from each experiment. The comparison shows that the results for each experiment improved from the previous and assisted with the automated detection of human identity deception on SMPs by introducing new engineered features. The fact that the same machine learning algorithm built the best model for each experiment confirmed that the successes seen in later experiments were purely because of the introduction of the newly engineered features and not because of a specific algorithm that was better at handling certain data.

Furthermore, the entropy was used as a guide to add new features to improve the results from one experiment to the next. A summarised overview of the entropy results over all experiments is given in Figure 9.13, with the greyed areas indicating those attributes or features of importance. For Experiment 1, FRIENDS_COUNT and FOLLOWERS_COUNT were shown to be important in a model detecting deceptive humans. The same features were also important for the detection of bots [80] and therefore the researcher engineered features based on knowledge obtained from related work in bot detection to apply in Experiment 2. Interesting though was that the FF_RATIO was not used in Experiment 2 – probably because the data had already been cleaned from non-human accounts during data preparation. The deduction made here is that the same features applying to humans did not necessarily apply to bots. This deduction should however be made with caution, as the results from Experiment 1 were not good. It would be advisable not to deduce anything from the results of Experiment 1's machine learning model.

New engineered features, indicated as being important during Experiment 2, were closely related to work in the field of psychology showing what people lie about. Experiment 2 especially highlighted the fact that 'name' is an identity attribute being lied about. Therefore, additional features based on research work in psychology were engineered to be used for Experiment 3. Experiment 3 confirmed that name was a feature indicating deception and added age as an important feature when detecting identity deception by humans on SMPs.

Table 9.15: Overview of the best research per experiment

Experiment	Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	AUC-ROC (%)	PR-AUC (%)	Cost (seconds)
1	rf	79.31	23.13	23.03	57.15	32.83	76.11	28.57	98.164
2	rf	87.11	43.16	37.98	72.11	49.75	90.24	49.90	131.810
3	rf	97.49	84.86	83.89	88.72	86.24	98.91	93.00	157.801

Attributes contributing most towards deception in experiment 1

FOLOWERS_COUNT
FRIENDS_COUNT
LANGUAGE
LISTED_COUNT
PROFILE_IMAGE
STATUS_COUNT
TIMEZONE
UTC_OFFSET

Features contributing towards deception in experiment 2

ACCOUNT_AGE_IN_MONTHS
DUP_PROFILE
FF_RATIO
FOLLOWERS_COUNT
FRIENDS_COUNT
GEO_ENABLED
HAS_IMAGE
HAS_NAME
HAS_PROFILE
LANGUAGE
LISTED_COUNT
PROFILE_HAS_URL
STATUS_COUNT
USERNAME_LENGTH

Features contributing towards deception in experiment 3

COMPARE_AGE
COMPARE_GENDER
DISTANCE_LOCATION
DISTANCE_TZ
LEVENSTEIN

Figure 9.13: Entropy results across all three experiments

9.6 Conclusion

This chapter explored the detection of identity deception by humans through three experiments and supervised machine learning; the results of each experiment led to the next. It was found that engineered features used in the past to detect non-human or bot accounts performed better than when using the attributes found on SMPs alone. Furthermore, features engineered based on what psychology tells us about liars performed even better at developing a model that can assist in the automated detection of human identity deception on SMPs. It was noticeable that some features – specifically, the age and name of the user – contributed more towards the detection of identity deception than others.

It however remains essential to explain these predictions to some person or institution (e.g. law enforcement) that investigates identity deception. Most choices implemented by machine learning algorithms in the final models are difficult to explain. In the next chapter, the ‘detect’ component proposes to use the ‘discovered’ model that was presented in this chapter to assist with the automated detection of human identity deception on SMPs. Deceptive users will not only be detected through the use of a supervised machine learning model, but the reasons for a user being deemed deceptive or trustworthy will be intuitive.

Chapter 10

Prototype: Detect

“The trust of the innocent is the liar’s most useful tool” - Stephen King

10.1 Introduction

Identity deception is an example of a cyber threat found on Social Media Platforms (SMPs) where humans pose a threat to other humans. The research in hand proposes a model for detecting identity deception to be implemented through a prototype that assists in the automated detection of human identity deception on SMPs. The prototype consists of three main components. The first component *prepares* the data for experimentation. The second component experiments with the prepared data to *discover* a supervised machine learning model that can identify humans lying about their identity on SMPs. The third component uses the discovered models to assist in the automatic *detection* of deceptive humans on SMPs.

The previous chapter discussed the second main component of the prototype that ‘discovered’ various models to assist in the automated detection of human identity deception on SMPs. Three experiments were performed by using supervised machine learning. The first experiment used the attributes, available on SMPs only, to develop a model with which deceptive humans could be detected. The second experiment proposed to improve on the results obtained in the first experiment and engineered additional features that were known to have had success in detecting deception in related bot detection work. Features similar to those in Experiment 1 were engineered from the SMP attributes to improve on the previous model that used SMP attributes as is. The resultant model showed increased accuracy and also indicated that

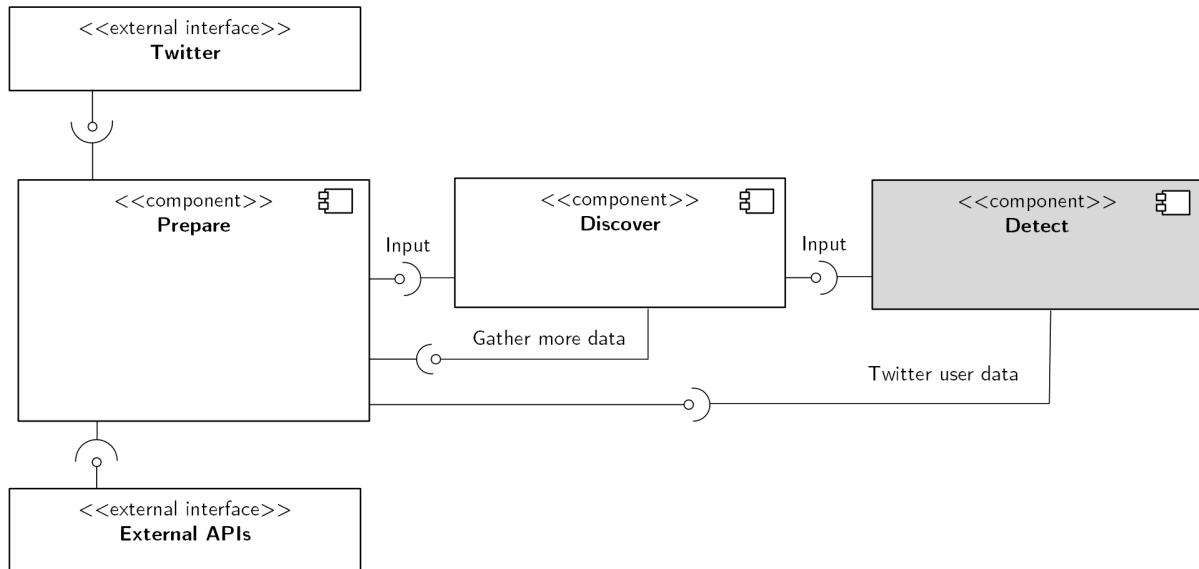


Figure 10.1: High-level overview of the prototype: ‘Detect’ component

knowledge from related work can aid in discovering more accurate models. Due to these reasons, a third experiment was conducted and the work done in Experiment 2 was used, together with results from related work in psychology, to develop a model that can assist in the automated detection of human identity deception. This model achieved the most accurate results, as demonstrated with evaluation metrics like accuracy, F1 score, and PR-ROC.

Chapter 10 discusses the ‘detect’ component – the last main component of the proposed prototype that solves for the expected requirements, namely that a human identity deception detection prototype should, among others, ensure that the machine learning results are reproducible, interpretative, and automated. The ‘detect’ component uses the results yielded by the previous components, as well as Twitter user data (see Figure 10.1) to assist in the automated detection of human identity deception on SMPs. This Twitter data differs from the original set that was used to discover a model for detecting human identity deception on SMPs. The reason for this is that once a model has been developed, it can be used to detect human identity deception on never seen before users. The developed supervised machine learning model was applied to this set of unlabelled SMP user data, similar to the data found at large in an SMP where it was not known upfront whether the user account was deceptive or trustworthy.

The ‘detect’ component consists of two sub-components aimed at presenting an Identity Deception Detection Model (IDDM). Firstly, the IDDM detected humans lying about

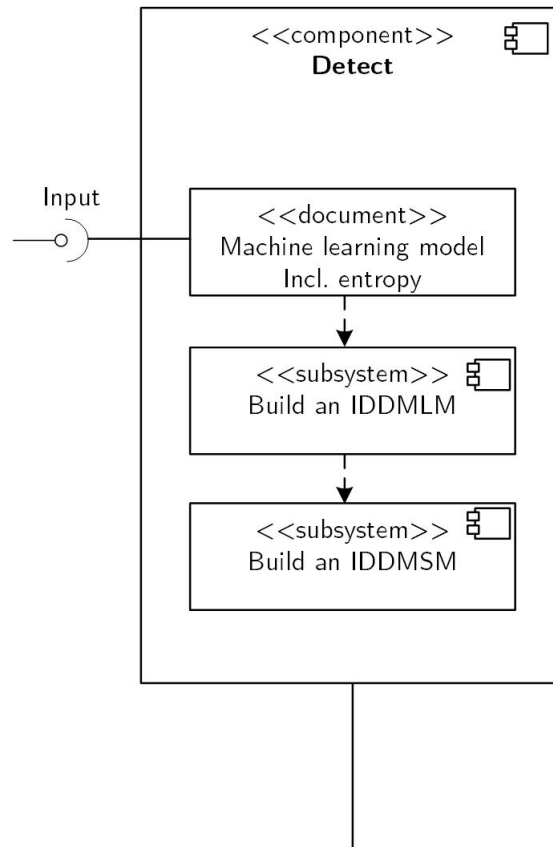


Figure 10.2: The ‘Detect’ component

their identities on SMPs through an Identity Deception Detection Machine Learning Model (IDDMLM) component. The results from the IDDMLM were then explained in an intuitive way through an Identity Deception Detection Score Model (IDSM) component. The two sub-components of the ‘detect’ component are illustrated in Figure 10.2.

This chapter starts off by discussing the current problems with the interpretation of supervised machine learning model results. This knowledge is required to explicate the decisions made for the IDDM. Each of the sub-components of the IDDM is then discussed in detail, before the chapter illustrates a working IDDM and compares it with other existing models.

10.2 Problems with using supervised machine learning to detect identity deception

Doshi-Velez and Been [101] present the need for interpretability in their research by inspecting what the machine learning models will be used for. They also propose a

common taxonomy with which interpretability can be measured. Ribeiro et al. [257] and Beillevaire [35] propose new methods for explaining predictions of machine learning models. Explaining decisions is important, as garbage inputs could result in garbage outputs without forewarning [17]. The correct interpretation is for example critical when the detection of deceptive users could have criminal consequences [57], ethics is involved [101], or the result has consequences for people's lives [257]. In May 2018, the interpretation or explanation of results became law in the European Union through the implementation of the General Data Protection Regulation (GDPR). This regulation ensures that European citizens have the right to remove their data from an organisation's database [132]. It also means that EU citizens have the right to receive an explanation on how decisions were reached about their personal data and algorithms [132].

Most machine learning models are difficult to interpret [28] [318], but some, like decision trees, have algorithms with simple logic that humans can understand [28]. Two of the many forms of decision tree machine learning algorithms are random forests and recursive partition (rpart), and the algorithms can be grouped as either regression type trees or classification type trees. Regardless of the type of tree or the simplicity of how decisions tree work, issues still arose during the interpretation of the results. To illustrate this fact, results from the last experiment (Experiment 3) were used. The results of the rpart algorithm were modelled as a decision tree (see Figure 10.3). A node represents a decision for a specific attribute or feature. For example, the first node split human accounts given whether their `DISTANCE_LOCATION` was less than/ greater than/ equal to -0.58. The intensity of a node's colour was proportional to the number of users classified as being either deceptive or trustworthy at that specific node. This means that a darker shade showed that the decision made at a specific node contributed more towards the detection of deceptive or trustworthy humans. Each node showed the following:

- The number at the top represented when each decision was made during the creation of the tree; with '1' being the first decision and so forth.
- The resultant class (deceptive or trustworthy).
- The probability of deception on the left and trustworthiness on the right.
- The percentage of total corpus observations included in the node.

The first node for example shows that 43% of the humans were found to be deceptive and 57% were found to be trustworthy. This also relates to findings from over 100 experiments by DePaulo et al. [93] who state that humans can detect lies from truth with a mean average of 54%.

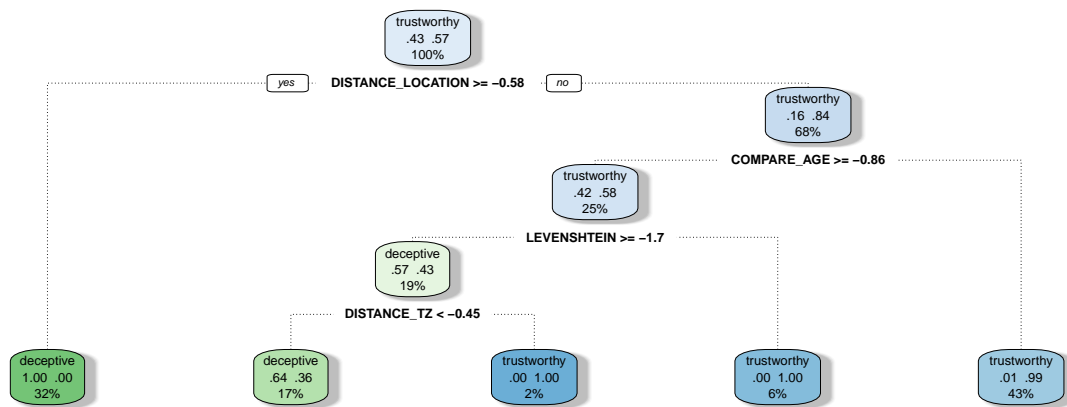


Figure 10.3: Tree representation of the rpart model that determines identity deception

Interpretation became less intuitive however, the further we traversed the tree. Figure 10.4 points out two nodes on the second level of the tree highlighted in red. Here it is shown that the node on the left contained 32% of all deceptive nodes and the node on the right contained 68% of all trustworthy nodes. This split was due to a `DISTANCE_LOCATION` feature value of -0.58. Together, these two nodes accounted for 100% of all nodes as expected. Some noteworthy interpretation issues were as follows:

- It is not intuitive what a `DISTANCE_LOCATION` (Figure 10.4) feature value of -0.58 means. The value was centred and scaled during data preparation as per the requirement for machine learning algorithms [191].
- Although there were less than 1% deceptive accounts in the corpus, the node on the left shows that 32% of humans were deceptive. This is due to the fact that oversampling was performed to cater for the skewness in the overall data set. Supervised machine learning models required labelled data to be equally represented in the dataset [339]. The results gave a false perception that 32% of all humans were deceptive.
- The tree was pruned to reduce overfitting. This is standard in most decision tree machine learning algorithms [245]. The pruning is visible by the missing numbers, like 4 and 5, at the top of the nodes. This means that some smaller decisions in the model were deemed irrelevant and omitted. There is no view on which of these decisions were omitted.

The above observations show the difficulty in interpreting results from machine learning.

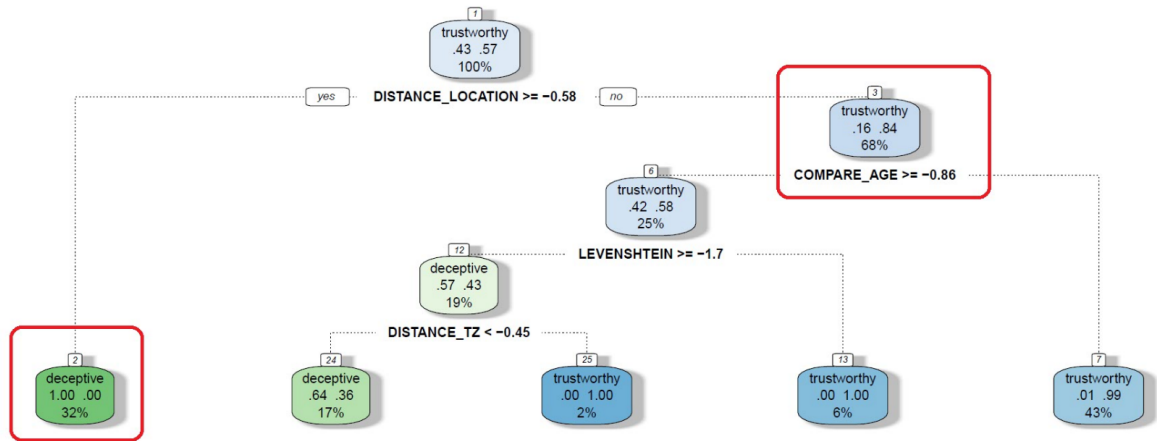


Figure 10.4: Explanation of the rpart tree

Even simple algorithms, which can intuitively be represented by trees, can sometimes lead to wrong interpretations. For this reason, the researcher looked towards other means of interpreting the results of a model that assists in the detection of human identity deception on SMPs.

10.2.1 Use entropy to explain identity deception detection results

Various methods to interpret machine learning models, such as LIME [257] and Shapley values [209], were presented in Chapter 5. The interpretation methods, proposed by related work, explained the results obtained from supervised machine learning models at an SMP user level (local) and for all SMP users in general (global). Both were important to this research. The first explained why specific SMP users were deemed to be deceptive. The second indicated, at large, those attributes or engineered features most indicative of deception so that future SMP platforms would be improved. It was however found that computational overhead was associated with the calculation of the machine learning model interpretations currently proposed by related work. For SMPs, the researcher proposed an alternative method using entropy, more specifically Shannon Entropy [273], to interpret machine learning models. Entropy indicates how much information is gained or lost when a new attribute or feature is introduced into a dataset [273]. This information was used to determine which attributes and engineered features were more important when deceptive identities were identified on SMPs.

Entropy is an outcome of the ‘discover’ component that preceded the ‘detect’ component

in the prototype proposed for this research. This means that the computation associated with model interpretation could be decreased as the information required to calculate the interpretation was already readily available. This was particularly valuable on an SMP platform like Twitter. Law enforcement, for example, required explanations about why a human was perceived as being deceptive to promptly protect humans at risk of malicious individuals lying about their identity.

10.3 Building an interpretable identity deception score

To assist in detecting human identity deception on SMPs, an IDDM was proposed. Based on the experimental results discussed in the ‘discover’ component and the problems encountered in interpreting supervised machine learning models, IDDM was structured to consist of the following two sub-components:

- Identity Deception Detection Machine Learning Model (IDDMLM): The IDDMLM used the supervised machine learning model developed by the ‘discover’ component. During experimentation with the ‘discover’ component, it was found that the random forest machine learning model, which used attributes and engineered features from related work in bots and psychology, most accurately detected humans lying about their identities on SMPs. The IDDMLM applied this random forest supervised machine learning model on new Twitter data that had not been used during previous experiments. The new Twitter data resembled SMP user data that was readily available in Twitter. By applying this model to the new Twitter data, these never-before-seen SMP users were scored as being potentially deceptive or trustworthy. The score was a value between 0% and 100% and indicated the level of that SMP user’s perceived deceptiveness (with 100% being perceived most as deceptive and 0% as trustworthy). At this point, the predicted deception score cannot be explained for reasons stated earlier.
- Identity Deception Detection Score Model (IDDSM). The IDDSM used the outputs of the IDDMLM together with entropy information available for the random forest model. The entropy values were applied ‘as is’ to the attribute and engineered feature values of the user. The final result was a score between 0 and 100 for each attribute and engineered feature that indicated their contribution to

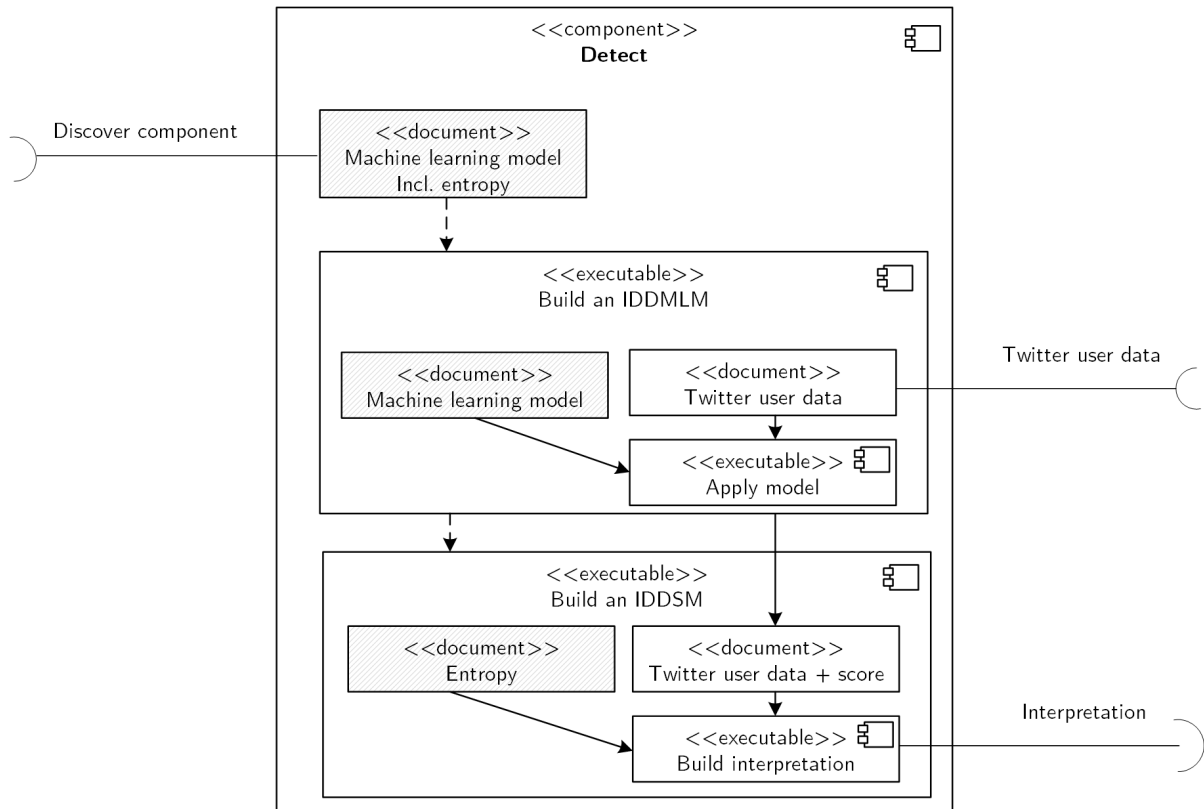


Figure 10.5: Detailed UML component diagram of the ‘detect’ component

the IDDMLM score. A score of 100 showed a high contribution and a score of 0 indicated no contribution of a particular attribute or engineered feature to the perceived deceptiveness score calculated by the IDDMLM.

The Unified Modeling Language (UML) model presented earlier in Figure 10.2 was extended in Figure 10.5 to illustrate the detail about the IDDMLM and IDSM sub-components.

10.3.1 The IDDMLM sub-component

The IDDMLM used the Random Forest algorithm, which is a collection of randomised decision trees [45]. The Random Forest algorithm was used as it performed the best during experimentation completed by the ‘discover’ component of the prototype. The ‘discover’ component iteratively used attributes and engineered features from related work in bots and psychology to discover the best model that could assist in the detection of human identity deception using SMP attributes.

The IDDMLM is presented in Algorithm 2.

Algorithm 2 The IDDMLM

Let $SMP = \{SMP_i: SMP_i \text{ is a Social Media Platform}\}$

Let $A = \{a_1, a_2, \dots, a_n\}$ be a subset of SMP_i attributes,

with $n \leq$ the number of attributes in SMP_i

Where:

A is periodically created,

$A = A^1 \cup A^2$,

$A^1 = \{a_1, a_2, \dots, a_n\}$ is a random extracted training data set with number
of deceptive examples = number of not deceptive examples,

$A^2 = \{a_1, a_2, \dots, a_n\}$ is a random extracted test data set

with Σ deceptive examples = Σ not deceptive examples,

Note: It is typical for A to be created and thereafter split into training and test data where A^1 contains 75% of A and the remaining 25% belongs to A^2

[220]

Let $F = \{f_1, f_2, \dots, f_m\}$ be a set of features,

$m =$ number of engineered features

Where:

$f_i \in A \vee f_i = f(a_j, \dots, a_k)$

Where:

$j \geq 1$,

$k \leq n$.

Let $RF = h(x| \theta_1), h(x| \theta_2), \dots, h(x| \theta_t)$

[52]

Where:

$RF =$ Random Forest algorithm,

$t =$ number of decision trees,

$h(x| \theta_i) =$ a single decision tree

Where:

$\theta_1 \subseteq ((F|A_1) \cup A_1)$,

$x =$ the values of $A_1 \vee F$ given θ_i ,

$1 \leq i \leq t$.

Note: For the final classification, each decision tree $h(x| \theta_1)$ casts a vote for the most popular output, given input x . The class with the most votes wins. These voting results are not visible or retrievable. This issue is known as the machine learning interpretability problem

[257].

Algorithm 2 The IDDMLM (continued)

Let $\text{RF}_{Results} = (f1_i, e_i)$: calculated for $\forall (a_i \vee f_i) \in \theta$ [45]

Where:

$\text{RF}_{Results}$ = Results of Random Forest,
 $f1_i$ = an F1 value, [170]

e_i = an Entropy value. [258]

Let $A^3 = a_i \vee f_i$: selected based on optimum values out of the set generated by $f(f_i, e_i)$ [52]

Where:

$1 \leq i \leq n$,
 $A^3 \subseteq A$.

Let M_i = final Identity Deception Score (IDS) for U_p

Where:

U_p is a user of SMP_i ,
 $M_i = \text{RF}_p = \{ h(x_p | \theta_1), h(x_p | \theta_2), \dots, h(x_p | \theta_t) \}$,
 x_p = values of $a_i \vee f_i \in A^3$ for U_p .

10.3.2 The IDDSM sub-component

The IDSM component used the output of the IDDMLM component. IDSM included an interpretation as to why the identity of a SMP user was perceived as deceptive or not.

The IDSM is presented as follows [314]:

Algorithm 3 The IDDSM

Let S_i be the Identity Deception Score (IDS) for U_p

Then

$$S_i = \sum_{i=1}^m f(w, x_p)_i$$

Where:

m = number of elements $a_i \vee f_i$ in A^3 ,

$f(w, x_p) = w|x_p|$,

$w \in [0,100]$,

x_p = values of $\{a_i \vee f_i\} \in A^3$

and x_p is derived for each a_i such that $(a_i \in U_2 - a_i \in U_1) = x_p$,

$w = e_i \in A^3$

If $S_i \sim M_i$ then w , together with x_p can be used to interpret the results of M_i for U_p .

10.4 Illustrating the working of IDDM

Table 10.1 shows IDDMMLM results for two Twitter users, with some features obfuscated due to privacy and ethical reasons. IDDMMLM detected one user as deceptive (U_1) and the other as trustworthy (U_2).

Looking at the results, it is not clear why U_1 was found to be deceptive with 94.80% certainty. To validate the IDDMMLM results shown in Table 10.1, a subset of tweets for each individual were presented for clarity. This is shown in Figure 10.6.

Based on the tweets shown in Figure 10.6, it is clear why the first individual could be perceived as trolling the profiles of celebrities and being deceptive, and the second not. Although the conclusion was still perhaps subjective, the IDDMMLM model was able to identify potential identity deceptiveness.

The IDSM attempted to explain the decisions given by IDDMMLM. Figure 10.7 shows the entropy values determined by RFResult in the form of A_3 . The entropy results were indicated by values between 0 and 100, with the latter being most influential. A_3 was subsequently used in the IDSM in order to detect identity deception.

Using the same examples (Table 10.1) as for the IDDMMLM sub-component, Table 10.2 presents the results for U_1 compared to U_2 according to the IDSM sub-component. The entropy is presented by filled bars to highlight those features most indicative of the U_1 deceptiveness compared to U_2 . The IDSM results therefore added an interpretation

Table 10.1: Results obtained from the IDDMLM model

SMP Attributes and Features (A)	Deceptive (U_1)	Trustworthy (U_2)
** ID	???	???
** SCREENNAME	???	???
** PROFILE_IMAGE	???	???
TIMEZONE	Quito	Pacific Time (US & Canada)
LOCATION	Narnia	Portland, OR
ACCOUNT_AGE_IN_MONTHS	76	79
COMPARE_AGE	10.57	40.72
COMPARE_GENDER	TRUE	FALSE
DISTANCE_LOCATION	N/A	N/A
DISTANCE_TZ	4 416.31	1 382.00
DUP_PROFILE	FALSE	FALSE
FF_RATIO	0.38	1.08
FOLLOWERS_COUNT	278	191 356
FRIENDS_COUNT	728	175 893
GEO_ENABLED	FALSE	TRUE
HAS_IMAGE	TRUE	TRUE
HAS_NAME	1	1
HAS_PROFILE	1	1
LEVENSHTEIN	9	9
LISTED_COUNT	0	4
PROFILE_HAS_URL	FALSE	FALSE
STATUS_COUNT	224	104 956
USERNAME_LENGTH	10	12
*IDDMLM	94.80%	2.40%

*IDt: high % = more deceptive

**Obfuscated for ethical reasons

feature to our IDDM model.

The results show that there were engineered features indicative of U_1 being perceived as deceptive. COMPARE_AGE, COMPARE_GENDER and DISTANCE_TZ were the most relevant indicators. Table 10.2 shows that U_1 was 10 years old when the Twitter account was opened, although the legal age to open a Twitter account is 13 [310]. Furthermore,

@infectionmalik. Follow me pleasee
@iswaaag. Follow me please
@VictorManuel03_ follow me please
@KevinArzate thanks
@KevinArzate follow me please
@VictorManuel03_ follow me please
@cesarraejepsen follow back please
@PedroNegrini follow me please!
@justinbieber follow me please please. I love u too

(a) Deceptive (U_1)

@s thank you... merry Christmas!!!! Have the best holiday.
@PM ..perfect timing.. lol
@a thank you!!! Happy Thursday!!!
Have a very merry blessed Thursday..
#Oregon #GoDucks fans
Love doesn't make the world go round
@et ..power was problem for most.. some areas still had solar and cell phones
@et ...yep... 110% She wasn't the only.. just the most famed

(b) Trustworthy (U_2)**Figure 10.6:** Tweets for individual users (U)

U_1 did not reveal their true location and it was not possible to determine the gender of U_1 from their profile image, which is an additional indicator of the fact that U_1 was potentially deceptive.

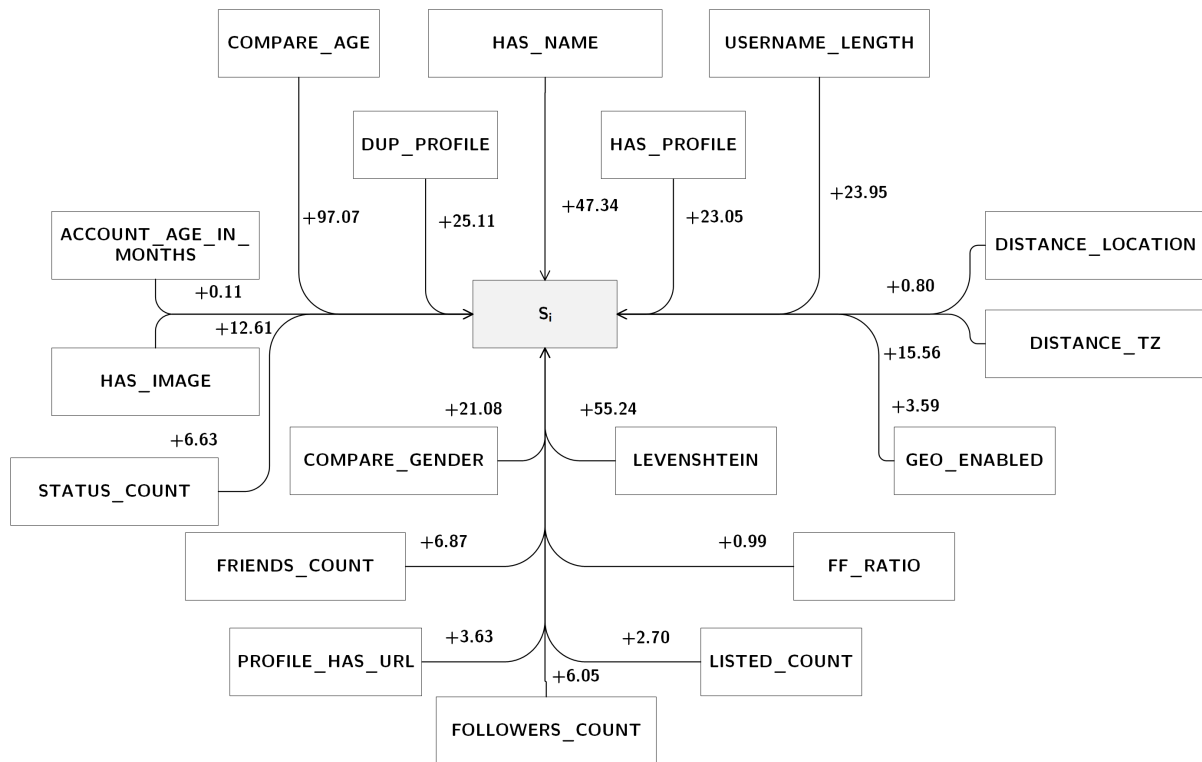


Figure 10.7: The IDDSM sub-component

Table 10.2: Results obtained from the IDDSM model

SMP Attributes and Features (A_3)	Explained (U_1)	Entropy
ACCOUNT_AGE_IN_MONTHS		0.48
COMPARE_AGE	<div style="width: 100%; height: 10px; background-color: red;"></div>	71.88
COMPARE_GENDER	<div style="width: 35%; height: 10px; background-color: red;"></div>	21.09
DISTANCE_LOCATION		0.80
DISTANCE_TZ	<div style="width: 20%; height: 10px; background-color: red;"></div>	10.69
DUP_PROFILE		0.00
FF_RATIO		0.64
FOLLOWERS_COUNT	<div style="width: 10%; height: 10px; background-color: red;"></div>	6.50
FRIENDS_COUNT	<div style="width: 10%; height: 10px; background-color: red;"></div>	6.84
GEO_ENABLED	<div style="width: 5%; height: 10px; background-color: red;"></div>	3.59
HAS_IMAGE		0.00
HAS_NAME		0.00
HAS_PROFILE		0.00
LEVENSHTTEIN		0.00
LISTED_COUNT	<div style="width: 5%; height: 10px; background-color: red;"></div>	2.70
PROFILE_HAS_URL		0.00
STATUS_COUNT	<div style="width: 10%; height: 10px; background-color: red;"></div>	6.62
USERNAME_LENGTH	<div style="width: 5%; height: 10px; background-color: red;"></div>	3.99

10.5 Comparing IDDSM with other interpretation methods

The IDSM component differed from other proposed machine learning interpretation methods in that the results required no further machine learning or repetitive processing. The LIME model [257], for example – if applied to the research data – took one user’s attributes and engineered features and then iteratively adjusted the values of these to create many examples of users with similar but not exactly the same attributes and engineered features as the original. A linear machine learning model was subsequently developed for that one user to explain the importance of each attribute or feature for the specific user’s perceived deceptiveness. If the original cost of one iteration of interpretation was represented as $O(1)$, then the cost of the LIME model would be $O(n)$ for n users, as a separate machine learning model was developed additionally for each user. Similar to LIME, Lundberg and Lee’s [209] method that uses Shapley values of game theory could be used to explain the perceived deceptiveness of an SMP user. The Shapely value [275] calculates the feature contribution by randomly omitting attributes or engineered features to develop new linear machine learning models and deceptive score results for the same user over many iterations. The combined result of all these models showed the overall contribution of each attribute or engineered feature for each SMP user. Computationally, this method’s cost could be expressed as $O(nk)$ where n was the number of SMP users and k the number of features that were included to explain the model. Due to this high cost, k was sometimes limited to not test for all combinations of features, despite the knowledge that the interpretative results might not be perfect in such scenarios.

Other machine learning interpretation models were dependent on the machine learning algorithm used and additional sample data was not required. An example of such a model was the ‘tree interpreter’ [263], which uses the knowledge gathered from all decision trees in a random forest machine learning model to interpret the final score produced by the random forest model for a user. For this method, each of the decisions trees produced by the random forest were developed on their own so that the combined individual decision tree results could explain the perceived deceptiveness of a user. In this scenario, the cost was represented as $O(t)$ where t was the number of decision trees generated by the model during development.

The IDSM proposed for this research, on the other hand, only used the results from the original developed supervised machine learning model. These results could be

periodically updated as better identity deception detection models were developed, but this did not influence the tree interpretation cost. The cost of the IDSM was thus only to compare one user with another – which translated to $O(2)$. This low cost was imperative for SMPs that dealt with large volumes of data and required prompt action to be taken when a human was at risk of being attacked by another malicious individual on an SMP.

10.6 Conclusion

In this chapter, the ‘detect’ component of the prototype was represented as an IDDM. The IDDM was proposed not only to assist with detection, but also to interpret perceived deceptiveness. The IDDM consisted of two sub-components: the first, IDDMLM, used input from prior experiments to score the perceived deceptiveness of new SMP users that had never been seen before. The second component, IDSM, interpreted the results obtained from IDDMLM by means of a simple weighted linear formula, given the known entropy of the features involved. The IDSM results highlighted those features that a specific user was found to be most likely deceptive about. This was invaluable in use cases where that particular user had to be investigated further.

The next chapter concludes the research by revisiting the research questions to show how this research addressed each. It also states the contribution of this research work, and presents final thoughts on potential future research directions to pursue so as to improve on the detection of the cyber threat of human identity deception on SMPs.

Part IV

Conclusion

Chapter 11

Conclusion

“The journey of a thousand miles begins with one step” - Lao Tzu

11.1 Introduction

This thesis addressed the cyber threat of identity deception by humans on Social Media Platforms (SMPs) and proposed a model that can assist in the automated detection of human identity deception on SMPs. The research showed that many attributes exist across various SMPs to describe the identity of a human. Identity attributes were found to differ from the content that SMP account holders post on a daily basis, as the identity attributes are mostly added when an account is opened, and they are only changed on an *ad hoc* basis. For example, if a woman gets married, she might change her surname on the SMP. On the other hand, a person’s birth date will never change. Besides these attributes, the researcher looked at related work performed to detect deceptive non-humans or bots on SMPs, as well as at findings from the field of psychology about the issues that humans lie about the most. This knowledge was transformed into additional engineered features to propose a model that could assist in the automated detection of human identity deception on SMPs.

The model was built following a number of research steps. Each of these steps was implemented via a bespoke prototype consisting of three components. The first component prepared the data for the model, while the second component used the prepared data to discover a model by experimenting with various supervised machine learning models and combinations of attributes and engineered features. The last component proposed a model to detect human identity deception, also known as the

Identity Deception Detection Model (IDDM). The IDDM consisted of two sub-models. The IDDM first used a supervised machine learning model to detect deception by means of the Identity Deception Detection Machine Learning Model (IDDMLM). The IDDM subsequently provided an interpretation of the results by scoring the attributes and features used to detect identity deception through an Identity Deception Detection Score Model (IDSM). The IDSM indicated which attributes or features contributed most to a human being indicated as potentially deceptive.

In this final chapter, the research questions are revisited to evaluate the extent to which the primary and secondary research problems have been addressed in this thesis. This is followed by an assessment of the main contributions of the research. The chapter concludes by suggesting directions for future research forthcoming from this work.

11.2 Revisiting the problem statement

The overall objective of this research was to address the cyber threat of identity deception by humans on SMPs by proposing a model that assists in the automated detection of identity deception. Secondary to that, new features that emerged from related work were evaluated to understand their potential application to the domain of humans lying about their identity on SMPs. These objectives were accomplished by answering the identified research questions:

Research Question 1: What are the cyber threats found on SMPs and why is it important to find a solution to the problem of identity deception by humans on SMPs as opposed to bots?

Chapter 2 showed that all threats either result from some form of malware, abuse some known network flaw, or are personal in that they are aimed at a human or SMP account. For the purposes of this research, those threats aimed at a human were of particular interest. With personal threats, individuals are vulnerable to a range of attacks that are possible on SMPs. Cyber threats can be in the form of identity theft, trolling, flaming, identity deception, cyber stalking, cyber bullying, grooming, or phishing. SMPs have increased the risk to individuals merely by the extent of their exposure to these threats.

Identity deception was found to be an important cyber threat aimed at humans on SMPs. Identity deception involves humans lying about who they are and not necessarily about what they say or post on the SMP. The challenge is that it is very difficult to prove that

people are who they say they are. Deceivers present themselves as someone else to gain the trust of innocent individuals and lure them for some malicious purpose. Related work has done much to detect deception at large from non-human or bot accounts, and examples of such bot accounts are readily available on SMPs.

The research in hand proposed to focus on detecting malicious humans, because these malicious humans target specific individuals whilst bots are mostly found to target groups. According to the researcher, this targeted threat has a very specific focus, for instance the cyber bullying of an individual, which could have severe consequences for the target – in some cases even death. Detecting the identity deception committed by these malicious individuals warrants the adoption of different approaches, as the objective is to find these few malicious individuals – as opposed to bots, which are generated at mass scale.

Research Question 2: What attributes are available in SMPs that have the potential to be used for identity deception by humans? A thorough study (reported on in Chapter 3) was performed in respect of the attributes found within the current top six SMPs in the world. The great similarity that was found between the attributes of these SMPs will be beneficial for future research, as it should be possible to apply the current research to other SMP platforms than Twitter.

The research in hand revealed that the SMP attributes either describe the human, their account, behaviour, relationships, or content. Deception is mostly prevalent in the description of the human and the content. The other attributes are either not editable, like the start date of the account, or they depend on their activity on the SMP, such as befriending someone or posting content only over weekends. It was also noticeable that most humans generated content on a daily basis. More content would require more time to detect deception, because in order to analyse content, different techniques like Natural Language Processing (NLP) are required. NLP is still considered difficult [238]. An example is the detection of sarcasm where researchers achieved an accuracy just better than random on a biased dataset from Twitter [238]. This additional overhead in processing the content is not acceptable in a scenario where deception detection is required in real time because the individual's life is being threatened. In addition, related work in bot detection has found that by ignoring content on SMPs, similar accuracy could be achieved in the detection of deception [80] [317]. Therefore, this research focused on identity deception found in the attributes that describe the human only, and content was excluded.

Research Question 3: What are the requirements for a model that will assist

in the automated detection of human identity deception on SMPs and how can such a model be implemented?

Chapter 3 discussed the requirements for a model to assist in the automated detection of human identity deception on SMPs. These requirements were aimed at finding those humans (as opposed to bot accounts) who are being deceptive about their identity for malicious purposes on SMPs. The reason for this was that humans target mostly other individuals, with malicious intent. Since such cyber attacks by humans are difficult to detect in the volumes of SMP data produced daily, the researcher decided to use only those attributes that define a user account for the purposes of detecting identity deception. The content could be ignored as it was found in related research that identity attributes were just as accurate in detecting deception and that the additional overhead of processing content was not required. Features should be engineered in such a way that they complement the automated detection of identity deception. In other words, features were directly engineered from attributes usually available on SMPs. Lastly, the model was required to be interpretive and reproducible.

The requirements mentioned were proposed to be implemented through various research steps that would culminate in a prototype. The research steps towards developing this prototype were discussed in detail in Chapter 5 and the three prototype components were described in Chapter 6. The first component, discussed in Chapter 8, involved the gathering of data from Twitter and its preparation for supervised machine learning. The preparation included the cleaning of the data from non-human accounts, the combination of the labelled examples of ‘deceptive’ accounts with the gathered set of ‘trustworthy’ accounts, and the preparation of the data for machine learning. The second component, discussed in Chapter 9, involved the discovering of a model with which to detect human identity deception through experimentation with various supervised machine learning models and combinations of attributes and features. The results obtained from these experiments were evaluated to be able to understand which attributes and features contributed most towards identifying those individuals lying about their identity on SMPs. The last component, discussed in Chapter 10, used the knowledge gathered during discovery, to automatically detect identity deception by humans on SMPs.

The proposed model (IDDM) was divided into two sub-components. The first sub-component (IDDMLM) used the machine learning model to detect potential deceptive individuals. The second sub-component (IDDSM) explained the results from the IDDMLM by scoring those attributes and features used to detect potential

deceptive individuals. These scores provided valuable input as to why a specific individual was perceived as being deceptive or not.

Research Question 4: Can features from related research in the detection of non-human or bot accounts and knowledge about deception in the field of the social sciences contribute towards the detection of identity deception?

Chapter 4 discussed two research fields (social sciences and bots) in which deception is prevalent. Related work from these fields showed how features could be engineered to detect non-human accounts on SMPs. Based on research in the field of psychology, humans are most likely to lie about features like their image, name, location, age, and gender. These features (proposed by the related work in non-human accounts and psychology) were engineered for the current research as part of the ‘prepare’ component of the prototype (discussed in Chapter 8) to make these new features available for experimentation. The ‘discover’ component of the prototype (discussed in Chapter 9) used the prepared engineered features together with the existing SMP attributes in various experiments. For each experiment, the same supervised machine learning algorithms were used, and only different combinations of attributes and features were introduced. It is however important to note that only standard SMP attributes were used. This strategy enabled the researcher to determine whether certain features contribute towards the detection of identity deception.

Chapter 9 presented the results of three experiments as part of the implementation of the ‘discover’ component of the prototype. The first experiment used the attributes found on SMPs without any change. The results showed that identity deception could not be detected using these attributes alone, and the random forest machine learning model achieved an F1 score of 32.83%. (An F1 score of at least 50% is expected for a model to be able to predict identity deception at random.) The second experiment proposed that features from past research be used in the detection of non-human accounts, and the random forest machine learning algorithm achieved an F1 score of 49.75%. This result shows that when features used to detect non-humans were added, identity deception by humans could be detected with greater precision, but the result was still worse than a random classifier. The third experiment added observations from past research work in psychology and the reasons why people lie. The results from this random forest machine learning algorithm showed an F1 score of 86.24%. The researcher concluded from the third experiment that new features related to the detection of non-human accounts and knowledge about deception from psychology together contributed to the detection of human identity deception on SMPs.

Research Question 5: Can we explain the model results in a format that is interpretable without any prior knowledge of machine learning, to show which attributes and features were most valuable in the detection of human identity deception?

It was shown, through experimentation, that identity deception by humans on SMPs can be detected through supervised machine learning and the right combination of SMP attributes and features. Unfortunately, the results obtained from supervised machine learning models are not intuitive. In a potential criminal scenario this is critical, and the predictive results from a machine are needed for guidance and prioritisation.

Chapter 10 proposed an IDDM as solution and divided the IDDM into two sub-models. The first sub-model used the fact that the ‘discovery’ component of the prototype found that attributes and features from related work in the detection of non-human accounts and in psychology contributed to and assisted with the automated detection of human identity deception on SMPs. The random forest machine learning algorithm performed the best, given the F1 scores. Entropy indicated which attributes and features were used to develop the supervised machine learning model. The IDDMMLM used this knowledge to predict whether Twitter accounts that had not been used during the experimentation were potentially deceptive or not.

The second sub-model used the entropy results produced in the model discovery component to explain the reasons for the predictions of the IDDMMLM. The IDSM presents a breakdown of the attributes and features most important to classify a particular individual as ‘deceptive’ or ‘trustworthy’. By using this information, it was possible to explain the model results in a format that is interpretable, without any prior knowledge of machine learning.

11.3 Main contributions

The *first* contribution made by the current research is that various attributes were identified, across multiple SMPs, that have the potential to develop a model that can assist in the automated detection of human identity deception by humans on SMPs. Even though this research was performed only with data gathered from Twitter, the findings can be applied to other SMPs due the similarity found in their available attributes. Furthermore, the research showed that not all data was required to detect identity deception. The content could for example be omitted, and it would still be

possible to provide an accurate prediction with less processing time and complexity. This is important in scenarios of identity deception like paedophilia and cyber bullying, where speed is of the essence.

The *second* contribution made by the current research is that it identified features that assist with the automated detection of identity deception by humans on SMPs. The research showed that the attributes found on SMPs alone cannot successfully detect identity deception by humans. Additional engineered features presented in related research work aimed at the detection of non-human accounts and in psychology increased the performance of the identity deception detection models. In fact, the identified features performed better than a model that would have predicted identity deception at random. This research showed the value of using knowledge from other research fields where similar deception problems are encountered.

The *third* contribution made by the current research work was the presentation of an IDDM. The IDDM is not only able to detect identity deception by humans on SMPs in an automated way by using a IDDMMLM, it also explains the results in an intuitive way. The IDSM predicts each human account by considering the contribution of each attribute and feature used. The attributes and features are scored to indicate their relative importance. In this way, one would intuitively understand the prediction of the IDDMMLM without having any prior knowledge of machine learning.

The *fourth* contribution made by this research involves the presentation of a prototype to assist in the automated detection of human identity deception on SMPs. This prototype can be implemented to predict identity deception by humans in an automated way. The researcher believes that the prototype can help to detect identity deception beforehand and therefore to prevent crime, rather than to react – only when it is already too late.

The *fifth* contribution was to create a research environment that enables experimentation with SMP data for the purposes of detecting identity deception. The same steps can be followed to conduct further related research. It took the researcher more than two years to set up an appropriate environment and other researchers can in the future accelerate their experimental work by learning from this research.

The *sixth* contribution was that the research results presented in this thesis show great promise in searching for solutions to cyber-security problems by focusing on the convergence of cyber security, big data and data science.

11.3.1 Advancing the state of the art

This research differed from other approaches in the following way:

- The research focused on finding deceptive humans rather than bots. This strategy can be compared to finding a needle in a haystack. Each individual is able to behave differently, whereas bots are usually created in groups and they all share some characteristic (e.g. their names are very similar).
- Only those attributes usually available on SMPs were required. Other research, like a model to predict loan propensity, requires a person's financial details, location, details about their family, etc.
- The research does not focus on content. People can tell a white lie, be sarcastic or even be honest, but still fake who they are for some malicious purpose. It is difficult to tell from content alone whether someone is lying about their identity. Think of online dating where the chats between people are honest, but most often the people lie about something that identifies them, like their age, hair colour, length, etc.
- The research requires no manual intervention. In some studies crowdsourcing is used to classify content in order to be able to develop a model. This research allowed for appending known deceptive human example accounts without crowdsourcing. Deceptive accounts were automatically generated through available Application Program Interfaces (APIs) and their deceptiveness was confirmed with statistical tests.
- The results of the research are a prototype that is implementable in the real world.
- The results from the models are interpretable. Many machine learning algorithms are regarded as black box. In the current research we added a model that interprets the machine learning results and approximates why a human was potentially perceived as being deceptive.

The research in hand furthermore used the knowledge gained from related research in the detection of non-human accounts on SMPs and in psychology. Knowing how other researchers from other research fields managed to detect deception was shown to benefit the outcome of this research. Using the knowledge from related research helped the researcher to

- understand which SMP attributes could be more indicative of deception;

- indicate the SMP attributes that could potentially be discarded earlier, as related work had already proved them to be irrelevant to deception;
- engineer new features that worked in the respective fields (e.g. friend to follower ratio worked well to find non-human accounts);
- understand how to work with data at scale (finding non-human accounts on SMPs containing large volumes of data);
- understand human deception intricacies (e.g. psychology research work focuses on why humans behave in a certain way).

11.4 Future work

This research proposed a final prototype that can assist in the automated detection of human identity deception on SMPs. Although the results of the prototype manifest in an intuitive way that is interpretable, there is still much room for future research work to supplement and support the current findings.

In general, the following research is proposed to be conducted in future:

- Investigating more features to be used in the detection of identity deception by humans. These features could either come from current SMPs, external sources or other research fields, or the human could even be asked to confirm certain facts (verification).
- Performing the same research on another SMP to evaluate and compare the results with the existing Twitter platform.
- Investigating the combination of a human's accounts across various SMPs. It would be interesting to know whether identity deception can be detected by using the attributes from another SMP as validation. The difficulty would be to match a human's account from one SMP to the next.
- Humans potentially deceiving differently across their different SMP accounts. The identity deception detection mechanism should cater for this fact.
- Humans potentially deceiving differently across different time periods. The human can for example tell a lie most times and then update the profile only when they want to deceive. The Identity Deception Score (IDS) should take cognisance of time as a factor.

- Considering how discrete and continuous attributes might benefit from different strategies in data cleaning but also detecting deceptiveness in them.
- Further investigating how existing SMPs can be re-engineered to improve the detection of identity deception.

In terms of data preparation, the following future research is proposed:

- Other mathematical methods could be applied to validate the data samples as representative of the population. The work performed for this research was very labour intensive. Better automated and more sophisticated methods could be investigated.
- Unsupervised learning could be used to label the gathered dataset. The idea is to create clusters of similar data from the original dataset gathered in Twitter and use these clusters as the labels. Future research could try to use supervised machine learning to predict these labels and, with the entropy results, try to understand the make-up of each cluster. One of these clusters could potentially be indicative of deceptive accounts.
- Some form of Turing test or Winograd Schema could be used to label the data. Perhaps a new form of test could be devised to identify trustworthy individuals.

In terms of machine learning and model interpretation, the following future research is proposed:

- Using a larger number of supervised machine learning algorithms.
- Changing the hyperparameters used in the algorithms. For this research, only the default hyperparameters were used.
- Defining another method for model interpretation and comparing with the IDSM.
- Adding confidence intervals for deception (similar to the Wilson score used in regression problems) to the results obtained from these deception detection classification models.
- Proposing new methods to determine the optimal data size required to develop the machine learning model. The elbow method currently used in clustering could be adapted for this purpose.
- Using adversarial machine learning techniques to generate fake individuals that could fool the current proposed IDDM. This could strengthen the model to cater for more serious future attacks.

Bibliography

- [1] Robert Abbott, Brian Ecker, Pranav Anand, and Marilyn Walker. Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *Language Resources and Evaluation Conference, LREC*, 2016.
- [2] ABI Research. More than 30 billion devices will wirelessly connect to the internet of everything in 2020. *ABI research*, 2013.
- [3] Kemal Veli Acar. Sexual extortion of children in cyberspace. *International Journal of Cyber Criminology*, 10(2):110–126, 2016.
- [4] Sareh Aghaei, Mohammad Ali Nematbakhsh, and Hadi Khosravi Farsani. Evolution of the world wide web: From web 1.0 to web 4.0. *International Journal of Web & Semantic Technology (IJWesT)*, 3(1), 2012.
- [5] Mohiuddin Ahmed, Abdun Naser Mahmood, and Jiankun Hu. A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60:19–31, 2016.
- [6] Airbnb Engineering. Data infrastructure at Airbnb, 2016. Available online: <https://medium.com/airbnb-engineering/data-infrastructure-at-airbnb-8adfb34f169c> (Accessed: 4 March 2018).
- [7] Mohammed Ali Al-garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the Twitter network. *Computers in Human Behavior*, 63:433–443, 2016.
- [8] Jan Philipp Albrecht. How the GDPR will change the world. *Eur. Data Prot. L. Rev.*, 2:287–289, 2016.
- [9] Abbas Raza Ali. Real-time big data warehousing and analysis framework. In *Big Data Analysis (ICBDA), IEEE 3rd International Conference on*, pages 43–49. IEEE, 2018.

- [10] Azliza Mohd Ali, Plamen Angelov, and Xiaowei Gu. *Detecting Anomalous Behaviour Using Heterogeneous Data*, pages 253–273. Springer, 2017.
- [11] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. Report, National Bureau of Economic Research, 2017.
- [12] Jalal S Alowibdi, Ugo A Buy, S Yu Philip, Sohaib Ghani, and Mohamed Mokbel. Deception detection in Twitter. *Social Network Analysis and Mining*, 5(1):1–16, 2015.
- [13] Amazon. Mechanical Turk, 2017. Available online: <https://www.mturk.com/> (Accessed: 8 January 2018).
- [14] Amazon. Amazon web services (AWS), 2018. Available online: <https://aws.amazon.com/big-data/> (Accessed: 31 May 2018).
- [15] Yoo Jung An, Kuo-chuan Huang, and James Geller. Naturalness of ontology concepts for rating aspects of the semantic web. *Communications of the IIMA*, 6(3):63–76, 2015.
- [16] Davide Anguita, Luca Ghelardoni, Alessandro Ghio, Luca Oneto, and Sandro Ridella. The ‘k’ in k-fold cross validation. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 441–446, 2012.
- [17] Anonymous. Ceding powers of decision to AI presents a paradox, 2017. Available online: <https://www.ft.com/content/63542534-ebf6-11e7-bd17-521324c81e23> (Accessed: 4 January 2018).
- [18] Apache Software Foundation. The Hadoop distributed file system: Architecture and design, 2014. Available online: <http://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html> (Accessed: 16 Jul 2014).
- [19] Apache Software Foundation. Apache Hadoop, 2018. Available online: <http://hadoop.apache.org/> (Accessed: 4 Mar 2018).
- [20] Apache Software Foundation. Apache Impala, 2018. Available online: <https://impala.apache.org/> (Accessed: 5 March 2018).
- [21] Scott D Applegate and Angelos Stavrou. Towards a cyber conflict taxonomy. In *5th International Conference on Cyber Conflict (CyCon)*, pages 1–18. IEEE, 2013.

- [22] Darren Scott Appling, Erica J Briscoe, and Clayton J Hutto. Discriminative models for predicting deception strategies. In *Proceedings of the 24th International Conference on World Wide Web*, pages 947–952. ACM, 2015.
- [23] S. S. Aravinth, A. Haseenah Begam, S. Shanmugapriyaa, S. Sowmya, and E. Arun. An efficient Hadoop frameworks Sqoop and Ambari for big data processing. *IJIRST –International Journal for Innovative Research in Science & Technology*, 1(10):252–255, 2015.
- [24] Muhammad Hassan Arif, Jianxin Li, Muhammad Iqbal, and Kaixu Liu. Sentiment analysis and spam detection in short informal text using learning classifier systems. *Soft Computing*, pages 1–11, 2017.
- [25] Keith Armstrong and Arron Hunt. Random user generator, 2017. Available online: <https://randomuser.me/> (Accessed: 8 January 2018).
- [26] Taylor B Arnold and John W Emerson. Nonparametric goodness-of-fit tests for discrete null distributions. *The R Journal*, 3(2):34–39, 2011.
- [27] By Press Association. Woman jailed for abusing boy after grooming him over Facebook, 2017. Available online: <http://www.dailymail.co.uk/wires/pa/article-5206541/Woman-jailed-abusing-boy-grooming-Facebook.html> (Accessed: 22 December 2017).
- [28] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.
- [29] Jimit Bagadiya. 171 amazing social media statistics you should know, 2018. Available online: <https://www.socialpilot.co/blog/social-media-statistics> (Accessed: 4 March 2018).
- [30] Bingdong Bai, Jing Chen, Mei Wang, and Jingjing Yao. Application research on big data in energy conservation and emission reduction of transportation industry. In *IOP Conference Series: Earth and Environmental Science*, volume 69, page 012029. IOP Publishing, 2017.
- [31] Simran Bajaj, Niharika Garg, and Sandeep Kumar Singh. A novel user-based spam review detection. *Procedia computer science*, 122:1009–1015, 2017.

- [32] Kapil Bakshi. Considerations for big data: Architecture and approach. In *IEEE Aerospace Conference*, Series Considerations for big data: Architecture and approach, pages 1–7, 2012.
- [33] Anil Bandhakavi, Nirmalie Wiratunga, P Deepak, and Stewart Massie. Generating a word-emotion lexicon from emotional tweets. In *Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics (SEM)*, 2014.
- [34] David Bartok, Estee van der Walt, Jan Lindemann, Johannes Eschrig, and Max Plauth. *Proceedings of the Third HPI Cloud Symposium "Operating the Cloud" 2015*. Universitätsverlag, 2015.
- [35] Marc Beillevaire. *Inside the Black Box: How to Explain Individual Predictions of a Machine Learning Model*. Thesis, KTH Royal Institute of Technology, 2017.
- [36] Gema Bello-Orgaz, Jason J Jung, and David Camacho. Social big data: Recent achievements and new challenges. *Information Fusion*, 28:45–59, 2016.
- [37] Mohamed Ben Khalifa, Rebeca P Díaz Redondo, and Ana Fernández Vilas. Why are these people there? an analysis based on Twitter. In *6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, pages 1–6. IEEE, 2015.
- [38] Mohamed Ben Khalifa, Rebeca P Díaz Redondo, Ana Fernández Vilas, and Sandra Servia Rodríguez. Identifying urban crowds using geo-located social media data: a Twitter experiment in New York city. *Journal of Intelligent Information Systems*, pages 1–22, 2016.
- [39] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida. Detecting spammers on Twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, volume 6, pages 12–21, 2010.
- [40] Emilia Bergen, Julia Davidson, Anja Schulz, Petya Schuhmann, Ada Johansson, Pekka Santtila, and Patrick Jern. The effects of using identity deception and suggesting secrecy on the outcomes of adult-adult and adult-child or-adolescent online sexual interactions. *Victims & Offenders*, 9(3):276–298, 2014.
- [41] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First Monday*, 12(11), 2016.

- [42] Nikita Bhagat, Ginni Bansal, and Bikrampal Kaur. Trends in cloud computing and big data. *International Journal of Science, Technology and Management*, 5(2), 2015.
- [43] Anol Bhattacharjee. *Social science research: Principles, methods, and practices*. CreateSpace Independent Publishing Platform, 2nd edition, 2012.
- [44] Sasanka Bhimavaram and P Govindarajulu. An enhanced approach for ontology based classification in semantic web technology. *Structure*, 4(2), 2015.
- [45] GÅŠrard Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13(Apr):1063–1095, 2012.
- [46] Netflix Technology Blog. Evolution of the Netflix data pipeline, 2016. Available online: <https://medium.com/netflix-techblog/evolution-of-the-netflix-data-pipeline-da246ca36905> (Accessed: 4 March 2018).
- [47] Netflix Technology Blog. Scaling time series data storage — part I, 2018. Available online: <https://medium.com/netflix-techblog/scaling-time-series-data-storage-part-i-ec2b6d44ba39> (Accessed: 4 March 2018).
- [48] Dasha Bogdanova, Paolo Rosso, and Thamar Solorio. Exploring high-level features for detecting cyberpedophilia. *Computer Speech & Language*, 28(1):108–120, 2014.
- [49] Charles F Bond and Bella M DePaulo. Accuracy of deception judgments. *Personality and social psychology Review*, 10(3):214–234, 2006.
- [50] Cheryl Lynn Booth and Shuyuan Mary Ho. Get a clue! some truths about online deception. In *IConference Proceedings*, 2016.
- [51] Taylor L Booth. *Sequential machines and automata theory*. John Wiley & Sons Inc, 1967.
- [52] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [53] Norman E Breslow and David G Clayton. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25, 1993.
- [54] Axel Bruns, Dr Dr Katrin Weller, Michael Zimmer, and Nicholas John Proferes. A topology of Twitter research: Disciplines, methods, and ethics. *Aslib Journal of Information Management*, 66(3):250–261, 2014.

- [55] Zhan Bu, Zhengyou Xia, and Jiandong Wang. A sock puppet detection algorithm on virtual spaces. *Knowledge-Based Systems*, 37:366–377, 2013.
- [56] David B Buller and Judee K Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.
- [57] Jenna Burrell. How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):1–12, 2016.
- [58] Canonical Ltd., Ubuntu community. Ubuntu, 2017. Available online: <https://www.ubuntu.com/> (Accessed: 31 July 2018).
- [59] Avner Caspi and Paul Gorsky. Online deception: Prevalence, motivation, and emotion. *CyberPsychology & Behavior*, 9(1):54–59, 2006.
- [60] Nick Castellina. SaaS and Cloud ERP trends, observations, and performance. *Analyst Inside*, 2011.
- [61] James L Cebula and Lisa R Young. A taxonomy of operational cyber security risks. Report, Carnegie-Mellon University Pittsburgh Software Engineering Institute, 2010.
- [62] Andrea Ceron, Luigi Curini, and Stefano Maria Iacus. *Politics and Big Data: Nowcasting and Forecasting Elections with Social Media*. Taylor & Francis, 2016.
- [63] Paul E Ceruzzi. *A history of modern computing*. MIT press, 2003.
- [64] Dave Chaffey. Global social media research summary, 2018. Available online: <https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/> (Accessed: 23 June 2018).
- [65] R Chandramouli. Emerging social media threats: Technology and policy perspectives. In *Cybersecurity Summit (WCS)*, pages 1–4. IEEE, 2011.
- [66] Thibaud Chardonens. *Big Data analytics on high velocity streams*. Thesis, University of Fribourg (Switzerland), 2013.
- [67] William R Cheswick, Steven M Bellovin, and Aviel D Rubin. *Firewalls and Internet security: repelling the wily hacker*. Addison-Wesley Longman Publishing Co., Inc., 2003.
- [68] Hyunyoung Choi and Hal Ronald Varian. Predicting the present with Google trends. *The Economic Record*, 88(S1):2–9, 2012.

- [69] Neelam Choudhary and Ankit Kumar Jain. *Towards Filtering of SMS Spam Messages Using Machine Learning Based Technique*, pages 18–30. Springer, Singapore, 2017.
- [70] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Who is tweeting on Twitter: human, bot, or cyborg? In *Proceedings of the 26th annual computer security applications conference*, pages 21–30. ACM, 2010.
- [71] Younghak Chun, Hyun Suk Hwang, and Chang Soo Kim. Development of a disaster information extraction system based on social network services. *International Journal of Multimedia and Ubiquitous Engineering*, 9(1):pp.255–264, 2014.
- [72] Roger Clarke. Human identification in information systems: Management challenges and public policy issues. *Information Technology & People*, 7(4):6–37, 1994.
- [73] Niall J Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake. In *Proceedings of the 78th ASIST Annual Meeting: Information Science with Impact: Research in and for the Community*, 2015.
- [74] Josh Constine. Facebook now has 2 billion monthly users and responsibility. *TechCrunch.com*, 27, 2017.
- [75] David M Cook. Birds of a feather deceive together: The chicanery of multiplied metadata. *Journal of Information Warfare*, 13(4):85–96, 2014.
- [76] David M Cook, Benjamin Waugh, Maldini Abdipanah, Omid Hashemi, and Shaquille Abdul Rahman. Twitter deception and influence: Issues of identity, slacktivism, and puppetry. *Journal of Information Warfare*, 13(1):58–71, 2014.
- [77] CoreFTP. CoreFTP, 2018. Available online: <http://www.coreftp.com/> (Accessed: 22 June 2018).
- [78] Economic Council and Social Research. What is social science?, 2018. Available online: <http://www.esrc.ac.uk/about-us/what-is-social-science/> (Accessed: 2 August 2018).
- [79] David Cournapeau. scikit-learn, 2017. Available online: <http://scikit-learn.org/stable/> (Accessed: 21 June 2018).

- [80] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Fame for sale: efficient detection of fake Twitter followers. *Decision Support Systems*, 80:56–71, 2015.
- [81] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. Social fingerprinting: detection of spambot groups through DNA-inspired behavioral modeling. *IEEE Transactions on Dependable and Secure Computing*, 2017.
- [82] Douglas Crockford. The application/json media type for javascript object notation (json). *RFC4627*, 2006.
- [83] Andrea Dal Pozzolo, Olivier Caelen, Serge Waterschoot, and Gianluca Bontempi. Racing for unbalanced methods selection. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 24–31. Springer, 2013.
- [84] Coral J Dando, Ray Bull, Thomas C Ormerod, and Alexandra L Sandham. Helping to sort the liars from the truth-tellers: The gradual revelation of information during investigative interviews. *Legal and Criminological Psychology*, 20(1):114–128, 2015.
- [85] Big Data. Principles and best practices of scalable realtime data systems. *N. Marz J. Warren. Henning*, 2014.
- [86] Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pages 233–240. ACM, 2006.
- [87] Krystal D’Costa. Catfishing: The truth about deception online. *Scientific American*, 2014.
- [88] Andrea De Mauro, Marco Greco, and Michele Grimaldi. A formal definition of big data based on its essential features. *Library Review*, 65(3):122–135, 2016.
- [89] James de Villiers. Suspects use fake Facebook profile to lure women, rape and kill them, 2017. Available online: <https://www.news24.com/SouthAfrica/News/suspects-use-fake-facebook-profile-to-lure-women-rape-and-kill-them-20171104> (Accessed: 4 November 2017).
- [90] Frans de Waal. Intentional deception in primates. *Evolutionary Anthropology: Issues, News, and Reviews*, 1(3):86–92, 1992.

- [91] Jeffrey Dean and Sanjay Ghemawat. MapReduce: A flexible data processing tool. *Communications of the ACM*, 53(1):72–77, 2010.
- [92] Lenka Dedkova. Stranger is not always danger: The myth and reality of meetings with online strangers. *Living in the digital age*, pages 78–94, 2015.
- [93] Bella M DePaulo, Kelly Charlton, Harris Cooper, James J Lindsay, and Laura Muhlenbruck. The accuracy-confidence correlation in the detection of deception. *Personality and Social Psychology Review*, 1(4):346–357, 1997.
- [94] Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979–995, 1996.
- [95] John P Dickerson, Vadim Kagan, and VS Subrahmanian. Using sentiment to detect bots on Twitter: Are humans more opinionated than bots? In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, pages 620–627. IEEE, 2014.
- [96] Francis X. Diebold. A personal perspective on the origin(s) and development of “big data”: The phenomenon, the term, and the discipline, 2012. Available online: <https://economics.sas.upenn.edu/sites/default/files/filevault/13-003.pdf> (Accessed: 31 July 2018).
- [97] Remco M Dijkman, Marlon Dumas, and Chun Ouyang. Semantics and analysis of business process models in bpmn. *Information and Software technology*, 50(12):1281–1294, 2008.
- [98] Chao Ding, Hsing Kenneth Cheng, Yang Duan, and Yong Jin. The power of the “like” button: The impact of social media on box office. *Decision Support Systems*, 94:77–84, 2017.
- [99] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.
- [100] Judith S Donath. Identity and deception in the virtual community. *Communities in cyberspace*, 1996:29–59, 1999.
- [101] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

- [102] Benedict Drasch, Johannes Huber, Sven Panz, and Florian Probst. Detecting online firestorms in social media. In *Proceedings of the 36th International Conference on Information Systems (ICIS)*, 2015.
- [103] Michelle Drouin, Daniel Miller, Shaun MJ Wehle, and Elisa Hernandez. Why do people lie online? "because everyone lies on the internet". *Computers in Human Behavior*, 64:134–142, 2016.
- [104] Mohammadreza Ebrahimi, Ching Y Suen, Olga Ormandjieva, and Adam Krzyzak. Recognizing predatory chat documents using semi-supervised anomaly detection. *Electronic Imaging*, 2016(17):1–9, 2016.
- [105] Allen L Edwards. *Experimental design in psychological research*. Harpercollins College Division, subsequent edition, 1950.
- [106] Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Compa: Detecting compromised accounts on social networks. In *The Network and Distributed System Security Symposium (NDSS)*, 2013.
- [107] Paul Ekman and Maureen O’sullivan. Who can catch a liar? *American psychologist*, 46(9):913–920, 1991.
- [108] Paul Ekman, Maureen O’Sullivan, and Mark G Frank. A few can catch a liar. *Psychological science*, 10(3):263–266, 1999.
- [109] Facebook. The Facebook graph API, 2017. Available online: <https://developers.facebook.com/docs/graph-api/overview> (Accessed: 8 January 2018).
- [110] Daniel Fallman. Design-oriented human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 225–232. ACM, 2003.
- [111] Tom Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
- [112] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems. *J. Mach. Learn. Res.*, 15(1):3133–3181, 2014.
- [113] Emilio Ferrara, Wen-Qiang Wang, Onur Varol, Alessandro Flammini, and Aram Galstyan. Predicting online extremism, content adopters, and interaction

- reciprocity. In *International Conference on Social Informatics*, pages 22–39. Springer, 2016.
- [114] Michael Fire, Roy Goldschmidt, and Yuval Elovici. Online social networks: threats and solutions. *IEEE Communications Surveys & Tutorials*, 16(4):2019–2036, 2014.
- [115] Michael Fire, Dima Kagan, Aviad Elyashar, and Yuval Elovici. Friend or foe? fake profile identification in online social networks. *Social Network Analysis and Mining*, 4(1):1–23, 2014.
- [116] Martin Fowler. *UML distilled: a brief guide to the standard object modeling language*. Addison-Wesley Professional, 2004.
- [117] Yoav Freund, Robert Schapire, and Naoki Abe. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(5):771–780, 1999.
- [118] Charles Fried. The new first amendment jurisprudence: A threat to liberty. *The University of Chicago Law Review*, 59(1):225–253, 1992.
- [119] Sanford Friedenthal, Alan Moore, and Rick Steiner. *A practical guide to SysML: the systems modeling language*. Morgan Kaufmann, 2014.
- [120] FSOC. The HPI Future SOC lab, 2018. Available online: <https://hpi.de/en/research/future-soc-lab.html> (Accessed: 8 June 2018).
- [121] Mélissa Gaillard. CERN data centre passes the 200-petabyte milestone, 2017. Available online: <https://phys.org/news/2017-07-cern-centre-petabyte-milestone.html> (Accessed: 4 March 2018).
- [122] Patxi Galán-García, José Gaviria De La Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. *Logic Journal of IGPL*, 24(1):42–53, 2016.
- [123] Roberto García, Heiko Paulheim, and Paola Di Maio. Special issue on semantic web interfaces. *Semantic Web*, 6(3):1–2, 2015.
- [124] Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 25(4):734–750, 2013.

- [125] Gartner. Gartner research, 2014. Available online: <https://www.gartner.com/it-glossary/big-data> (Accessed: 15 July 2018).
- [126] Wajeb Gharibi and Maha Shaabi. Cyber threats in social networking websites. *CoRR*, abs/1202.2420, 2012.
- [127] David R Gibson. Enduring illusions: The social organization of secrecy and deception. *Sociological Theory*, 32(4):283–306, 2014.
- [128] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. An in-depth characterisation of bots and humans on Twitter. *CoRR*, abs/1704.01508, 2017.
- [129] Global Deception Research Team. A world of lies. *Journal of cross-cultural psychology*, 2016.
- [130] Uri Gneezy. Deception: The role of consequences. *American Economic Review*, 95(1):384–394, 2005.
- [131] Ashish Goel, Aneesh Sharma, Dong Wang, and Zhijun Yin. Discovering similar users on Twitter. In *11th Workshop on Mining and Learning with Graphs*, Series Discovering Similar Users on Twitter, 2013.
- [132] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". In *ICML workshop on human interpretability in machine learning*, Series European Union regulations on algorithmic decision-making and a "right to explanation", 2016.
- [133] Google. Google trends, 2014. Available online: <http://www.google.com/trends/> (Accessed: 21 June 2014).
- [134] Google+. Google+ API, 2017. Available online: <https://developers.google.com/+/web/api/rest/> (Accessed: 8 January 2018).
- [135] Google. Google vision API, 2017. Available online: <https://cloud.google.com/vision/> (Accessed: 8 January 2018).
- [136] Google. Google scholar, 2018. Available online: <https://scholar.google.com> (Accessed: 6 January 2019).
- [137] Guofei Gu, Roberto Perdisci, Junjie Zhang, and Wenke Lee. Botminer: Clustering analysis of network traffic for protocol-and structure-independent botnet detection. In *USENIX security symposium*, volume 5, pages 139–154, 2008.

- [138] Pramod Kumar Gudipati. *Implementing a Lambda Architecture to perform real-time updates*. Thesis, Kansas State University, USA, 2016.
- [139] Pritam Gundecha and Huan Liu. Mining social media: a brief introduction. *Tutorials in Operations Research*, 1(4), 2012.
- [140] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. Faking Sandy: characterizing and identifying fake images on Twitter during hurricane Sandy. In *Proceedings of the 22nd international conference on World Wide Web*, pages 729–736. ACM, 2013.
- [141] Supraja Gurajala, Joshua S White, Brian Hudson, and Jeanna N Matthews. Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In *Proceedings of the International Conference on Social Media & Society*, pages 9–16. ACM, 2015.
- [142] Supraja Gurajala, Joshua S White, Brian Hudson, Brian R Voter, and Jeanna N Matthews. Profile characteristics of fake Twitter accounts. *Big Data & Society*, 3(2), 2016.
- [143] Michael Guta. Exploring video social network trends, 2016. Available online: <https://smallbiztrends.com/2016/03/video-social-networks.html> (Accessed: 4 March 2018).
- [144] Alon Halevy, Peter Norvig, and Fernando Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, 2009.
- [145] Rony Halevy, Shaul Shalvi, and Bruno Verschuere. Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, 40(1):54–72, 2014.
- [146] Richard W Hamming. Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160, 1950.
- [147] Jeffrey T Hancock. *Digital deception*. Oxford handbook of internet psychology. Oxford University Press, 2007.
- [148] Jeffrey T Hancock, Jennifer Thom-Santelli, and Thompson Ritchie. Deception and design: The impact of communication technology on lying behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 129–134. ACM, 2004.

- [149] Jeffrey T Hancock and Catalina L Toma. Putting your best face forward: The accuracy of online dating photographs. *Journal of Communication*, 59(2):367–386, 2009.
- [150] Simon Hansman and Ray Hunt. A taxonomy of network and computer attacks. *Computers & Security*, 24(1):31–43, 2005.
- [151] Md. Ansarul Haque and Tamjid Rahman. Sentiment analysis by using fuzzy logic. In *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, volume 4, 2014.
- [152] Ibrahim Abaker Targio Hashem, Ibrar Yaqoob, Nor Badrul Anuar, Salimah Mokhtar, Abdullah Gani, and Samee Ullah Khan. The rise of “big data” on cloud computing: review and open research issues. *Information Systems*, 47:98–115, 2015.
- [153] Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L Sporer. Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19(4):307–342, 2015.
- [154] Valerie Hauch, Jaume Masip, Iris Blandon-Gitlin, and Siegfried L Sporer. Linguistic cues to deception assessed by computer programs: A meta-analysis. In *Proceedings of the workshop on computational approaches to deception detection*, pages 1–4. Association for Computational Linguistics, 2012.
- [155] Pedram Hayati and Vidyasagar Potdar. Toward spam 2.0: an evaluation of web 2.0 anti-spam methods. In *7th IEEE International Conference on Industrial Informatics (INDIN)*, pages 875–880. IEEE, 2009.
- [156] Changlin He. Survey on nosql database technology. *Journal of Applied Science and Engineering Innovation Vol*, 2(2), 2015.
- [157] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.
- [158] Florian Heimerl, Steffen Lohmann, Simon Lange, and Thomas Ertl. Word cloud explorer: Text analytics based on word clouds. In *47th Hawaii International Conference on System Sciences (HICSS)*, pages 1833–1842. IEEE, 2014.
- [159] Lori Hil. These are the most engaging tweets of 2017. *Forbes*, 2017.

- [160] Shuyuan Mary Ho, Paul Benjamin Lowry, Merrill Warkentin, Yanyun Yang, and Jonathan Hollister. Lie to me: A multifactorial analysis of gender deception in asynchronous online communication. *Information Processing & Management*, 2016.
- [161] Eric Horvitz, David E Heckerman, Susan T Dumais, Mehran Sahami, and John C Platt. Technique which utilizes a probabilistic classifier to detect "junk" e-mail by automatically updating a training and re-training the classifier based on the updated training set. Patent, Google Patents, 2000.
- [162] John D Howard. An analysis of security incidents on the Internet 1989-1995. Report, Carnegie-Mellon Univ Pittsburgh PA, 1997.
- [163] IEEE. IEEE explore digital library, 2014. Available online: <http://ieeexplore.ieee.org/Xplore/guesthome.jsp> (Accessed: 21 June 2014).
- [164] Ross Ihaka and Robert Gentleman. R, 2017. Available online: <https://cran.r-project.org/> (Accessed: 31 July 2018).
- [165] Instagram. Instagram API, 2017. Available online: <https://www.instagram.com/developer/> (Accessed: 8 January 2018).
- [166] Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50, 1912.
- [167] Elin K Jacob. Classification and categorization: a difference that makes a difference. *In Library Trends*, 52(3):515–540, 2004.
- [168] Lori Jahnke and Andrew Asher. The problem of data. Report, Council of library and information services, 2012.
- [169] Antonio J. Jara, Yann Bocchi, and Dominique Genoud. Determining human dynamics through the internet of things. In *IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, pages 109–113, 2013.
- [170] László A Jeni, Jeffrey F Cohn, and Fernando De La Torre. Facing imbalanced data—recommendations for the use of performance metrics. In *Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 245–251. IEEE, 2013.

- [171] Xiaolong Jin, Benjamin W Wah, Xueqi Cheng, and Yuanzhuo Wang. Significance and challenges of big data research. *Big Data Research*, 2015.
- [172] Xin Jin, Chi Wang, Jiebo Luo, Xiao Yu, and Jiawei Han. Likeminer: a system for mining the power of "like" in social media networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 753–756. ACM, 2011.
- [173] Kedar R Joshi and Sonali S Patil. Processing big data applications using mapreduce. *Journal of Data Mining and Management*, 1(1, 2, 3), 2017.
- [174] Louise Marie Jupe, Aldert Vrij, Galit Nahari, Sharon Leal, and Samantha Ann Mann. The lies we live: Using the verifiability approach to detect lying about occupation. *Journal of Articles in Support of the Null Hypothesis*, 13(1):1–13, 2016.
- [175] ME Kabay, Eric Salveggio, Robert Guess, and Russell D Rosco. Anonymity and identity in cyberspace. *Computer Security Handbook, Sixth Edition*, pages 70.1–70.37, 2014.
- [176] D. Kahle and H. Wickham. ggmap: Spatial visualization with ggplot2, 2018. Available online: <https://cran.r-project.org/web/packages/ggmap/index.html> (Accessed: 6 May 2018).
- [177] Deborah A Kashy and Bella M DePaulo. Who lies? *Journal of Personality and Social Psychology*, 70(5):1037–1051, 1996.
- [178] Benjamin Keen. Generate data, 2017. Available online: <http://www.generatedata.com/> (Accessed: 8 January 2018).
- [179] Leo Kelion. Facebook: Data haul had private messages. *BBC News*, 10 Apr 2018 2018.
- [180] Simon Kemp. Digital in 2018: world's internet users pass the 4 billion mark. *New York, We Are Social*, 30, 2018.
- [181] Mudassir Khan. Big data analytics evaluation. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, 5(2):25–28, 2018.
- [182] Sylvia Kierkegaard. Cybering, online grooming and ageplay. *Computer Law & Security Review*, 24(1):41–55, 2008.

- [183] Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of Twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI Workshop on Microblogging*, 2010.
- [184] Mariam Kiran, Peter Murphy, Inder Monga, Jon Dugan, and Sartaj Singh Baveja. Lambda architecture for cost-effective batch and speed big data processing. In *IEEE International Conference on Big Data (Big Data)*, pages 2785–2792. IEEE, 2015.
- [185] Lyudmyla Kirichenko, Tamara Radivilova, and Anders Carlsson. Detecting cyber threats through social network analysis: short survey. *SocioEconomic Challenges*, 1(1):20–34, 2017.
- [186] Jytte Klausen. Tweeting the jihad: Social media networks of western foreign fighters in Syria and Iraq. *Studies in Conflict & Terrorism*, 38(1):1–22, 2015.
- [187] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conferences on Artificial Intelligence (IJCAI)*, volume 14, pages 1137–1145. Stanford, CA, 1995.
- [188] John Kopanakis. Best times to post on social media in 2018, 2018. Available online: <https://www.mentionlytics.com/blog/best-times-to-post-on-social-media-2018/> (Accessed: 2 June 2018).
- [189] Chakravanti Rajagopalachari Kothari. *Research methodology: Methods and techniques*. New Age International, 2004.
- [190] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.
- [191] M. Kuhn, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, and Z. Mayer. caret: Classification and regression training, 2016. Available online: <https://cran.r-project.org/package=caret> (Accessed: 15 July 2018).
- [192] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.
- [193] Alexandros Labrinidis and H. V. Jagadish. Challenges and opportunities with big data. In *Proceedings of the VLDB Endowment*, volume 5, page 2. VLDB Endowment, 2012.

- [194] Avinash Lakshman and Prashant Malik. Apache Cassandra, 2018. Available online: <http://cassandra.apache.org/> (Accessed: 4 Mar 2018).
- [195] Doug Laney. 3d data management: Controlling data volume, velocity and variety. *META Group Research Note*, 6(70), 2001.
- [196] Sharon Leal, Aldert Vrij, Galit Nahari, and Samantha Mann. Please be honest and provide evidence: deterrents of deception in an online insurance fraud context. *Applied Cognitive Psychology*, 30(5):768–774, 2016.
- [197] Kyumin Lee, James Caverlee, and Steve Webb. The social honeypot project: protecting online communities from spammers. In *Proceedings of the 19th international conference on World wide web*, pages 1139–1140. ACM, 2010.
- [198] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710, 1966.
- [199] Emma E Levine and Maurice E Schweitzer. Prosocial lies: When deception breeds trust. *Organizational Behavior and Human Decision Processes*, 126:88–106, 2015.
- [200] Timothy R Levine, Mohamed Vaqas Ali, Marleah Dean, Rasha A Abdulla, and Karina Garcia-Ruano. Toward a pan-cultural typology of deception motives. *Journal of Intercultural Communication Research*, 45(1):1–12, 2016.
- [201] Jiexun Li and Alan G Wang. A framework of identity resolution: evaluating identity attributes and matching algorithms. *Security Informatics*, 4(1):1, 2015.
- [202] Jundong Li and Huan Liu. Challenges of feature selection for big data analytics. *IEEE Intelligent Systems*, 32(2):9–15, 2017.
- [203] Xiaoshan Li, Zhiming Liu, Jifeng He, and Quan Long. Generating a prototype from a UML model of system requirements. In *International Conference on Distributed Computing and Internet Technology*, pages 255–265. Springer, 2004.
- [204] Yixuan Li, Oscar Martinez, Xing Chen, Yi Li, and John E Hopcroft. In a world that counts: Clustering and detecting fake social engagement at scale. In *Proceedings of the 25th International Conference on World Wide Web*, pages 111–120. International World Wide Web Conferences Steering Committee, 2016.
- [205] Ilaria Liccardi, Monica Bulger, Hal Abelson, Daniel J Weitzner, and Wendy Mackay. Can apps play by the COPPA rules? In *Twelfth Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–9. IEEE, 2014.

- [206] LinkedIn. LinkedIn developers, 2017. Available online: <https://developer.linkedin.com/> (Accessed: 8 January 2018).
- [207] Huan Liu, Jiawei Han, and Hiroshi Motoda. Uncovering deception in social media. *Social Network Analysis and Mining*, 4(1):1–2, 2014.
- [208] Tetyana Lokot and Nicholas Diakopoulos. News bots: Automating news and information dissemination on Twitter. *Digital Journalism*, 4(6):682–699, 2016.
- [209] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.
- [210] Chuang Ma, Hao Helen Zhang, and Xiangfeng Wang. Machine learning for big data analytics in plants. *Trends in plant science*, 19(12):798–808, 2014.
- [211] Feng Mai, Qing Bai, Zhe Shan, Xin Shane Wang, and Roger Chiang. The impacts of social media on Bitcoin performance. *Journal of Management Information Systems*, 35(1):19–52, 2016.
- [212] Gōrg Mallia. A concise timeline of printing milestones, 2000. Available online: (Accessed: 31 May 2014).
- [213] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
- [214] Bernard Marr. Big data-driven decision-making at domino’s pizza. *Forbes*, 2016.
- [215] Bernard Marr. The amazing ways coca cola uses artificial intelligence and big data to drive success. *Forbes*, 2017.
- [216] Jim McCambridge, Kypros Kypri, Preben Bendtsen, and John Porter. The use of deception in public health behavioral intervention trials: a case study of three online alcohol trials. *The American journal of bioethics*, 13(11):39–47, 2013.
- [217] Lori McCay-Peet and Anabel Quan-Haase. What is social media and what questions can social media research help us answer? *The SAGE Handbook of Social Media Research Methods*, page 13, 2017.
- [218] John H McDonald. *Handbook of biological statistics*, volume 2. Sparky House Publishing Baltimore, MD, 2009.

- [219] Ali M Meligy, Hani M Ibrahim, and Mohamed F Torky. Identity verification mechanism for detecting fake profiles in online social networks. *International Journal Computer Network and Information Security*, 1:31–39, 2017.
- [220] Giovanna Menardi and Nicola Torelli. Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, 28(1):1–31, 2014.
- [221] Jorg Michael. Names dictionary, 2008. Available online: <https://www.heise.de/ct/ftp/07/17/182/> (Accessed: 30 October 2017).
- [222] Microsoft Corporation. Windows Azure, 2014. Available online: <https://azure.microsoft.com/en-us/> (Accessed: 8 August 2014).
- [223] Microsoft Corporation. Microsoft Excel, 2016. Available online: <https://products.office.com/en-gb/excel> (Accessed: 21 June 2018).
- [224] Microsoft Corporation. Microsoft Power BI, 2017. Available online: <https://powerbi.microsoft.com/en-us/> (Accessed: 31 May 2018).
- [225] Microsoft Corporation. Windows, 2017. Available online: <https://www.microsoft.com/en-us/windows> (Accessed: 19 June 2018).
- [226] Peter Middleton. Forecast analysis: Internet of things — endpoints, worldwide, update, 2017. Available online: <https://www.gartner.com/doc/3841268/forecast-analysis-internet-things-> (Accessed: 15 July 2018).
- [227] Sean Miller and Curtis Busby-Earle. The impact of different botnet flow feature subsets on prediction accuracy using supervised and unsupervised learning methods. *Journal of Internet Technology and Secured Transactions (JITST)*, 5(12):474–485, 2016.
- [228] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. Is the sample good enough? comparing data from Twitter’s streaming API with Twitter’s Firehose. In *The International AAAI Conference on Web and Social Media (ICWSM)*, 2013.
- [229] David Mortenson. Year in review: Data centers, 2018. Available online: <https://code.facebook.com/posts/392743124493876/2017-year-in-review-data-centers/> (Accessed: 4 March 2018).
- [230] Rebecca B Morton and Kenneth C Williams. *Experimental political science and the study of causality: From nature to the lab*. Cambridge University Press, 2010.

- [231] Frank Muller and Frank Thiesing. Social networking APIs for companies—an example of using the Facebook API for companies. In *International Conference on Computational Aspects of Social Networks (CASoN)*, pages 120–123. IEEE, 2011.
- [232] Shimon Y Nof. *Handbook of industrial robotics*, volume 1. John Wiley & Sons, 1999.
- [233] Richard J Oentaryo, Arinto Murdopo, Philips K Prasetyo, and Ee-Peng Lim. On profiling bots in social media. In *International Conference on Social Informatics*, pages 92–109. Springer, 2016.
- [234] OpenStreetMap Foundation. OpenStreetMap API, 2018. Available online: https://wiki.openstreetmap.org/wiki/Main_Page (Accessed: 31 May 2018).
- [235] Gerard O’Regan. *Unified Modelling Language*, pages 225–238. Springer, 2017.
- [236] Oxford. The English Oxford Dictionary, 2012. Available online: <http://www.oxforddictionaries.com/> (Accessed: 21 Jun 2018).
- [237] Oxford. Oxford university - academic departments, 2018. Available online: <http://www.ox.ac.uk/> (Accessed: 5 May 2018).
- [238] Natalie Parde and Rodney Nielsen. Detecting sarcasm is extremely easy. In *Proceedings of the Workshop on Computational Semantics beyond Events and Roles*, pages 21–26, 2018.
- [239] Hee Sun Park and Timothy R Levine. Base rates, deception detection, and deception theory: A reply to Burgoon (2015). *Human Communication Research*, 41(3):350–366, 2015.
- [240] Radhika Patel, Rajvi Bhagat, Palaumi Modi, and Harshil Joshi. Privacy and security issues in social online networks. In *National Conference on Latest Trends in Networking and Cyber Security (IJIRST)*, pages 130–134, 2017.
- [241] Tina R Patil and SS Sherekar. Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, 6(2):256–261, 2013.
- [242] Sai Teja Peddinti, Keith W Ross, and Justin Cappos. Mining anonymity: Identifying sensitive accounts on Twitter. In *International AAAI Conference on*

- Web and Social Media*, Series Mining Anonymity: Identifying Sensitive Accounts on Twitter, 2017.
- [243] Jaclyn Peiser. A bot that makes Trump's tweets presidential. *The New York Times*, 15 Oct 2017 2017.
- [244] Sarah Perez. Top 8 web 2.0 security threats. *Read Write Enterprise*. Retrieved March, 9, 2011.
- [245] Petra Perner. Big data analysis and reporting with decision tree induction. In *Advances in Information Science and Computer Engineering*, pages 25–34, 2015.
- [246] Jillian Peterson and James Densley. Cyber violence: What do we know and where do we go from here? *Aggression and violent behavior*, 34:193–200, 2017.
- [247] Thu Zar Phyu and Nyein Nyein Oo. Performance comparison of feature selection methods. In *MATEC Web of Conferences*, volume 42. EDP Sciences, 2016.
- [248] Pinterest. Pinterest API, 2017. Available online: <https://developers.pinterest.com/> (Accessed: 8 January 2018).
- [249] PostgreSQL Global Development Group. PostgreSQL: The world's most advanced open source database, 2018. Available online: <https://www.postgresql.org/> (Accessed: 12 March 2018).
- [250] Devika Prabhu. *Application of web 2.0 and web 3.0: an overview*. Lap Lambert Academic Publishing, 2017.
- [251] Pragya Pradhan, Santanu Misra, and Tawal Koirala. A survey on data security in social networking sites. *International Journal of Computer Applications*, 155(7), 2016.
- [252] Tom Preston-Werner, Chris Wanstrath, and PJ Hyett. GitHub, 2017. Available online: <https://github.com/> (Accessed: 31 May 2018).
- [253] Eftychios Protopapadakis. *Decision making via semi-supervised machine learning techniques*. Thesis, Technical University of Crete, 2016.
- [254] Nicole M Radziwill and Morgan C Benton. Bot or not? deciphering time maps for tweet interarrivals. *CoRR*, abs/1605.06555, 2016.

- [255] V Ramesh. An efficient survey of managing big data analytics using swarm intelligence. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 2(4):933–936, 2017.
- [256] Bahman Rashidi and Carol Fung. Bottracer: Bot user detection using clustering method in recdroid. In *Network Operations and Management Symposium (NOMS), IEEE/IFIP*, pages 1239–1244. IEEE, 2016.
- [257] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM, 2016.
- [258] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.
- [259] C Carl Robusto. The Cosine-Haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.
- [260] Rong Rong, Daniel Houser, and Anovia Yifan Dai. Money or friends: Social identity and deception in networks. *European Economic Review*, 90:56–66, 2016.
- [261] Victoria L Rubin. On deception and deception detection: Content analysis of computer-mediated stated beliefs. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–10, 2010.
- [262] Victoria L Rubin. *Deception Detection and Rumor Debunking for Social Media*, book section 21. SAGE, UK, 2017.
- [263] Ando Saabas. Package for interpreting scikit-learn’s decision tree and random forest predictions., 2018. Available online: <https://pypi.org/project/treeinterpreter/> (Accessed: 7 April 2018).
- [264] Takaya Saito and Marc Rehmsmeier. Precrec: fast and accurate precision–recall and ROC curve calculations in R. *Bioinformatics*, 33(1):145–147, 2017.
- [265] Christopher Sandy, Patrice Rusconi, and Shujun Li. Can humans detect the authenticity of social media accounts? In *3rd IEEE International Conference on Cybernetics (CYBCONF)*, Series Can Humans Detect the Authenticity of Social Media Accounts?, 2017.

- [266] SAP. SAP HANA, 2017. Available online: <https://www.sap.com/uk/developer/topics/sap-hana.html> (Accessed: 31 August 2018).
- [267] Ankur Saxena, Shivani Singh, and Chetna Shakya. *Concepts of HBase Archetypes in Big Data Engineering*, pages 83–111. Springer, 2018.
- [268] Anja Schulz, Emilia Bergen, Petya Schuhmann, Jürgen Hoyer, and Pekka Santtila. Online sexual solicitation of minors how often and between whom does it occur? *Journal of Research in Crime and Delinquency*, page 0022427815599426, 2015.
- [269] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.
- [270] Graham G Scott, Elizabeth A Boyle, Kamila Czerniawska, and Ashleigh Courtney. Posting photos on Facebook: The impact of narcissism, social anxiety, loneliness, and shyness. *Personality and Individual Differences*, 2017.
- [271] Surendra Sedhai and Aixin Sun. Semi-supervised spam detection in Twitter stream. *IEEE Transactions on Computational Social Systems*, 5(1):169 – 175, 2017.
- [272] Saeed Shahrivari and Saeed Jalili. Beyond batch processing: Towards real-time and streaming big data. *Computers*, 3:117–129, 2014.
- [273] Claude E Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.
- [274] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Alessandro Flammini, and Filippo Menczer. The spread of fake news by social bots. *arXiv preprint arXiv:1707.07592*, 2017.
- [275] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.
- [276] D Shenk. *Data Smog - Surviving the information glut*. HarperCollins, 1997.
- [277] BP Shumaker and RW Sinnott. Astronomical computing: Computing under the open sky and virtues of the Haversine. *Sky and telescope*, 68:158–159, 1984.

- [278] Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler. The Hadoop distributed file system. In *IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST)*, pages 1–10. IEEE, 2010.
- [279] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, and Vishanth Weerakkody. Critical analysis of big data challenges and analytical methods. *Journal of Business Research*, 70:263–286, 2017.
- [280] A Skendzic and B Kovacic. Open source system openvpn in a function of virtual private network. In *IOP Conference Series: Materials Science and Engineering*, volume 200, page 012065. IOP Publishing, 2017.
- [281] Luke Sloan, Jeffrey Morgan, Pete Burnap, and Matthew Williams. Who tweets? deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. *PloS one*, 10(3):e0115545, 2015.
- [282] Oliver Smith. Mapped: The world according to Internet connection speeds. *The Telegraph*, 7 Apr 2017 2017.
- [283] Saini Jacob Soman and S Murugappan. Detecting malicious tweets in trending topics using clustering and classification. In *International Conference on Recent Trends in Information Technology (ICRTIT)*, pages 1–6. IEEE, 2014.
- [284] Kasey Stanton, Stephanie Ellickson-Larew, and David Watson. Development and validation of a measure of online deception and intimacy. *Personality and Individual Differences*, 88:187–196, 2016.
- [285] Margaret C. Stewart and B. Gail Wilson. The dynamic role of social media during hurricane Sandy. *Computers in Human Behavior*, 54:639–646, 2016.
- [286] Sirko Straube and Mario M Krell. How to evaluate an agent’s behavior to infrequent events?—reliable performance estimation insensitive to class distribution. *Frontiers in computational neuroscience*, 8(43):1–6, 2014.
- [287] Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna. Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference*, pages 1–9. ACM, 2010.
- [288] Ying Tan, Yuhui Shi, and Ben Niu. *Advances in swarm intelligence*. Springer, 2016.

- [289] Andrew S Tanenbaum and Albert S Woodhull. *Operating systems: design and implementation*, volume 2. Prentice-Hall Englewood Cliffs, NJ, 1987.
- [290] Rongrong Tao, Baojian Zhou, Adil Alim, Feng Chen, David Mares, Patrick Butler, and Naren Ramakrishnan. Social media as a sensor for censorship detection in news media. *CoRR*, abs/1611.06947 (2016), 2016.
- [291] Simon Tatham. Putty, 2017. Available online: <http://www.putty.org/> (Accessed: 1 August 2018).
- [292] John R Taylor. *Linguistic categorization*. Oxford University Press, 2003.
- [293] Dafna Tener, Janis Wolak, and David Finkelhor. A typology of offenders who use online communications to commit sex crimes against minors. *Journal of Aggression, Maltreatment & Trauma*, 24(3):319–337, 2015.
- [294] Sin G. Teo, Shuguo Han, and Vincent C. S. Lee. Privacy preserving support vector machine using non-linear kernels on Hadoop Mahout. In *IEEE 16th International Conference on Computational Science and Engineering*, Series Privacy Preserving Support Vector Machine Using Non-linear Kernels on Hadoop Mahout, pages 941–948. IEEE, 2013.
- [295] Gillian Tett. Trump, Cambridge Analytica and how big data is reshaping politics. *Financial Times*, 2017.
- [296] Terry M Therneau and Elizabeth J Atkinson. An introduction to recursive partitioning using the rpart routines. Report, Technical report Mayo Foundation, 1997.
- [297] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. Suspended accounts in retrospect: an analysis of Twitter spam. In *Proceedings of the ACM SIGCOMM conference on Internet measurement conference*, pages 243–258. ACM, 2011.
- [298] Kurt Thomas, Damon McCoy, Chris Grier, Alek Kolcz, and Vern Paxson. Trafficking fraudulent accounts: The role of the underground market in Twitter spam and abuse. In *USENIX Security*, volume 13, pages 195–210. Citeseer, 2013.
- [299] T. J. Thomson and Keith Greenwood. I “like” that: Exploring the characteristics that promote social media engagement with news photographs. *Visual Communication Quarterly*, 24(4):203–218, 2017.
- [300] A Toffler. *Future Shock*. Random House, 1970.

- [301] Catalina L Toma, Jeffrey T Hancock, and Nicole B Ellison. Separating fact from fiction: An examination of deceptive self-presentation in online dating profiles. *Personality and Social Psychology Bulletin*, 34(8):1023–1036, 2008.
- [302] Martin Traverso. Presto: Interacting with petabytes of data at Facebook, 2013. Available online: <https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920/?s=keen-io> (Accessed: 4 March 2018).
- [303] Seema D Trivedi, Chandani Kathad, Tosal Bhalodiya, and Tanvi Pandya. Analytical study of cyber threats in social networking. In *International Conference on Computer Science Networks and Information Technology*, 2016.
- [304] Michail Tsikerdekis. Identity deception prevention using common contribution network data. *IEEE Transactions on Information Forensics and Security*, 12(1):188–199, 2017.
- [305] Michail Tsikerdekis and Sherali Zeadally. Multiple account identity deception detection in social media using nonverbal behavior. *Information Forensics and Security, IEEE Transactions on*, 9(8):1311–1321, 2014.
- [306] Michail Tsikerdekis and Sherali Zeadally. Detecting and preventing online identity deception in social networking services. *Internet Computing, IEEE*, 19(3):41–49, 2015.
- [307] Tayfun Tuna, Esra Akbas, Ahmet Aksoy, Muhammed Abdullah Canbaz, Umit Karabiyik, Bilal Gonen, and Ramazan Aygun. User characterization for online social networks. *Social Network Analysis and Mining*, 6(1):104–131, 2016.
- [308] Doug Turnbull. Solving data “variety” with Postgres’s NoSQL extensions, 2014. Available online: <http://opensourceconnections.com/blog/2014/09/26/solving-data-variety-with-postgress-nosql-extensions/> (Accessed: 27 October 2014).
- [309] Simranjit Kaur Tuteja. A survey on classification algorithms for email spam filtering. *International Journal of Engineering Science*, 5937, 2016.
- [310] Twitter. Twitter API, 2018. Available online: <https://dev.twitter.com/overview/api> (Accessed: 30 March 2018).
- [311] Sonja Utz. Types of deception and underlying motivation: What people think. *Social Science Computer Review*, 23(1):49–56, 2005.

- [312] E. Van der Walt and J.H.P. Eloff. Using machine learning to detect fake identities - bots versus humans. *IEEE Access*, 6:6540 – 6549, 2018.
- [313] Estee van der Walt and J. H. P. Eloff. Creating an environment for detecting identity deception. In *5th HPI Symposium on Operating the Cloud*, Series Creating an environment for detecting Identity Deception, 2017.
- [314] Estee van der Walt, JHP Eloff, and Jacomine Grobler. Cyber-security: Identity deception detection on social media platforms. *Computers & Security*, 2018.
- [315] José Van Dijck. *The culture of connectivity - A critical history of social media*. Oxford University Press, 2013.
- [316] Diederik Van Liere. How far does a tweet travel?: Information brokers in the twitterverse. In *Proceedings of the International Workshop on Modeling Social Media*, page 6. ACM, 2010.
- [317] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh International AAAI Conference on Web and Social Media*, pages 280–289, 2017.
- [318] Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *European Symposium on Artificial Neural Networks (ESANN)*, volume 12, pages 163–172, 2012.
- [319] Sridhar Venkatesan, Massimiliano Albanese, Ankit Shah, Rajesh Ganesan, and Sushil Jajodia. Detecting stealthy botnets in a resource-constrained environment using reinforcement learning. In *Workshop on Moving Target Defense*, pages 75–85. ACM, 2017.
- [320] Bimal Viswanath, Muhammad Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. Towards detecting anomalous user behavior in online social networks. In *Usenix Security*, volume 14, 2014.
- [321] Matthias Volk, Stefan Hart, Sascha Bosse, and Klaus Turowski. How much is big data? A classification framework for IT projects and technologies. In *Twenty-second Americas Conference on Information Systems*, Series How much is Big Data? A Classification Framework for IT Projects and Technologies, 2016.

- [322] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.
- [323] Aldert Vrij. *Guidelines to catch a liar*. The Detection of Deception in Forensic Contexts. Cambridge University Press, 2004.
- [324] G Alan Wang, Hsinchun Chen, Jennifer J Xu, and Homa Atabakhsh. Automatically detecting criminal identity deception: An adaptive detection algorithm. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, 36(5):988–999, 2006.
- [325] Gang Wang, Hsinchun Chen, and Homa Atabakhsh. Criminal identity deception and deception detection in law enforcement. *Group Decision and Negotiation*, 13(2):111–127, 2004.
- [326] WenJie Wang, Yufei Yuan, and Norm Archer. A contextual framework for combating identity theft. *IEEE Security & Privacy*, 4(2):30–38, 2006.
- [327] Michelle Wetzler. Architecture of giants: Data stacks at Facebook, Netflix, Airbnb, and Pinterest. *The Event Log*, 2017.
- [328] Hadley Wickham. ggplot2, 2017. Available online: <http://ggplot2.org/> (Accessed: 30 June 2018).
- [329] Wikipedia. Big data, 2014. Available online: http://en.wikipedia.org/wiki/Big_data (Accessed: 21 June 2014).
- [330] Wikipedia. IEEE explore, 2014. Available online: http://en.wikipedia.org/wiki/IEEE_Explore (Accessed: 21 June 2014).
- [331] Wikipedia. Prototype, 2017. Available online: <https://en.wikipedia.org/wiki/Prototype> (Accessed: 30 April 2018).
- [332] Nancy E Willard. *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press, 2007.
- [333] Shirley A Williams, Melissa M Terras, and Claire Warwick. What do people study when they study Twitter? classifying Twitter related academic papers. *Journal of Documentation*, 69(3):384–410, 2013.
- [334] Michael Wood. *In search of the Trojan War*. Univ of California Press, 1998.

- [335] Wei Wu, Jaime Alvarez, Chengcheng Liu, and Hung-Min Sun. Bot detection using unsupervised machine learning. *Microsystem Technologies*, pages 1–9, 2016.
- [336] Xindong Wu, Vipin Kumar, J Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J McLachlan, Angus Ng, Bing Liu, and S Yu Philip. Top 10 algorithms in data mining. *Knowledge and information systems*, 14(1):1–37, 2008.
- [337] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. Data mining with big data. *Knowledge and Data Engineering, IEEE Transactions on*, 26(1):97–107, 2014.
- [338] Yingcai Wu, Nan Cao, David Gotz, Yap-Peng Tan, and Daniel A Keim. A survey on visual analytics of social media data. *IEEE Transactions on Multimedia*, 18(11):2135–2148, 2016.
- [339] Peter Xenopoulos. Introducing deepbalance: Random deep belief network ensembles to address class imbalance. In *IEEE International Conference on Big Data (Big Data)*, Series Introducing DeepBalance: Random Deep Belief Network Ensembles to Address Class Imbalance. IEEE, 2017.
- [340] Cao Xiao, David Mandell Freeman, and Theodore Hwa. Detecting clusters of fake accounts in online social networks. In *The 8th ACM Workshop on Artificial Intelligence and Security*, pages 91–101. ACM, 2015.
- [341] Gaogang Xie, Zhenyu Li, Mohamed Ali Kaafar, and Qinghua Wu. Access types effect on internet video services and its implications on CDN caching. *IEEE Transactions on Circuits and Systems for Video Technology*, 2017.
- [342] Mosa Yahyazadeh and Mahdi Abadi. Botonus: An online unsupervised method for botnet detection. *The ISC International Journal of Information Security*, 4(1):51–62, 2012.
- [343] Yale. Yale university - Academic departments, 2018. Available online: <https://www.yale.edu/academics/departments-programs> (Accessed: 5 May 2018).
- [344] Chao Yang, Robert Harkreader, and Guofei Gu. Empirical evaluation and new design for fighting evolving Twitter spammers. *IEEE Transactions on Information Forensics and Security*, 8(8):1280–1293, 2013.
- [345] Tal Yarkoni and Jacob Westfall. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(6):1100–1122, 2016.

- [346] Juan Ye, Stamatia Dasiopoulou, Graeme Stevenson, Georgios Meditskos, Efstratios Kontopoulos, Ioannis Kompatsiaris, and Simon Dobson. Semantic web technologies in pervasive computing: A survey and research roadmap. *Pervasive and Mobile Computing*, 2015.
- [347] James Yonan. OpenVPN, 2017. Available online: <https://openvpn.net/> (Accessed: 30 June 2018).
- [348] YouTube. YouTube API, 2017. Available online: <https://developers.google.com/youtube/> (Accessed: 4 March 2018).
- [349] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning*, pages 856–863, 2003.
- [350] Reza Zafarani and Huan Liu. Evaluation without ground truth in social media research. *Communications of the ACM*, 58(6):54–60, 2015.
- [351] Ke Zeng, Xiao Wang, Qingpeng Zhang, Xinzhan Zhang, and Fei-Yue Wang. Behavior modeling of internet water army in online forums. *World Congr*, 19:9858–9863, 2014.
- [352] Zhiyong Zhang and Brij B Gupta. Social media security and trustworthiness: overview and new direction. *Future Generation Computer Systems*, 2016.
- [353] Lina Zhou, Judee K Burgoon, Douglas P Twitchell, Tiantian Qin, and Jay F Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20(4):139–166, 2004.
- [354] Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2(3):4–63, 2005.

Appendix A

Glossary of terms and definitions

The following terms that were used in this thesis are briefly defined to avoid potential misinterpretation:

Attribute vs Feature

According to the Oxford Dictionary [236], an ‘attribute’ is defined as “a piece of information which determines the properties of a field or tag in a database or a string of characters in a display”. ‘Attributes’ describe an object and usually consist of a key value pair where the name is the name given to the attribute and the value describes that attribute. Examples are the name of the SMP account holder and their profile image. A ‘feature’, on the other hand, is defined as “a distinctive attribute or aspect of something” [236] and could be one attribute, the combination of many attributes, or the creation of a new attribute from existing attributes or other information. For the purpose of this thesis, the word ‘attribute’ will be used to describe information as found directly within the SMP about a human or their account. ‘Feature’ will be used to describe any engineered information from the SMP attributes. An example of a feature is where the gender is derived from the profile image.

Machine learning algorithm vs machine learning model

Machine learning is “the capacity of a computer to learn from experience” [236]. A machine learning algorithm consists of a formula with various input parameters that aim to predict some output [99]. Given the input parameters and data, the machine learning algorithm will determine a model that best describes the data. A machine learning model is thus the result of a machine learning algorithm combined with example training data [99]. The machine learning model can be applied to describe new data. A machine learning algorithm is more general, whereas a machine learning model will

predict some output, given its input and prior training.

Entropy vs importance vs information gain

In this thesis, the term ‘Shannon entropy’ is used to describe how much information is available in an attribute [273]. This is synonymous with ‘information gain’ [336]. Entropy, importance, and information gain are used interchangeably throughout the thesis and refer to the same concept. For example, the experimental machine learning results could show that by including the SMP account holder’s geographic location during the development of the model, the machine learning model’s predictive success has improved. Or, in other words, by including the SMP account holder’s geographic location, more information was gained to determine whether the person was potentially deceptive or not.

Cyber security

Cyber security is defined in a popular book by Cheswick and Bollovin [67] as a means to keep someone from doing things you do not want them to do on any electronic device. This definition is very vague as it refers to ‘someone’ doing ‘some things’ on ‘any’ device. For the purposes of this thesis, cyber security will specifically refer to the protection of humans who use SMPs against other malicious humans who present themselves with an identity different from the truth. The latter threat is also known as identity deception [306].

Impersonation vs identity deception vs masquerading vs social engineering vs fake accounts

Impersonation is the act of pretending to be another [236], while identity deception is when the identity is not representative of the truth [304]. Identity deception is a more focused form of impersonation [326]. Both impersonation and identity deception can be seen as masquerading, where masquerading means to pretend [236]. Masquerading is not specific to identities found on SMPs, whereas social engineering is a form of masquerading specific to creating fake identities [81]. For the purposes of this research, these terms will be used interchangeably to refer to fake accounts created on SMPs.

SMP account holder vs SMP user

An SMP account is opened by a user. The words SMP account holder and SMP user will be used interchangeably in this research as they are one and the same.

Appendix B

Acronyms

The following acronyms were used in this thesis. The acronyms are listed alphabetically with the meaning and page locations alongside.

API	Application Program Interface 39, 61, 69, 70, 73–75, 78, 93, 94, 110, 111, 122, 124, 137, 138, 140, 204
AUC	Area Under Curve 47, 84, 159, 164, 170
AWS	Amazon Web Services 18, 19, 21
BC	Before Christ 15
BPMN	Business Process Modelling Notation 93
CDA	Communications of Decency Act 39
CDN	Content Delivery Network 21
CIPA	Children Internet Protection Act 39
COPA	Child Online Protection Act 39
COPPA	Children Online Privacy Protection Act 39
CPU	Central Processing Unit 104–106
DOPA	Deleting Online Predators Act 39
EDA	Exploratory Data Analysis 123, 124, 127, 132, 134
ERD	Entity Relationship Diagrams 93
FN	False Negative 83
FP	False Positive 83
GDPR	General Data Protection Regulation 39, 181
HDFS	Hadoop Distributed File System 110
HPI	Hasso Plattner Institute 104
ID	Identifier 25, 71, 134

IDDM	Identity Deception Detection Model 8, 11, 12, 85, 106, 179, 180, 184, 192, 194, 198, 202, 203, 206
IDDMLM	Identity Deception Detection Machine Learning Model 85, 86, 99, 180, 184–186, 190, 191, 194, 198, 200, 202, 203
IDS	Identity Deception Score 205
IDSM	Identity Deception Detection Score Model 85, 86, 99, 180, 184–186, 190, 191, 193, 194, 198, 202, 203, 206
IDT	Interpersonal Deception Theory 36, 48, 89
IEEE	Institute of Electrical and Electronic Engineers 29, 30, 32
IET	Institute of Engineering and Technology 29, 30
IOT	Internet of Things 15
JSON	JavaScript Object Notation 21, 22, 110
LAN	Local Area Network 109
LHC	Large Hydron Collider 18
NLP	Natural Language Processing 199
NoSQL	Not only SQL 114
OED	Oxford English Dictionary 36
PAL	Predictive Analytics Library 107
PR	Precision Recall 84
REST	Representational State Transfer 75
ROC	Receiver Operator Characteristic 81, 84
ROSE	Random Over-Sampling Examples 80
SMOTE	Synthetic Minority Oversample Technique 80, 81, 97
SMP	Social Media Platform 2–12, 14, 15, 19–22, 24, 26–57, 59, 61–64, 66–70, 74, 77, 78, 80, 85–95, 97, 99–102, 104–111, 113, 114, 117, 119, 121–124, 127, 130–132, 137, 138, 141, 144, 145, 149–158, 162–164, 166, 169, 173, 175, 177–180, 183–186, 191, 193–195, 197–206, 261
SQL	Structured Query Language 21, 111
STD	Standard Deviation 145
SVI	Search Volume Index 31
SVM	Support Vector Machine 44–46, 79, 80, 82, 156, 157
SysML	Systems Modelling Language 93
TDT	Truth Deception Theory 36
TN	True Negative 83
TP	True Positive 83, 114

UK	United Kingdom 30
UML	Unified Modeling Language 11, 91–94, 99, 100, 185
USA	United States of America 30
VM	Virtual Machine 105, 106, 108, 109, 113, 114
VPN	Virtual Private Network 108, 113, 117
XS	Extended Application Services 107, 111

Appendix C

Ethical Clearance

This appendix provides the final application and approval from the Engineering, Built environment and Information Technology faculty of the University of Pretoria for the research presented in this thesis. The final request for approval was submitted on 19 January 2017 and approval received on 22 February 2017.

C.1 Application for ethical clearance

For office use only	
Assigned EBIT tracking number	EBIT/ /
Date received	

<p>UNIVERSITY OF PRETORIA</p> <p>FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY</p> <p>FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY</p> <p>(EBIT Ethics Committee)</p>
<p>APPLICATION FOR APPROVAL OF A RESEARCH PROJECT</p>
<p>This application form must be read with the relevant UP regulations, as documented in the Code of Ethics for Scholarly Activities, and the Policy and Procedures for Responsible Research. By completing and submitting this form, you declare that you have read these two documents and understand the regulations.</p> <p>Important: Each item must be completed.</p> <p>Complete the form in your word processor. Forms completed in handwriting are not accepted.</p> <p>Where applicable, underline the correct answer (e.g. <u>Yes</u> or No).</p>

1. RESEARCHER DETAILS: (Please include your Supervisor details in this section if you are a student)			
Applicant details:		Supervisor details:	
Initials and surname:	E VAN DER WALT	Initials and surname:	JHP ELOFF
Title:	MISS	Title:	PROF
Email:	estee.vanderwalt@gmail.com	Email:	eloff@cs.up.ac.za
Phone:	***	Phone:	***
Employee/student number:	***	Employee number:	***
Department:	Computer Science	Department:	Computer Science
Are you a student (yes or no):	yes		

2. RESEARCH PROJECT TITLE (use a descriptive title)
Identity Deception Detection on Social Media Platforms

3. RESEARCH PROJECT DETAILS

3.1 Provide a complete but concise description (no more than 5000 characters, including spaces) of the study objectives and study design, so that the relevant ethical aspects can be identified.

- From this, please identify the aspects clearly that you believe require ethics clearance.
- Please note: do NOT submit a complete research proposal. The Ethics Committee will not consider this, but will only consider the documents required for submission of an application.

Human deception detection has been around for many years. Most past research has been psychological in nature. It was found that people lie for various purposes.

Big data platforms, like social media, is not immune from deception. Big data is known by the 3Vs; Volume, Velocity and Variety. The nature of big data makes it very difficult to detect deception as the volumes of data is too much, the data is produced too fast and data is of a heterogeneous nature.

Deception can vary from something harmless, like improving an online social status (Squicciarini and Griffin, 2014), to something harmful like a paedophile grooming minors or terrorist recruitment. The research at hand is interested in identity deception. Known identity features in social media are presented through literature reviews with which potential deceptive individuals can be identified. Examples of such features are:

- What the overall sentiment is of the person.
- Whether people upload a profile image or still use the default.
- Their number of friends and followers.

Ethical clearance is required for the use of Twitter data as part of the research experiment towards identity deception detection, taking cognizance of the study design and measures that will be protect the privacy of individuals. The Twitter data will be used to identify those features that could indicate deception.

Study design:

The research experiment at hand proposes to use the scientific method. The hypothesis is that if we know what identity features lead to deception, we can identify potential deceptive accounts.

The experiment will gather public data from Twitter to build a corpus of data. This will be done through a freely available Twitter API. The data gathered will consist of all individuals tweeting about 'school' or 'homework' including their network of friends and followers. The data will be cleansed to not include any retweets, inactive accounts or accounts from celebrities. The belief is that this will create a corpus of data from adolescents (Schwartz et al., 2013).

Machine learning (e.g. SOM maps) can identify those outlier features leading to potential identity deception. These features are matched to what is already known from the literature review and will then be used towards proving the hypothesis.

Deceptive dummy accounts are injected into the corpus next. These accounts will be manually created and are not actual Twitter accounts. The final part of the experiment hopes to identify these harmful accounts from the corpus and refine the results through appropriate weighting of the identity features.

Measures taken to protect individuals:

Many research papers were considered to understand the ethics around working with big data and social media data in research (Zimmer, 2010) (Rivers and Lewis, 2014) (Li and Wang, 2015).

With this in mind and to protect the privacy of individuals during the research, various methods will be applied to obfuscate identity:

- The corpus will be used to pick those identity features deemed best for the experiment based on the literature review and results from machine learning.
- Final identified features will be in an aggregated format. (Li and Wang, 2015) proposed that data clustering is an additional good mechanism towards privacy preservation. With clustering, the rules used during the experiment is hidden and individuals' identities are irrelevant. The same principles are suggested by (Rivers and Lewis, 2014).
- The research proposes to show that it can identify the manually created, deceptive accounts and not those of actual Twitter individuals. Even if other actual Twitter accounts were identified, they will be omitted from the results.
- The research will at no point require to engage or befriend individuals or gather information from other social media sites using the Twitter data (Rivers and Lewis, 2014).
- Information not required for the research at hand will not be gathered (Rivers and Lewis, 2014) (Zimmer, 2010).
- During the research, privacy preservation techniques defined by Xu and Wang (Li and Wang, 2015) will be applied. To be more specific, the names of individuals will be scrambled.

Finally:

I do not expect to find any criminal activities from the Twitter data used during the study. However, if I do at any point suspect or find any potential crime, it will be reported to the authorities as soon as possible.

The hope is that the research could produce valuable insights for law enforcement agencies in the future; leading to a concept for better protecting the innocent.

Bibliography

Li, J. & Wang, A. G. 2015. A framework of identity resolution: evaluating identity attributes and matching algorithms.

Rivers, C. M. & Lewis, B. L. 2014. Ethical research standards in a world of big data.

Schwartz, H. A. et al 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach.

Squicciarini, A. & Griffin, C. 2014. Why and how to deceive: game results with sociological evidence.

Zimmer, M. 2010. "But the data is already public": on the ethics of research in Facebook.

<p>3.2 Will a research questionnaire/survey be used?</p> <ul style="list-style-type: none"> • If Yes, please answer the next question. If No, ignore the next question. • Please submit your questionnaire, survey questions or interview questions with your application. This will be a separate file that should be submitted as a pdf file, using this filename format: Questionnaire.pdf or Survey.pdf 	Yes	<u>No</u>
<p>3.2.1 Does your questionnaire/survey include any personal questions? (including ANY of the following: name, address, email address, any other information by which a respondent can be identified, gender, age, race, income, medical status)?</p>	Yes	No
<p>3.3 Are employees of a firm, organisation or institution questioned as informant in this study?</p> <ul style="list-style-type: none"> • If Yes, please submit letter(s) of permission from this entity to carry out this study. It should be clear that the person giving permission is authorised to do so and should be on a company letterhead and should include the date and that person's signature. • Where required, your application cannot be considered without this permission. • This letter should be submitted as a pdf file, using this filename format: CompanyPermissionLetter.pdf 	Yes	<u>No</u>
<p>3.4 Will you be surveying or questioning UP students or UP personnel in this study?</p> <ul style="list-style-type: none"> • If Yes, you need to submit a letter or email from the Dean that provides permission for you to include UP personnel or students as participants in your study. • Where this is required, your application cannot be considered without this permission letter. • This letter should be submitted as a pdf file, using this filename format: DeanPermissionLetter.pdf 	Yes	<u>No</u>

4. RESEARCH SUBJECTS		
Does the project involve people as participants, either individually or in groups? If Yes, please answer questions 4.1 to 4.7. If No, continue to section 5.	<u>Yes</u>	No
4.1 Does the study involve people as informants, or does it involve people as research subjects? <i>Informants</i> are people of whom you require an opinion, e.g. people that are interviewed or that take part in a survey. <i>Research subjects</i> are people that actively take part in research, e.g. where biological measurements are made (e.g. heart rate) or where people take part in behavioural tasks (e.g. listening tasks)	Informants	<u>Subjects</u>
4.2 Describe possible safety and health implications that participation in the project may pose. None – gathering of social media data		
4.3 What is the expected duration of participation of people in the project? No participation is required. We will merely collect data publicly available		
4.4 Describe the manner in which confidential information will be handled and confidentiality assured. <ul style="list-style-type: none"> • Only gather public information • Features identified will be in an aggregated format and thus not reveal individual information • Personal information not required for the study will not be gathered • Obfuscation of personal information, like the scrambling of names • Results will not reveal individual's personal information, only those of the deceptive dummy accounts created. 		
4.5 Please explain how and where data will be stored. <ul style="list-style-type: none"> • Data will be stored on a SAP HANA server in Potsdam Germany. Upon completion of the research, the data will be deleted • Data is also downloaded on a personal laptop and storage drive to help in situations of slow remote Internet connection speeds or server downtime • Data will be gathered using the Twitter4J Java API which is available for free use 		
4.6 Is remuneration offered to subjects for participation? If yes, please expand. Not applicable		
4.7 INFORMED CONSENT/ASSENT Informed consent is a requirement for <i>all</i> studies. All participants need to provide individual informed consent, which the researcher should keep on record. An example for an informed consent form appears on the website, but this should be adapted to be very specific about your study and what you will require of participants. Please submit your informed consent form (example) with your application. This should be submitted as a pdf file, using this filename format: InformedConsent.pdf		
4.7.1. Please describe how you will obtain informed consent/assent from your participants (or their caregivers in the case of underage participants). I request to waive this requirement for the research at hand. <ul style="list-style-type: none"> • The study uses public social media data and adheres to the usage and privacy policy as stated by the social media platform used (Twitter - https://twitter.com/privacy?lang=en) • The research will at no point report on individuals from the gathered social media feed or any content which could identify individuals from the feed. 		

<p>4.7.2 Detail the measures you will take to ensure voluntary participation.</p>
--

5. ENVIRONMENTAL IMPACT and HAZARDOUS MATERIALS		
5.1 Does the project have a potentially detrimental environmental impact, or are hazardous materials used in the project?	Yes	<u>No</u>
<ul style="list-style-type: none"> If Yes, you will need to submit a letter of approval from the Department of Facilities and services, Occupational Health and Safety division, before the Ethics Committee can consider your application. If section 5 (this section) is the only aspect of your project for which you require clearance from the Ethics Committee (i.e. no people or animals are included in your study), you should not apply to the Ethics Committee, but should apply for clearance directly to the Occupational Health and Safety division. If No, continue to section 6. 		

6. DISSEMINATION OF DATA
6.1 How and where will your results be published and/or applied?
<p>The results, without direct reference to any individual, will potentially be published in conference and journal papers. Finally, the results will be presented in the final thesis towards a PhD degree. The data might also be given to students at the Computer Science department for further analysis and research help but under the same rules as stated in this application.</p>

7. DECLARATION (Tick the relevant boxes)	
✓	I accept and will adhere to all stipulations pertaining to ethically sound research as locally, nationally and internationally established.
✓	I will conduct the study as specified in the application and will be principally responsible for all matters related to the research.
✓	I shall communicate all changes to the application or any other document before any such is executed in my research, to obtain the necessary permissions from the Ethics Committee.
✓	I will not exceed the terms of reference of the research application or any other documents submitted to the Ethics Committee.
✓	I confirm that I'm not seeking ethics clearance for research that has already been carried out.
✓	I affirm that all relevant information has been provided and that all statements made are correct.
✓	I have familiarised myself with the University of Pretoria's policy regarding plagiarism http://www.aibrary.up.ac.za/plagiarism/index.htm . Plagiarism is regarded as a serious violation and may lead to suspension from the University.
<p style="color: red;">Please submit the completed Declaration By The Researcher form with your application. Please submit this as a pdf file with this filename format: Declaration.pdf</p>	

8. SUBMISSION CHECKLIST

Each item to be submitted should be submitted as a separate pdf file, using the naming convention given earlier in this document or below.

8.1 Have you submitted your application form (this form)? Please submit as a pdf file with this filename format: ApplicationForm.pdf	<u>Yes</u>	No	
8.2 Have you submitted your survey questions, questionnaire or interview questions (where applicable)?	Yes	No	<u>N/A</u>
8.3 Have you submitted the required Informed Consent Form?	Yes	No	<u>N/A</u>
8.4 Have you submitted the Declaration By The Researcher form?	<u>Yes</u>	No	
8.5 Have you submitted permission letters from firms, institutions or organisations where required?	Yes	No	<u>N/A</u>
8.6 Have you submitted a permission letter from the Dean where required?	Yes	No	<u>N/A</u>

C.2 Ethical clearance approval



Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en
Inligtingtegnologie / Lefapha la Boetšenere,
Tikologo ya Kago le Theknolotši ya Tshedimošo

Reference number: EBIT/6/2017

22 February 2017

Ms E van der Walt
Department Computer Science
University of Pretoria
Pretoria
0028

Dear Ms Van der Walt,

FACULTY COMMITTEE FOR RESEARCH ETHICS AND INTEGRITY

Your recent application to the EBIT Research Ethics Committee refers.

Approval is granted for the application with reference number that appears above.

1. This means that the research project entitled "Identity deception detection on social media platforms" has been approved as submitted. It is important to note what approval implies. This is expanded on in the points that follow.
2. This approval does not imply that the researcher, student or lecturer is relieved of any accountability in terms of the Code of Ethics for Scholarly Activities of the University of Pretoria, or the Policy and Procedures for Responsible Research of the University of Pretoria. These documents are available on the website of the EBIT Research Ethics Committee.
3. If action is taken beyond the approved application, approval is withdrawn automatically.
4. According to the regulations, any relevant problem arising from the study or research methodology as well as any amendments or changes, must be brought to the attention of the EBIT Research Ethics Office.
5. The Committee must be notified on completion of the project.

The Committee wishes you every success with the research project.

Prof JJ Hanekom

Chair: Faculty Committee for Research Ethics and Integrity
FACULTY OF ENGINEERING, BUILT ENVIRONMENT AND INFORMATION TECHNOLOGY

Appendix D

The FSOC proposal

Resources from the HPI Future SOC lab in Potsdam, Germany was used for this research from 2014 to 2018. The resources is highly available and includes access to large distributed storage which is crucial for big data research projects. This appendix shows an example of a request for lab resources required on an ongoing six-monthly period. In addition, research projects are required to provide a technical progress report and research poster at the end of each six months. Examples of these deliverables are given next.

D.1 HPI Future SOC resource example request

HPI Future SOC Lab - Call for Projects Fall 2016



The HPI Future SOC Lab offers researchers free of charge access to a powerful infrastructure to conduct their research activities. Over 310 research projects were supported by our lab since it was launched together with EMC, Fujitsu, Hewlett Packard Enterprise (HPE), and SAP in 2010.

Topics of interest for the HPI Future SOC Lab include but are not limited to:

- **Service-oriented computing** (SOC) and efficient management of concurrency and memory,
- **Cloud Computing** for engineering and hosting distributed applications,
- New applications exploring new ways of data processing, such as **in-memory technology**,
- **Multicore architectures / GPU**,
- **Hybrid Computing**,
- **On-demand and delivery models** for business applications,
- Applications in the area of **Machine Learning**.

Researchers can apply to gain access to this infrastructure by submitting project proposals. Project proposals are reviewed and approved by a Steering Committee which is led by the head of HPI, Prof. Dr. Christoph Meinel, and which comprises representatives from HPI and the industrial partners, EMC, Fujitsu, HPE, and SAP.

All proposals must be submitted by October 06, 2016.

The Steering Committee decisions will be announced as part of the **HPI Future SOC Lab Day on November 03, 2016**.

Projects are approved for the length of one Future SOC Lab period, which starts November 04 and is approximately six months long. Proposals for follow-up projects can be submitted in subsequent call for projects, with the next one being published in Spring 2017.

Project proposal

Your project proposals should be short and concise. Please use the following template for your submissions.

Please do not hesitate to get back to us in case of questions or comments.

Contact and Proposal Submission: futuresoc-lab-info@hpi.de

CONTACT INFORMATION

PRINCIPAL INVESTIGATOR:

NAME (INCL. TITLE):

EMAIL ADDRESS:

RESEARCH INSTITUTION:

STREET ADDRESS:

CITY: ZIP CODE:

STATE:

CONTACT AUTHOR:

NAME (INCL. TITLE):

EMAIL ADDRESS:

BIOGRAPHY OF CONTACT AUTHOR (MAX. 1000 CHARACTER):
(OPTIONAL: BIOGRAPHIES OF ADDITIONAL PARTICIPANTS)

OPTIONAL: ADDITIONAL PARTICIPANTS AND INSTITUTIONS:

PROJECT INFORMATION

PROJECT TITLE:

MANAGEMENT SUMMARY (MAX. 350 CHARACTER):

PLEASE ATTACH TO THIS FILE A SHORT ABSTRACT AND FURTHER INSIGHT INTO THE PROPOSED RESEARCH (MAX. 2500 CHARACTER, PDF FORMAT)
(APPROACH TO ADDRESS THE RESEARCH TOPIC, PROJECT PLAN, REFERENCES TO PRIOR PRACTICAL EXPERIENCE OR RELATED PUBLICATIONS)

WHICH FUTURE SOC LAB RESOURCES DO YOU REQUIRE FOR YOUR RESEARCH?

DO YOU NEED ACCESS TO THE CLUSTER? YES NO

IF YES, HOW MANY NODES? (1-25)

HPE CLOUDSYSTEMS 9:

DO YOU NEED ACCESS TO THE HPE CLOUDSYSTEMS 9? YES NO IF YES, HOW MANY BLADES? (1-16):

INFRASTRUCTURE ORCHESTRATION: YES NO

MULTI-CORE SERVER:

RAM: 16 GB 32 GB 64 GB 256 GB 1 TB 2 TB

CPU (CORES): 24 32 64

GPU ACCESS: YES

VIRTUAL MACHINES - PLEASE SPECIFY THE AMOUNT AND SIZING:

STORAGE - PLEASE SPECIFY YOUR REQUIREMENTS IN DETAIL:

DO YOU NEED A SAP HANA INSTANCE? YES NO IF YES: SAP HANA PREDICTIVE ANALYSIS LIBRARY (PAL) R

OTHER SOFTWARE:

REQUIRED FUTURE SOC LAB RESOURCES (HARDWARE AND SOFTWARE) FOR THE SIX MONTHS TO COME (THE FUTURE SOC LAB PERIOD)

IS EXCLUSIVE ACCESS REQUIRED? YES NO

FOR HOW LONG AND AS OF WHEN DO YOU PLAN TO ACCESS TO THE REQUIRED RESOURCES? (E. G. 40 HOURS FOR TESTING AND ANOTHER 80 HOURS TO RUN EXPERIMENTS BEGINNING IN APRIL 2016 OR THREE MONTHS BEGINNING IN MAY 2016)

PLEASE DO NOT HESITATE TO GET BACK TO US IN CASE OF QUESTIONS OR COMMENTS.

CONTACT AND PROPOSAL SUBMISSION: futuresoc-lab-info@hpi.de

WEB LINK: http://www.hpi.de/future_soc_lab

D.2 Technical research progress report example

A Big Data Science Experiment - Protecting Minors on Social Media Platforms

Estée van der Walt
Department of Computer Science
Security & Data Science Research Group
University of Pretoria, South Africa
estee.vanderwalt@gmail.com

Prof J.H.P. Eloff
Department of Computer Science
Security & Data Science Research Group
University of Pretoria, South Africa
eloff@cs.up.ac.za

Abstract

The protection of individuals on big data platforms, like social media, is a challenge. This is in large due to the nature of these platforms allowing individuals the freedom to create and use any persona online without regulation. The victims, that includes minors, are exposed to various cyber threats of which identity deception is an example. In past research, various features extracted from social media platforms have been proposed to identify identity deception as found in online bot accounts. These same features were found to not have the same success when applied to the detection of deceptive online human accounts. We looked towards the field of social sciences, and more specifically psychology, to identify those features most likely to indicate lying humans. Supervised machine models were built in the hope to predict potential identity deception found for human accounts on social media platforms using these additional newly found features. For the research at hand all past results are compared to evaluate the hypothesis that these features, added through knowledge gained from psychology, improved the detection of identity deception in human on social media platforms. Finally, an Identity Deception Score (IDS) is presented in the hope to further explain why specific individuals are perceived as being potentially deceptive.

1 Project idea

Social media allows individuals to add content at will [1]. More data does however create more cyber threats [2]. The volume of data alone makes it difficult to monitor each online social media account. This volume, combined with data being added at great speed and in different formats, like video and image, adds to the challenge of protecting individuals against a plethora of cyber threats [3].

One such cyber threat is identity deception [4]. This is when a human lies about who they are. In a social

media context, this is when an online social media account's information (human or bot) is not correlating with the truth about the actual account holder [5]. An example would be a man in his 40s describing himself as a 14-year-old boy on Facebook.

Past research has tried to address this problem by finding those deceptive accounts generated by bots [6] [7] [8]. There are more examples of such fake accounts which allows for more research opportunity. Past research has also proposed various new features that could indicate deception, like the friend/follower ratio in Twitter [9]. No research has been found to date, presenting features, new or engineered, given what we have learnt about the deceptive nature of humans from other research fields like the social sciences.

Within the field of social sciences, and specifically psychology, much research has been done in trying to understand why humans lie. Not only were the humans' motive for lying investigated amongst others [10], but also what they lie about [11]. For this research we have focused on what humans have been found to lie about and what similar features are available or could be engineered on social media platforms. The hope is that these additional engineered features will aid in the detection of identity deception of humans on social media platforms.

Various research experiments have been executed to evaluate the hypothesis that these features from psychology, aid in the more successful detection of identity deception. The research project has been divided into various steps discussed in more detail during previous research papers [12] and follows a scientific approach. The focus of this phase of the research, highlighted in figure 1, was to evaluate the results from all previous experiments. The results are discussed in section 3 of this report. Lastly this phase also proposes an Identity Deception Score (IDS) to further explain why one human is perceived to be

deceptive and another not.

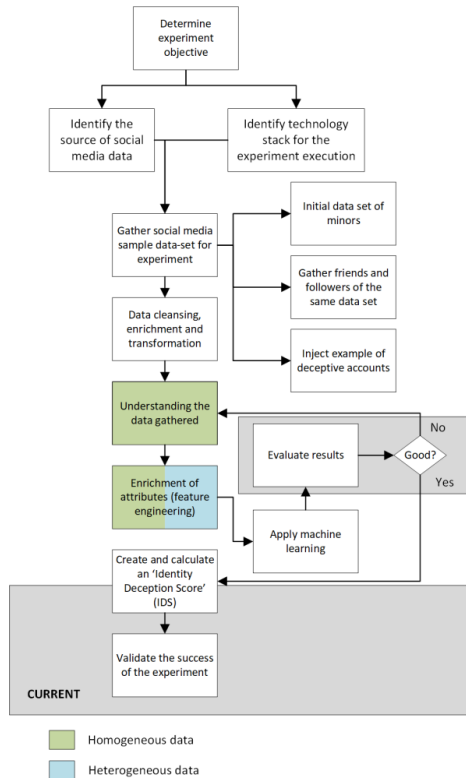


Figure 1: The project process diagram

1.1 Main deliverables

The main deliverables of the past six months were:

- To combine all results from previous experiments.
- To compare the performance of all machine learning algorithms used across all experiments.
- To create an IDS. The IDS is supposed to indicate whether an account is deceptive or not and should also include reasoning as to why that decision was reached.

2 Use of HPI Future SOC Lab resources

To reiterate past feedback, the following resources were used for the research at the HPI Future SOC lab:

- Twitter: The Twitter4j Java API was used to dump the data needed for the experiment in a big data repository [13].
- Hortonworks Hadoop 2.4 [14]: For the purposes of this experiment HDP Hadoop runs on an

Ubuntu Linux virtual machine hosted in “The HPI Future SOC”- research lab in Potsdam, Germany. This machine contains 4TBs of storage, 8GB RAM, 4 x Intel Xeon CPU E5-2620 @2GHz and 2 cores per CPU. Hadoop is well known for handling heterogeneous data in a low-cost distributed environment, which is a requirement for the experiment at hand.

Flume: Flume is used as one of the services offered in Hadoop to stream initial Twitter data into Hadoop and into SAP HANA.

Ambari: For administration of the Hadoop instance and starting/stopping the services like Flume.

- Java: Java is used to enrich the Twitter stream with additional information required for the experiment at hand and automate the data gathering process.
- SAP HANA [15]: A SAP HANA instance is used which is hosted in “The HPI Future SOC”- research lab in Potsdam, Germany on a SUSE Linux operating system. The machine contains 4TBs of storage, 2TB of RAM (1.4TB effective) and 32CPUs / 100 cores. The in-memory high-performance processing capabilities of SAP HANA enables almost instantaneous results for analytics.

The XS Engine from SAP HANA is used to accept streamed Tweets and populate the appropriate database tables.

- Machine learning APIs: Various tools are considered to perform classification, analysis and apply deep learning techniques on the data. These include the PAL library from SAP HANA, SciPy libraries in Python, Spark MLlib on Hadoop, and the Hadoop Mahout service. For the research, R was the final choice. This decision was made due to support on this platform and libraries freely being available on the web community at a large scale.
- An additional Linux machine was provided for the lab to aid in the running of the CPU and memory intensive machine learning algorithms. The VM has 8 cores and 64GB of RAM.
- Visualization of the results will be performed by the libraries in R [16] and PowerBI [17] where appropriate.

The following ancillary tools were used as part of the experiment:

- For connection to the FSOC lab we used the OpenVPN GUI as suggested by the lab.
- For connecting and configuration of the Linux VM instance we used Putty and WinSCP
- For connecting to the SAP HANA instance, we used SAP HANA Studio (Eclipse) 1.80.3

3 Findings in the Fall 2017 semester

For this phase of the research, the results from previous supervised machine learning results were compared against each other. This is done in the hope to understand which feature sets were best at identifying deceptive online humans. This evaluation was compared against the hypothesis that engineered features, borrowed from the field of psychology, can significantly increase the successful detection of identity deception.

The three experiments were as follows:

- The first experiment used attributes found in social media alone as the features to identify deceptive online accounts.
- The second experiment looked towards past research done in the detection of bot account to add more engineered features in the supervised model trained to detect deceptive online accounts.
- The last experiment engineered new features given past research from psychology on why and what humans lie about. These new features were added in the hope to train an even better supervised machine learning model.

The results from all three experiments are presented in Table 1. It shows how each successive experiment improved on the results from the previous.

Machine learning algorithm	Exp1 (Meta)	Exp2 (Bots)	Exp3 (Psychology)
svmRadial	15.09%	32.16%	89.18%
rf	31.77%	49.75%	96.15%
J48	27.16%	44.53%	91.64%
bayesglm	15.33%	33.41%	87.84%
knn	22.62%	40.43%	89.06%
Adaboost	29.24%	47.54%	94.44%
rpart	23.05%	32.37%	91.56%
nnet	27.90%	41.07%	89.43%

Table 1: Comparative results from experiments

Lastly, an IDS were produced. The IDS predict the potential of a human being deceptive. An example of such a scenario would be once the above machine learning model is applied, it produced only a prediction of 84% that person X is deceptive. This information is however insufficient to explain why this decision was reached. The IDS propose to also give additional information as to what features attributed to the 84% and how much. An example of one IDS is shown in Table 2.

Feature	Contribution
FOLLOWERS_COUNT	-0.12%
FRIENDS_COUNT	-0.09%
FF_RATIO	-0.01%
LISTED_COUNT	0.01%
USERNAME_LENGTH	12.84%
GEO_ENABLED	0.03%
PROFILE_HAS_URL	-0.43%
ACCOUNT_AGE_IN_MONTHS	1.84%
HAS_NAME	23.51%
HAS_IMAGE	0.05%
DUP_PROFILE	0.88%
HAS_PROFILE	0.08%
STATUS_COUNT	0.13%
DISTANCE_LOCATION	0.09%
DISTANCE_TZ	3.59%
COMPARE_GENDER	-0.56%
LEVENSHTEIN	24.51%
COMPARE_AGE	33.66%

Table 2: IDS contribution example result

With this information it is now possible to understand why a decision was reached.

4 Architecture

The SAP HANA instance, virtual machines and storage was provided by the HPI FSOC research lab and the following is worth mentioning:

- There were no issues in connection.
- The lab was always responsive and helpful in handling any queries.
- The environment is very powerful, and more than enough resources are available which makes the HPI FSOC research lab facilities ideal for the experiment at hand
- Without the additional VM with more cores, we would not have been able to perform the machine learning computations.

Overall, we found that the environment and its power enabled the collection and handling of a big dataset without issue. The support of the HPI FSOC research lab is greatly appreciated.

5 Next steps for 2018

The deliverables for this next phase are as follow:

- To explain the IDS results in more detail
- Investigate whether the IDS could lead us to understand which features online as most indicative of deception.

References

- [1] S. Kaisler, F. Armour, J. A. Espinosa, and W. Money, "Big Data: Issues and Challenges Moving Forward," pp. 995-1004, 2013.
- [2] R. P. Khandpur, T. Ji, S. Jan, G. Wang, C.-T. Lu, and N. Ramakrishnan, "Crowdsourcing Cybersecurity: Cyber Attack Detection using Social Media," arXiv preprint arXiv:1702.07745, 2017.
- [3] H. F. Lipson, "Tracking and tracing cyber-attacks: Technical challenges and global policy issues," CARNEGIE-MELLON UNIV PITTSBURGH PA SOFTWARE ENGINEERING INST2002.
- [4] M. Tsikerdekis and S. Zeadally, "Detecting and Preventing Online Identity Deception in Social Networking Services," *Internet Computing, IEEE*, vol. 19, pp. 41-49, 2015.
- [5] J. T. Hancock and C. L. Toma, "Putting your best face forward: The accuracy of online dating photographs," *Journal of Communication*, vol. 59, pp. 367-386, 2009.
- [6] B. Van den Belt. (2012), How to recognize Twitter bots: 7 signals to look out for. Available: <http://www.stateofdigital.com/how-to-recognize-twitter-bots-6-signals-to-look-out-for/>
- [7] B. Rashidi and C. Fung, "BotTracer: Bot user detection using clustering method in RecDroid," in *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*, 2016, pp. 1239-1244.
- [8] J. P. Dickerson, V. Kagan, and V. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, 2014, pp. 620-627.
- [9] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: efficient detection of fake Twitter followers," *Decision Support Systems*, vol. 80, pp. 56-71, 2015.
- [10] M. Drouin, D. Miller, S. M. Wehle, and E. Hernandez, "Why do people lie online? "Because everyone lies on the internet"," *Computers in Human Behavior*, vol. 64, pp. 134-142, 2016.
- [11] S. Utz, "Types of deception and underlying motivation: What people think," *Social Science Computer Review*, vol. 23, pp. 49-56, 2005.
- [12] E. Van der Walt and J. H. P. Eloff, "Protecting minors on social media platforms - A Big Data Science experiment " HPI Cloud Symposium "Operating the Cloud", 3 Nov 2015 2015.
- [13] Twitter. Twitter API. Available: <https://dev.twitter.com/overview/api>
- [14] Hortonworks. (2014). Hortonworks with SAP Hana. Available: <http://hortonworks.com/partner/sap/>
- [15] S. SE, "SAP HANA," vol. Enterprise Edition, pp. SAP HANA is an in-memory, column-oriented, relational database management system., 2017.
- [16] R. Ihaka and R. Gentleman, "R," p. R is an open source programming language and software environment for statistical computing and graphics, 2017.
- [17] Microsoft, "Microsoft Power BI," vol. August, 2017, p. Power BI is a business analytics service provided by Microsoft, 2017.

D.3 Research poster example

A Big Data Science Experiment Protecting Minors on Social Media Platforms

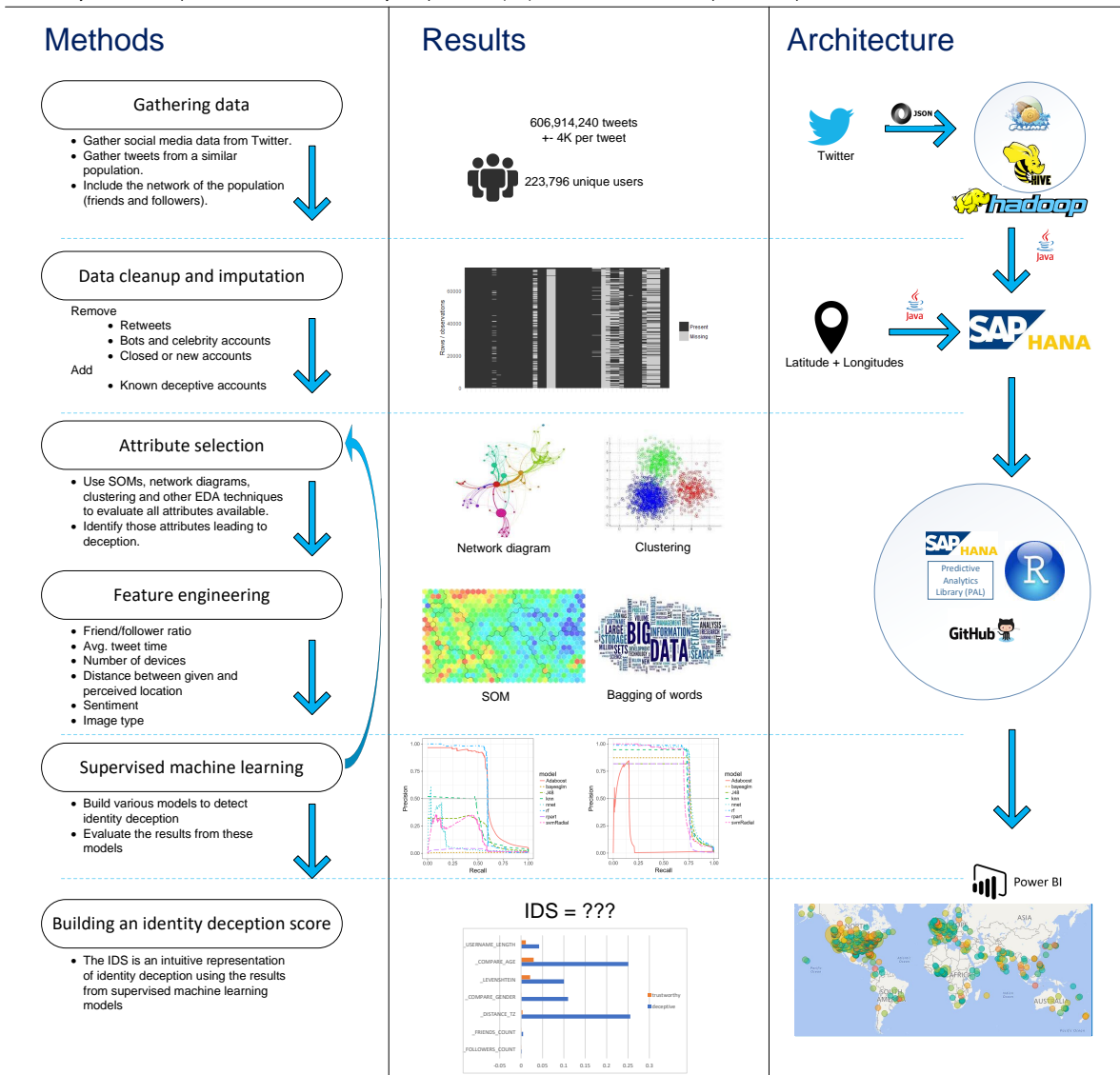
Estée van der Walt¹, J.H.P. Eloff²
estee.vanderwalt@gmail.com¹, eloff@cs.up.ac.za²



Cyber-security and Big Data Science research group, Department of Computer Science, University of Pretoria, South Africa

Overview

- Many people present a false identity for various purposes, whether to remain anonymous or for some other malicious purpose, like online grooming.
- The big data characteristics of social media make it not only easier for people to deceive others about their identity, but also harder to prevent or detect identity deception.
- The research proposes to use features from social sciences to detect identity deception.
- Lastly the results are presented in the form of an Identity Deception Score (IDS) of which the results are transparent and explainable.



Conclusion

- The research at hand identified and evaluated various features that could play a role in identity deception on a social media platform.
- It was found that engineered features previously used to detect non-human accounts (bots) did not perform well to apply directly to humans.
- It was found that engineered features built from knowledge in social sciences (psychology) could better the prediction of identity deception considerably.
- The results are difficult to explain due to the nature of machine learning models (usually black box models).

Future work:

- Determine which attributes or features contributed the most to identity deception during the previous supervised machine learning experiments.
- Build a simple, intuitive algorithm to score each user account knowing the contribution mentioned before. This score will be known as the Identity Deception Score (IDS).
- Use the IDS to explain the results in a more intuitive way than current black box machine learning models.

Appendix E

Publications and contributions

Throughout this study, results of the research have been published in the following journal, conference and technical report papers:

E.1 Journal papers

Van Der Walt, Estée and Eloff, J.H.P. Using Machine Learning to Detect Fake Identities: Bots vs Humans. IEEE Access, Volume 6 (2018), pages 6540-6549. ISSN: 2169-3536.

Abstract: There is a growing number of people who hold accounts on social media platforms (SMPs) but hide their identity for malicious purposes. Unfortunately, very little research has been done to date to detect fake identities created by humans, especially so on SMPs. In contrast, many examples exist of cases where fake accounts created by bots or computers have been detected successfully using machine learning models. In the case of bots these machine learning models were dependent on employing engineered features such as the 'friend-to-followers ratio'. These features were engineered from attributes, such as 'friend-count' and 'follower-count', which are directly available in the account profiles on SMPs. The research discussed in this paper applies these same engineered features to a set of fake human accounts in the hope of advancing the successful detection of fake identities created by humans on SMPs.

Estée van der Walt, J.H.P. Eloff, and Jacomine Grobler. Cyber-security: Identity Deception Detection on Social Media Platforms. Computers and Security Journal (2018). ISSN: 0167-4048.

Abstract: Social media platforms allow billions of individuals to share their thoughts,

likes and dislikes in real-time, without any censorship. This freedom, however, comes at a cyber-security risk. Cyber threats are more difficult to detect in a cyber world where anonymity and false identities are ever-present. The speed at which these deceptive identities evolve calls for solutions to detect identity deception. Cyber-security threats caused by humans on social media platforms are widespread and warrant attention. This research posits a solution towards the intelligent detection of deceptive identities contrived by human individuals on social media platforms (SMPs). Firstly, this research evaluates machine learning models by using attributes such as the “profile image” found on SMPs. To improve on the results delivered by these models, past research findings from the field of psychology, such as that humans lie about their gender, are used. Newly engineered features such as “gender-derived-from-the-profile-image” are evaluated to grasp whether these features detect deception with greater accuracy. Furthermore, research results from detecting non-human (also known as bot) accounts are also leveraged to improve on the initial results. These machine learning results are lastly applied to a proposed model for the intelligent detection and interpretation of identity deception on SMPs. This paper shows that the cyber-security threat of identity deception can potentially be minimized, should the vulnerability in the current way of setting up user accounts on SMPs be re-engineered in the future.

Estée van der Walt and Prof Jan Eloff. Unravelling ‘Big Data’ for the enterprise. Innovate, Faculty of Engineering, Built Environment, and Information Technology, University of Pretoria. Issue 11 (2016).

Abstract: The term ‘Big Data’ is pervasive in business discussions as it is seen as an innovation enabler. However, the term can be quite confusing. Is ‘Big Data’ really that different from ‘data’? As data has been around for a while, it might appear that Big Data is the result of data moving away from a strictly ‘text-based, structured format’ to a heterogeneous unstructured format.

E.2 Conference papers

Estée van der Walt and J.H.P. Eloff. Are attributes on Social Media Platforms usable for assisting in the automatic detection of Identity Deception? International Symposium on Human Aspects of Information Security & Assurance (HAISA). Dundee, Scotland. 29-31 August 2018. Pages 57-66. ISBN: 978-0-244-40254-9

Abstract: Social Media Platforms (SMPs) allow any person to easily communicate with their friends or the general public in the large. People can now be targeted at great scale, most often for malicious purposes. The mere fact that more people are using SMPs, exposes more people to various forms of cyber threats such as cyber-bullying. The problem is that many of these cyber-attacks involve some form of identity deception, where the attackers lie about who they are. The solution proposed in this paper is to work towards developing a model for Identity Deception Detection (IDD) on SMPs by identifying and using metadata that is freely available on SMPs. This metadata includes attributes that describes a user account on an SMP. The aim is to use only these attributes, as opposed to the contents of a communication, for determining if people are lying about their identities. A prototype is discussed that runs an experiment using the metadata (attributes) that defines the identity of a user on an SMP. The results show promise for further research in developing solutions for assisting with the automatic detection of identity deception.

Estée van der Walt and J.H.P. Eloff. Creating an environment for detecting Identity Deception. Potsdam, Germany. November 2017.

Abstract: In today's interconnected world we are all exposed to potentially harmful behaviour caused by fake identities, be they those of people or machines. It is difficult to discern whether you are communicating to another entity you can trust. Today fake identities and identity deception constitute a cybercrime threat to all people connected to cyber communities. Detecting Identity Deception in a small environment, say for example where there are 1000 people, is expected to be a feasible task. This is however not the case where the numbers of people in an environment are running into millions - if not billions - such as is the case on Social Media Platforms (SMPs). This paper focuses on answering the following question: "Is it at all possible to detect Identity Deception on big data platforms?" Furthermore, a discussion is provided about the type of environment that would be required to mine and process 'big data' in order to detect Identity Deception on social media platforms.

Estée van der Walt and J.H.P. Eloff. Identity Deception detection on social media platforms. The International Conference on Information Systems Security and Privacy (ICISSP). Porto, Portugal; 19-21 February 2017.

Abstract: The bulk of currently available research in identity deception focuses on understanding the psychological motive behind persons lying about their identity. However, apart from understanding the psychological aspects of such a mindset, it is

also important to consider identity deception in the context of the technologically integrated society in which we live today. With the proliferation of social media, it has become the norm for many people to present a false identity for various purposes, whether for anonymity or for something more harmful like committing paedophilia. Social media platforms (SMPs) are known to deal with massive volumes of big data. Big data characteristics such as volume, velocity and variety make it not only easier for people to deceive others about their identity, but also harder to prevent or detect identity deception. This paper describes the challenges of identity deception detection on SMPs. It also presents attributes that can play a role in identity deception detection, as well as the results of an experiment to develop a so-called Identity Deception Indicator (IDI). It is believed that such an IDI can assist law enforcement with the early detection of potentially harmful behaviour on SMPs.

Estée van der Walt and J.H.P. Eloff. An automated identity deception framework for big data platforms. Annual Conference of the South African Institute of Computer Scientists and Information Technologist 2016 (SAICSIT): M & D Symposium. Johannesburg, South Africa. 26 September 2016

Abstract: Identity deception is a regular occurrence on social media sites like Twitter and Facebook. Not only is it very hard to prevent but also could have severe consequences. The research at hand propose a framework for automated identity deception detection which culminates in an early warning identity deception indicator.

Estée van der Walt and J.H.P. Eloff. A Big Data Science experiment – Identity Deception Detection. 2015 International Conference on Computational Science and Computational Intelligence (CSCI). Pages 416-419. ISBN: 1467397954.

Abstract: Identity Deception Detection is a problem on social media platforms today. Not only are there challenges towards determining the authenticity of people, but also with analysing the data that forms part of the communications. These data are of heterogeneous type and include photos, videos and sound. Furthermore, most social media platforms are operating in an uncontrolled environment. Any person can contribute content and take part. Even though age restrictions do exist there are no enforcement of these laws and honesty of the public is expected. This is dangerous for minors specifically as they are either unaware of the dangers or not mature enough to be responsible for their actions online. Online predators are aware of this fact and

targeting this group specifically. This paper presents work-in progress towards developing an intelligent Identity Deception Indicator (IDI). It is envisaged that this work could eventually assist authorities in doing large-scale observation on publicly available social media platforms, such as Twitter. Of particular interest are those personas whose behaviour and online content does not fit with the age group they are conversing with.

Estée van der Walt and J.H.P. Eloff. Protecting minors on social media platforms - A Big Data Science experiment. HPI Cloud Symposium "Operating the Cloud". Potsdam, Germany. 3 Nov 2015

Abstract: Interpersonal communications on social media, hosted via cloud computing infrastructures, has become one of the most common online activities. This is especially so for children and adolescents (minors) who may be accidentally and intentionally exposed to cyber threats such as cyber bullying, pornography and paedophilia. Most of these unwanted activities deals with some form of identity deception. The problem is that existing countermeasures, such as plug-ins for safe browsing, are inadequate for protecting minors against the threat of identity deception. Furthermore, this problem is propounded by the complexities of the volume and type of data (big data) on social media platforms. This paper presents work-in-progress that leverages on the advances made in big data and data science to assist in the early detection of identity deception and thereby to protect minors using social media platforms. The output of this research project can assist authorities to pro-actively monitor social media feeds and identify potential online personas who are not who they pose to be.

E.3 HPI Future SOC technical research reports

The FSOC lab in Potsdam, Germany requires all projects to provide feedback on a six-monthly basis. The feedback is requested in the form of a technical progress report that includes a research poster. The research poster was displayed during their FSOC Spring and Autumn lab days with examples of each provided in the previous appendix. Feedback was provided for the following lab period (end date of the six month period in brackets):

- **2014 Spring (October 2014)** – The initial proposal for research submitted to use Shannon-Entropy as an uncertainty indicator for big data.
- **2014 Fall (March 2015)** – The research environment was prepared during this

period. This included the setup of a local instance of Hadoop and setup of Flume to gather Twitter data.

- **2015 Spring (October 2015)** – The main deliverables of this period were to gather a big dataset for the experiment consisting of minors. Data exploration was further performed on the gathered data to gain a better understanding of the SMP attributes available.
- **2015 Fall (March 2016)** – During this period additional Twitter data was gathered of friends and followers. The gathering process was automated, and data was stored in SAP HANA. Initial exploration with various machine learning libraries was started. Various machine learning algorithms were tested from the Caret package in R.
- **2016 Spring (October 2016)** - The main deliverables of this period were to prepare the data for machine learning. Furthermore, data was enriched using knowledge from related work in bots and psychology. External APIs, like the Google Face API, was used to extract the gender and age from profile images of a Twitter account holder.
- **2016 Fall (March 2017)** – During this period an initial version of an IDDM was produced. Three main experiments were performed: the first using SMP attributes only, the second using additional knowledge from related work in bots, and the last using additional knowledge from psychology on why humans lie.
- **2017 Spring (October 2017)** – During this period the IDDMLM and IDDSM models were built. Various ideas were tested to reach a final IDDSM that uses entropy to explain why a human can be perceived as being deceptive.
- **2017 Fall (March 2018)** – During this period a model that assist in the detection of human identity deception on SMPs were automated and its results explained using the IDDSM.

Appendix F

Disclosure of the machine learning results

F.1 Introduction

This appendix shows the detailed results from all machine learning experiments. The next few sections will present the results for all experiments used in the 'discover' component of the prototype. In this component the dataset was either used 'as-is' or enriched before using it to train various machine learning models to detect identity deception. Each experiment resulted in measure to evaluate the accuracy of the models. Each experiment was also run at least 30 times to ensure the results are consistent and also over various dataset sizes. These detailed results will be presented in the next few sections.

F.2 Results from Experiment 1

In this experiment the attributes found in SMPs were used 'as-is'. The intention was to understand if the attributes in SMPs alone are sufficient to detect identity deception. The experiment was run 30 times with the results shown in Tables D.1 to D.30.

F.3 Results from Experiment 2

In this experiment related features found to have had success in the detection of bots on SMPs are used to enrich the existing dataset. The experiment aims to determine whether these additional features improve the detection of human identity deception detection on SMPs. The experiment was run 30 times with the results shown in Tables D.31 to D.60.

F.4 Results from Experiment 3

In this experiment related features found to have had success in the detection of humans lying in general (psychology) are used to enrich the existing dataset. The experiment aims to determine whether these additional features improve the detection of human identity deception detection on SMPs. The experiment was run 30 times with the results shown in Tables D.61 to D.90.

Table F.71: Experiment 3 - Run 11

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.22	42.67	35.59	82.72	49.77	91.89	54.51	38.366
rf	91.35	60.80	50.63	91.81	65.27	95.70	64.86	62.737
J48	91.79	61.78	52.09	90.29	66.07	94.21	57.55	84.397
bayesglm	84.31	40.86	34.07	82.75	48.27	91.90	54.33	3.460
kknn	91.62	60.10	51.60	86.05	64.51	94.24	63.10	54.352
Adaboost	91.55	60.52	51.32	88.37	64.94	95.20	69.89	654.020
rpart	87.27	49.56	40.25	90.61	55.74	89.26	38.25	4.060
nnet	91.30	60.50	50.46	91.31	65.00	96.35	73.87	36.846

Table F.72: Experiment 3 - Run 12

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.70	41.73	34.75	83.07	49.00	92.14	53.13	38.767
rf	91.69	61.75	51.73	91.44	66.08	95.78	70.73	62.503
J48	92.22	63.16	53.59	90.27	67.25	95.73	67.89	85.074
bayesglm	84.67	41.74	34.73	83.25	49.01	92.05	52.53	3.252
kknn	91.80	60.94	52.19	87.07	65.26	94.26	63.66	54.920
Adaboost	92.17	61.80	53.59	85.76	65.96	95.28	70.58	634.882
rpart	87.21	49.24	40.08	90.00	55.46	89.01	37.97	3.417
nnet	91.61	61.36	51.47	90.91	65.73	96.37	74.26	37.453

Table F.73: Experiment 3 - Run 13

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.41	41.13	34.27	82.99	48.51	91.84	53.68	37.068
rf	91.35	60.63	50.63	91.20	65.11	95.59	69.31	62.220
J48	90.70	58.78	48.64	91.60	63.54	94.21	52.55	81.160
bayesglm	84.36	41.14	34.24	83.28	48.52	91.76	53.25	3.205
kknn	91.49	60.01	51.11	87.31	64.47	93.92	62.36	55.615
Adaboost	91.82	60.72	52.31	85.95	65.04	94.92	68.94	644.331
rpart	87.17	49.10	39.99	89.79	55.33	88.88	37.83	3.388
nnet	91.17	59.95	50.04	90.77	64.52	96.02	73.27	36.279

Table F.74: Experiment 3 - Run 14

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.50	41.29	34.41	82.91	48.64	91.95	54.52	36.743
rf	91.60	61.59	51.42	91.92	65.95	95.45	64.00	63.862
J48	91.71	61.60	51.82	90.59	65.92	95.80	70.95	83.514
bayesglm	84.68	41.77	34.75	83.33	49.04	91.93	54.22	3.738
kknn	91.81	60.80	52.24	86.45	65.13	94.33	64.33	54.174
Adaboost	91.95	60.85	52.81	84.93	65.13	95.08	68.76	668.122
rpart	85.35	45.97	36.93	92.64	52.80	88.89	35.46	4.361
nnet	91.51	61.20	51.12	91.57	65.61	96.32	75.25	36.610

Table F.75: Experiment 3 - Run 15

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.28	43.10	35.81	83.65	50.15	91.85	53.94	37.951
rf	91.79	62.18	52.06	91.84	66.45	95.67	65.59	62.131
J48	91.97	62.38	52.71	90.37	66.58	94.44	54.61	83.831
bayesglm	84.37	41.32	34.31	83.79	48.69	91.89	53.65	3.597
kknn	91.44	59.68	50.97	86.61	64.17	94.43	64.42	55.256
Adaboost	92.00	61.26	52.97	85.79	65.50	95.31	70.54	657.326
rpart	87.43	49.78	40.54	90.08	55.91	89.22	38.43	3.934
nnet	92.02	62.55	52.86	90.48	66.73	96.29	71.82	36.765

Table F.76: Experiment 3 - Run 16

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.59	43.60	36.29	83.23	50.55	92.00	53.94	37.443
rf	91.67	61.75	51.65	91.68	66.08	95.57	64.46	62.458
J48	91.59	61.34	51.39	91.12	65.72	95.42	60.88	83.702
bayesglm	84.80	42.12	34.99	83.68	49.34	92.09	54.12	3.284
kknn	91.94	61.33	52.72	86.75	65.58	94.68	65.33	55.780
Adaboost	92.14	62.28	53.37	87.95	66.43	95.43	70.79	656.540
rpart	87.14	49.23	39.98	90.48	55.46	89.30	38.04	3.850
nnet	91.50	60.99	51.10	90.85	65.41	96.37	74.65	36.571

Table F.77: Experiment 3 - Run 17

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.56	41.43	34.51	83.01	48.75	91.80	53.59	36.453
rf	91.75	62.01	51.90	91.76	66.30	95.28	65.65	62.664
J48	91.72	61.87	51.81	91.57	66.18	94.80	63.85	83.622
bayesglm	84.67	41.77	34.74	83.36	49.05	91.75	53.34	3.125
kknn	91.83	61.09	52.28	87.28	65.39	94.32	63.92	55.366
Adaboost	92.45	62.87	54.65	86.21	66.89	95.40	71.24	668.840
rpart	85.33	45.80	36.85	92.21	52.66	88.62	35.29	4.074
nnet	91.89	61.93	52.46	89.63	66.18	96.14	71.55	36.627

Table F.78: Experiment 3 - Run 18

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.47	41.15	34.33	82.67	48.51	91.86	53.97	38.026
rf	91.74	61.81	51.89	91.09	66.12	95.66	69.04	62.217
J48	90.82	58.99	48.99	91.07	63.71	94.21	53.01	83.903
bayesglm	84.53	41.34	34.45	82.93	48.68	91.80	53.43	3.349
kknn	91.67	59.98	51.79	85.04	64.38	94.07	63.94	55.007
Adaboost	92.24	61.93	53.90	85.31	66.06	95.11	69.17	661.464
rpart	87.08	48.87	39.80	89.71	55.14	88.82	37.65	4.220
nnet	91.85	61.72	52.32	89.33	65.99	96.00	71.43	35.885

Table F.79: Experiment 3 - Run 19

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.41	43.20	35.97	83.12	50.21	91.91	53.49	36.781
rf	91.64	61.60	51.54	91.55	65.95	95.62	65.95	61.766
J48	91.67	61.60	51.67	91.09	65.94	95.11	66.77	83.733
bayesglm	84.40	41.17	34.28	83.15	48.54	91.93	53.39	3.296
kknn	91.54	59.98	51.30	86.59	64.43	94.18	64.27	55.871
Adaboost	91.96	61.66	52.76	87.76	65.90	95.29	71.34	656.692
rpart	85.42	46.11	37.05	92.61	52.92	88.86	35.56	4.478
nnet	91.72	61.76	51.83	91.12	66.07	96.29	73.63	36.903

Table F.80: Experiment 3 - Run 20

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.03	42.18	35.21	82.43	49.35	91.57	53.88	38.232
rf	91.08	59.84	49.77	91.39	64.44	95.26	63.46	63.707
J48	90.94	59.32	49.35	90.96	63.98	93.98	66.85	83.518
bayesglm	84.29	40.87	34.07	82.85	48.28	91.60	53.70	3.789
kknn	91.68	60.33	51.81	86.19	64.72	94.17	64.80	56.890
Adaboost	91.89	61.21	52.52	86.96	65.49	95.13	71.15	623.094
rpart	87.18	49.21	40.04	90.08	55.44	88.97	37.93	3.828
nnet	90.37	57.57	47.69	90.61	62.49	95.29	66.97	38.629

Table F.81: Experiment 3 - Run 21

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.33	43.27	35.92	83.87	50.30	92.27	54.45	39.260
rf	91.64	61.83	51.52	92.43	66.16	95.67	64.12	62.003
J48	90.91	59.45	49.28	91.76	64.12	95.72	66.98	84.100
bayesglm	84.36	41.31	34.30	83.87	48.68	92.25	54.13	3.291
kknn	91.59	60.83	51.44	89.09	65.22	94.53	65.13	55.046
Adaboost	92.29	62.38	54.01	86.43	66.48	95.40	70.44	660.624
rpart	87.26	49.62	40.26	90.88	55.80	89.50	38.37	4.620
nnet	91.54	61.35	51.24	91.71	65.74	96.27	71.72	38.861

Table F.82: Experiment 3 - Run 22

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.37	43.11	35.89	83.17	50.14	91.85	53.61	39.104
rf	91.32	60.68	50.51	91.81	65.17	95.36	63.43	63.347
J48	91.08	59.98	49.77	91.92	64.57	95.28	62.77	84.186
bayesglm	84.50	41.42	34.46	83.33	48.76	91.89	53.45	3.702
kknn	91.72	60.59	51.91	86.75	64.95	94.30	65.02	55.304
Adaboost	91.65	60.79	51.66	88.24	65.17	95.19	69.63	648.385
rpart	85.63	46.52	37.39	92.51	53.26	88.92	35.86	4.697
nnet	91.08	59.96	49.78	91.81	64.55	96.31	74.89	38.224

Table F.83: Experiment 3 - Run 23

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.22	42.74	35.62	82.93	49.83	91.83	53.11	37.719
rf	91.64	61.65	51.57	91.63	65.99	95.85	70.78	62.130
J48	91.83	61.87	52.21	90.24	66.15	95.85	68.96	83.613
bayesglm	84.31	40.91	34.10	82.88	48.32	91.84	52.90	3.392
kknn	91.81	60.92	52.25	86.83	65.24	94.61	65.23	55.483
Adaboost	92.01	61.49	52.98	86.53	65.72	95.47	72.28	658.774
rpart	85.42	46.16	37.07	92.75	52.97	88.96	35.61	4.477
nnet	91.29	60.44	50.43	91.23	64.95	96.33	74.83	37.814

Table F.84: Experiment 3 - Run 24

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.62	41.81	34.71	83.79	49.09	91.94	53.12	37.274
rf	91.75	61.89	51.92	91.28	66.19	95.85	71.14	62.633
J48	91.38	60.58	50.72	90.67	65.05	95.05	63.89	81.826
bayesglm	84.67	41.88	34.78	83.71	49.14	91.94	52.97	3.174
kknn	91.85	61.18	52.35	87.39	65.47	94.45	64.73	55.845
Adaboost	91.94	61.56	52.68	87.65	65.81	95.25	71.07	631.323
rpart	87.02	48.89	39.73	90.27	55.17	89.04	37.71	4.224
nnet	91.31	60.61	50.50	91.60	65.11	96.07	70.94	37.777

Table F.85: Experiment 3 - Run 25

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	83.96	40.74	33.76	84.45	48.23	92.22	54.79	40.557
rf	91.42	61.04	50.85	91.95	65.48	95.46	64.01	62.581
J48	91.51	61.11	51.14	91.15	65.52	94.01	64.84	83.606
bayesglm	83.96	40.75	33.76	84.48	48.24	92.12	54.33	3.095
kknn	91.75	61.03	52.00	87.97	65.37	94.59	65.73	55.790
Adaboost	91.86	61.70	52.37	89.12	65.97	95.50	71.70	653.252
rpart	87.18	49.28	40.05	90.32	55.50	89.17	38.04	4.309
nnet	91.02	59.76	49.59	91.76	64.38	95.71	69.94	37.850

Table F.86: Experiment 3 - Run 26

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.63	41.96	34.78	84.24	49.23	92.22	53.97	38.732
rf	91.70	61.86	51.74	91.81	66.18	95.80	65.45	62.672
J48	91.26	60.22	50.33	90.72	64.74	95.70	70.55	84.573
bayesglm	84.59	41.65	34.62	83.52	48.95	92.20	53.64	3.432
kknn	91.80	60.75	52.23	86.32	65.08	94.55	65.64	55.618
Adaboost	92.43	62.51	54.65	85.07	66.55	95.37	71.76	659.962
rpart	85.39	46.09	37.01	92.77	52.91	88.94	35.56	4.499
nnet	91.04	59.67	49.66	91.15	64.29	95.69	68.04	36.855

Table F.87: Experiment 3 - Run 27

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.13	40.95	33.97	84.11	48.39	92.03	54.61	38.067
rf	91.29	60.73	50.41	92.35	65.22	95.40	63.35	64.858
J48	91.65	61.36	51.61	90.40	65.71	95.14	65.12	83.583
bayesglm	84.25	41.02	34.10	83.57	48.44	92.03	54.26	3.240
kknn	91.52	60.05	51.23	87.04	64.50	94.32	62.84	55.674
Adaboost	91.98	61.54	52.84	87.12	65.78	94.94	67.85	674.714
rpart	87.17	49.31	40.04	90.56	55.53	89.31	38.10	4.228
nnet	90.95	59.49	49.38	91.52	64.15	96.23	72.63	36.919

Table F.88: Experiment 3 - Run 28

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.25	42.83	35.67	83.04	49.91	92.15	53.77	39.587
rf	91.37	60.90	50.68	92.00	65.36	95.64	65.03	64.137
J48	91.33	60.72	50.55	91.81	65.20	93.55	55.23	84.828
bayesglm	84.35	40.99	34.16	82.93	48.39	92.08	53.31	3.125
kknn	91.45	59.88	50.97	87.31	64.37	94.33	63.48	54.539
Adaboost	92.08	61.71	53.22	86.53	65.91	95.36	70.59	674.573
rpart	85.56	46.48	37.31	92.91	53.24	89.14	35.88	4.552
nnet	90.98	59.56	49.47	91.44	64.21	95.56	66.66	37.089

Table F.89: Experiment 3 - Run 29

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	84.36	41.29	34.29	83.76	48.66	92.05	54.15	38.686
rf	91.73	61.89	51.85	91.49	66.19	95.67	65.38	63.970
J48	90.89	59.42	49.20	91.95	64.10	95.07	60.52	84.224
bayesglm	84.40	41.20	34.29	83.23	48.57	92.04	53.98	3.284
kknn	91.47	59.91	51.05	87.17	64.39	94.25	64.99	55.467
Adaboost	91.95	61.37	52.75	86.80	65.62	95.25	71.53	662.318
rpart	85.63	46.56	37.40	92.69	53.30	89.08	35.92	4.437
nnet	91.78	61.95	52.02	91.12	66.23	96.21	69.14	36.598

Table F.90: Experiment 3 - Run 30

Machine learning algorithm	Accuracy (%)	Kappa (%)	Precision (%)	Recall (%)	F1 score (%)	ROC-AUC (%)	PR-AUC (%)	Cost (seconds)
svmLinear	85.76	43.95	36.60	83.20	50.84	92.19	55.52	38.488
rf	91.76	61.89	51.96	91.15	66.19	95.44	70.54	63.015
J48	91.72	61.71	51.84	90.91	66.03	94.32	52.39	83.668
bayesglm	84.72	42.03	34.89	83.84	49.27	92.22	55.70	3.594
kknn	92.00	61.45	52.93	86.53	65.68	94.27	64.56	55.537
Adaboost	92.60	63.06	55.31	85.01	67.02	95.47	72.38	647.419
rpart	85.48	46.15	37.12	92.35	52.95	88.82	35.58	4.561
nnet	91.30	60.36	50.47	90.77	64.87	96.53	77.24	35.809