Feature selection using Benford's Law to support detection of malicious social media bots

Innocent Mbona[1] and Jan H.P. Eloff[1]

[1]Department of Computer Science, University of Pretoria, South Africa.

Corresponding author: Innocent Mbona (e-mail: u15256422@tuks.co.za)

ABSTRACT The increased amount of high-dimensional imbalanced data in online social networks challenges existing feature selection methods. Although feature selection methods such as principal component analysis (PCA) are effective for solving high-dimensional imbalanced data problems, they can be computationally expensive. Hence, an effortless approach for identifying meaningful features that are indicative of anomalous behaviour between humans and malicious bots is presented herein. The most recent Twitter dataset that encompasses the behaviour of various types of malicious bots (including fake followers, retweet spam, fake advertisements, and traditional spambots) is used to understand the behavioural traits of such bots. The approach is based on Benford's law for predicting the frequency distribution of significant leading digits. This study demonstrates that features closely obey Benford's law on a human dataset, whereas the same features violate Benford's law on a malicious bot dataset. Finally, it is demonstrated that the features identified by Benford's law are consistent with those identified via PCA and the ensemble random forest method on the same datasets. This study contributes to the intelligent detection of malicious bots such that their malicious activities, such as the dissemination of spam, can be minimised.

INDEX TERMS Benford's law; high-dimensional imbalanced dataset; malicious bots; feature selection; online social network

I. INTRODUCTION

Online social networks (OSNs) are web service platforms that enable users to interact virtually in real time through posting information and sending messages. This interaction (which is associated with the number of posts) enriches OSNs with large data volumes of end-user behaviour [1, 2]. OSNs are open platforms because anyone can create an account without being subjected to intensive verification processes [3]. In other words, OSNs implement simple authentication methods such as one-time passcodes and CAPTCHA methods [4] to prevent access to non-human users. However, these methods can be circumvented; hence, OSNs are being targeted for malicious activities such as the spread of fake news [47], trolling [1] [5], and Sybil attacks [2] [6], all of which impose serious threats. OSN users are not limited to humans; in fact, social media bots (bots) are widely prevalent [48].Half-human and half-bot accounts exist between humans and bots, and they are known as cyborg accounts [8, 9]. Bots can be used for not only legitimate purposes, such as news or weather updates, but also for malicious activities [10, 48]. According to the authors of [7], several malicious bots have been identified as engaging in the dissemination of misinformation on Instagram by posting false negative comments through *hashtag hijacking*. Spreading fake news on OSNs is a serious concern, as such activities can agitate users as well as influence public opinion [47]. Digital marketing is a vital strategy for many businesses and political parties; therefore, bots that engage in trolling activities to influence public opinion must be identified. The detection of malicious bots on OSNs is a well-known problem that has been investigated extensively by researchers, who subsequently proposed machine and deep learning models for detecting such bots [48]. When developing a predictive machine-learning model, a set of features that may include the daily average number of posts for an account is considered. It is crucial

[1] Trolling is an act – either by humans or bots – of distributing inflammatory content on the Internet.

[2] A Sybil attack is the use of multiple fake identities to control a substantial portion of online networks.

to consider the optimum set of features when designing a machine-learning model [5]. For example, consider a Twitter user account with many attributes, such as user-id, screen name, and location. These attributes are used to create features such as the screen name length [32]; subsequently, these attributes and features are used to design effective machine-learning models that can differentiate between malicious bot and human accounts [48]. The most well-known machine-learning-based bot detection tool is the *botometer* [33] tool. The high-dimensional and velocity aspects of OSN data necessitate an effortless feature selection method that can support machine-learning-based models for malicious bot detection [11, 12]. The majority of feature selection methods are embedded or based on machine learning, such as support vector machine (SVM) [13], neural networks (NN) [12], and ensemble methods [45]. However, embedded feature selection methods are not effective for solving binary classification problems involving an imbalanced dataset [13], such as the case study reported herein. Although machine-learning-based feature selection methods such as those mentioned above and PCA can be used to solve high-dimensional imbalanced dataset problems, their computational cost can be high [14, 46].

Hence, an effortless approach to identify meaningful features that can differentiate between human and malicious bot accounts is proposed herein. A feature is only meaningful if it is indicative of anomalous behaviour between human and malicious bot accounts [16, 23]; for example, malicious bot accounts may post content more frequently than humans. To identify meaningful features, we adopted Benford's law, which states that the distribution of the first significant leading digit (FSLD) on a "naturally occurring" dataset is non-uniform [15]. The meaning of a naturally occurring dataset, as opposed to a fabricated or inflated dataset [17, 35], is discussed later herein. For example, consider a feature constituting a Twitter dataset known as *status_count*, which counts the number of tweets an account contains at a particular discrete time *t*. Let us consider that 1000 Twitter users are selected randomly and the status_count of each user is examined. The first user may have **3**24 tweets, the second user may have **8**7 tweets, and so on, until the 1000th user. The following question arises: What is the distribution of the FSLD (in bold) for all 1000 users, or what is the likelihood that the FSLD for status_count begins with digits 1, 2,.., 9? Digit 1 is expected to occur approximately 30% more frequently as an FSLD than digit 9 [16, 44]. The same logic can be applied to examine the FSLD distribution of other numeric Twitter-based features, such as followers_count, which will be discussed in Sections I and IV. Benford's law can be implemented easily and does not require parameter fitting. Hence, it is superior to other non-uniform distributions, such as the power law and Zipf's law [17].

Research questions and objectives

(i) What are the behavioural traits of malicious bots and humans using a Twitter dataset?

(ii) Investigate features based on Twitter attributes to determine if they obey Benford's law.

(iii) Demonstrate that Benford's law can effectively identify meaningful features that can differentiate malicious bots from humans, even on a high-dimensional imbalanced dataset.

(iv) Demonstrate that features identified by Benford's law are consistent with prevalent feature selection methods, which include the PCA and the ensemble random forest, on the same Twitter datasets.

II. LITERATURE REVIEW

In this section, we discuss key literature pertaining to bot detection, feature selection, and Benford's law.

The authors of [16] were the first to apply Benford's law to data on OSNs[16]; they discovered that certain user features on OSNs conformed to Benford's law. Furthermore, they reported that Benford's law can be used to detect users who display anomalous behaviour (some of these were discovered to be bots). Datasets were extracted from different OSNs (including Facebook and Twitter) to demonstrate that the distribution of FSLDs of the number of friends, followers, and posts obeyed Benford's law. Subsequently, the authors [16] performed further investigations to demonstrate that Benford's law can be used to detect malicious bots using the Russian botnet (Twitter dataset), retweet bots, and 'like' fraud bots on Facebook [44]. The authors of [44] demonstrated that the abovementioned bots consistently violated the FSLD distribution of the friends_count feature, whereas humans obeyed it. It was concluded that the friend_count feature is significant for differentiating between malicious bots and humans. The main limitation of the study reported in [44] was the small sample size of retweets and fraud bots. Although the two studies [16, 44] mentioned above highlighted key valuable insights into using Benford's law on OSNs, the following question arises: Is there a set of features that can differentiate between human and malicious bot datasets in general? Furthermore, both [16, 44] studies showed that a three features were significant in differentiating between human and malicious bot datasets. Research shows that a prediction model with few features (such as three) on a high-dimensional imbalanced dataset can be affected by bias [18]. The primary objectives of this study are as follows: (i) Investigate various attributes and features on Twitter to determine whether they can be used to differentiate between human and malicious bots in general; (ii) demonstrate that Benford's law can effortlessly and consistently identify anomalous behaviour of different types of malicious bots, and (iii) demonstrate that features identified by Benford's law are consistent with prevalent feature selection methods.

Numerous feature selection methods exist, including filter, wrapper, and embedded methods [19]. Filter methods use a single feature (e.g. status_count) to determine the predictive power of a model. Wrapper methods are similar to filter methods, except that they use a combination of features (e.g. status_count and friends_count) to determine the predictive power of a model. Embedded methods use regression methods to obtain an optimal subset of features (e.g. status_count, friends_count, and followers_count) through repeated learning steps. The feature selection methods above perform effectively on balanced datasets (e.g. an equal number of human and malicious bot accounts) [12]. In cases where the dataset is high-dimensional and imbalanced (e.g. fewer malicious bots compared with humans), advanced techniques such as NN, SVM, and PCA are required [11]. The abovementioned methods partition both the majority (human) and minority (malicious bot) datasets into small subsets to obtain the optimum subset of features that can differentiate malicious bots from humans [11, 13, 20–21]. In this study, we compared features identified using Benford's law with those identified using PCA and the ensemble random forest to differentiate human and malicious bot datasets. Both PCA and the ensemble random forest are decent feature selection methods for high-dimensional imbalanced datasets such as OSNs [22,44], despite being susceptible to high computational costs. Next, we briefly discuss machine-learning models and significant features that have been proposed previously to detect bots on OSNs; their summaries are available in [3, 23-25].

Automation and bot detection

Chu et al. [8, 9] designed a machine-learning-based system that classifies Twitter users into three distinct groups: human, bot, and cyborg. Their model is based on entropy, spam detection, and account information. The account information provides account-related information, e.g. the tweeting device used. The entropy feature is used to determine the tweeting time frequency of an account. The model of Chu et al. [8, 9] claims that bot accounts are expected to exhibit a consistent tweeting pattern, whereas humans exhibit spontaneous behaviours. The spam feature is used to determine unsolicited content in the tweets. Such spam detection is based on a predefined set of words, e.g. 'cash, prizes, win' and so on, that are known to be spam. The challenge in spam detection is that spammers often use dynamic words that are difficult to detect. Moreover, extracting important features for spam detection can be computationally expensive [26, 27]; hence, we did not consider the cost in this study. Chu et al. [8, 9] applied machine-learning models and discovered that human accounts tended to interact with other human accounts via their tweets, retweets, mentions, hashtags, and direct messaging more often than bots. The current study demonstrates that such behaviours of humans and bots can be discovered using a simpler method and by applying Benford's law.

Chu et al. [8, 9] aggregated features from entropy, spam detection, and account information, and then used the random forest algorithm to classify Twitter users into three classes: human, bot, and cyborg. The authors observed that more than 53% of their dataset accounts belonged to humans, 36% to cyborgs, and 11% to bots. Their results further indicated that bot accounts were less than human accounts in a Twitter dataset. In general, their classification system successfully distinguished human accounts from bot accounts reasonably well. The entropy, URL ratio, and tweeting device features contributed the most to the final classification. The results of Chu et al. [8, 9] pertaining to the differentiation of bot accounts from cyborg and human accounts were mediocre. This might be because cyborgs can be assisted by either human or bots.

Dickerson et al. [29] applied sentiment features to classify human and bot users using a Twitter dataset. Their classification model relies on sentiment analysis from tweet syntax, tweet sentiment, user behaviour, and network properties. They successfully classified bots from humans reasonably well. Since the majority of the sentimental features were not real number values, we could not apply Benford's law directly to this type of dataset. The process of transforming non-numerical data or features into a numerical dataset requires further work that is beyond the scope of this study. In their attempt to classify Twitter followers as either human, bot, or neutral, Cresci et al. [30] verified 1950 human accounts and 1950 bot accounts. In total, their human dataset contained 2,631,730 tweets, 1,785,438 followers, and 908,935 friends. Meanwhile, the bot dataset contained 118,327 tweets, 34,553 followers, and 879,580 friends. Cresci et al. [30] used 49 distinct features in eight machine-learning models. They indicated that features describing account relationships (i.e. their neighbours) were key in classifying fake or real followers. The limitation of their methodology is that an account under investigation must fulfil several criteria, e.g. it must contain at least 30 followers and at least 50 tweets. The features used by Cresci et al. [30] were investigated in this study to determine whether they can differentiate between human and malicious bots using Benford's law. Cresci et al. [30] mined 3474 genuine human accounts on Twitter, 991 social spambots #1 (retweet spam), 3457 social spambots #2 (hashtag spam), social spambots #3 (advertisement spam), 1000 traditional spambots (evolving Twitter spambots), and 3351 fake followers (accounts that inflate the follower count). We used this dataset from Cresci et al. [31] in our experiments because it is the most recent and encompasses the behaviours of various types of malicious bot accounts (see Table 2). Cresci et al. [31] analysed the percentage of accounts that were still active, deleted, and suspended, separately, and discovered that a significant proportion of spambots remained active at the time of their investigation. Moreover, Cresci et al. [31] applied the datasets above to different well-known techniques for bot detection; however, their results suggested that bot detection was challenging. Hence, the findings of Cresci et al. [31] serve as motivation for the current study. In contrast to the investigations performed by the authors of [32] and [30], Ferrara et al. [10, 28] classified Twitter users as humans or bots by extracting more than *1150* features from a network (which described users' distribution of tweets, retweets, mentions, and hashtags), the user (account-related information), friends (number of friends and followers), temporal features (describing tweet timing patterns), content features (linguistic behaviour), and sentiment features (sentimental analysis and emoticon scores). The high number of features, i.e. *1150*, indicates that the OSN data are highly dimensional. Ferrara et al. [10, 28] applied a random forest model based on these features to successfully classify bot and human accounts. Ferrara et al.'s [28] tool is available for free online [33]. In the next section, we discuss features that were used in previous studies, including those used by Ferrara et al. [28], to distinguish bot accounts from human accounts; additionally, we discuss whether they obey Benford's law.

Previous features used in bot detection

The features shown in Table 1, which were used in previous studies, successfully detected bots. These features indicate the behavioural traits of users; for example, status_count indicates the active level of a user, and features such as followers_count and friend_count indicate the level of popularity on OSNs. In the current study, we used Benford's law to identify features and attributes that were indicative of anomalous behaviour between humans and malicious bots. For example, if a feature x exists that obeys Benford's law on the human dataset, whereas the same feature x violates Benford's law on the malicious bot dataset, then feature x is regarded as a good differentiator of malicious bot and human datasets [16, 44]. If a feature violates (or does not violate) Benford's law on both malicious bot and human datasets, then that feature is regarded as a good differentiator as it fails to separate the datasets.

Feature	**Description**	**Source**
Screen_name length	Number of characters in screen name.	[32]
Status_count	Number of tweets.	[16, 30]
Followers_count	Number of followers.	[16, 31]
Friends_count	Number of friends.	[16, 34]
Favourite_count (likes)	Number of tweets liked by user.	[28]
Listed_count	Number of groups (lists) subscribed by user.	[28, 32]
Re-tweet_count	Number of retweets.	[26, 34]
Reply_count	Number of replies to a tweet.	[31, 33]
Hashtag_count	Number of tweets and retweets for a specified # key phrase.	[31]
URL_count	Number of URL addresses in user profile.	[28]
Mention_count	Number of times an account is mentioned.	[31, 33]

Table 1. Sample of features used previously to detect Twitter malicious bots.

III. BENFORD'S LAW

In this section, we describe Benford's law and its dataset conditions. Additionally, we demonstrate empirically that the FSLD distribution of a Twitter dataset should conform to Benford's law. For this study, it suffices to demonstrate Benford's law empirically, since we are addressing the application of this law. A mathematical proof of Benford's law is available in [15, 17]. Benford's law was discovered in 1881 by the astronomer Simon Newcomb; since then, it has been applied in various areas, including forensic accounting [17], fraud detection [15], and OSNs [16, 44].

A Twitter dataset and Benford's dataset conditions

Not all datasets are expected to obey Benford's law [17, 35]. Herein, we discuss the cases in which a dataset is expected to obey or disobey Benford's law. We highlight the general conditions and demonstrate that the Twitter dataset satisfies those conditions.

• *All leading digits from 1 to 9 must be possible in a dataset*: In the case of Twitter, consider, for example, the followers_count feature. Each account will have a number of followers where FSLDs 1 to 9 are all possible.
• *A dataset should have more small numbers than large numbers*: This has been demonstrated by the authors of [16]. In this study, it appears that the Twitter dataset in fact satisfies this condition. Small numbers, as compared with large numbers, tend to occur more frequently on different features [16].
• *A dataset must occur naturally*: This implies that the number of OSN features is expected to increase naturally as opposed to being inflated. Specifically, bots have been shown to fabricate features such as followers_count through black markets [7] such that their accounts appear more popular.
• *Numbers should not occur sequentially*: Each user on Twitter will have different numbers of friends, followers, etc. that are displayed randomly.
• *Numbers should not have predefined boundaries (i.e. a minimum or maximum, except for zero)*: This applies to Twitter datasets.
• *A dataset must be sufficiently large (typically a sample size above one thousand)*: OSNs such as Twitter contain millions of users; hence, a large dataset of users is available.

Based on the above, it is evident that the Twitter dataset satisfies Benford's law dataset conditions. Hence, the features listed in Table 1 satisfy these conditions.

Property 1

The FSLD. Let D be a positive real number on (Ω, Ƒ, ℙ). The logarithmic density function for the first leading digit is expressed as

$$\mathbb{P}(\boldsymbol{D} = \boldsymbol{d}) = \boldsymbol{log_b}\left(\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{d}}\right), \qquad (\mathbf{1})$$

where b = 10 and d ∈ {1,2,...9}.

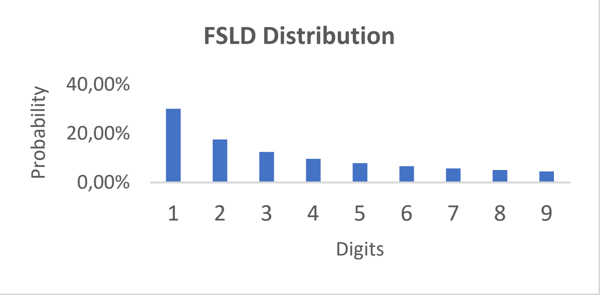


Figure 1. FSLD distribution adapted from [15]

In Figure 1, the FSLD distribution suggests that numbers beginning with smaller digits (1–5) have a higher occurrence probability than the larger digits (6–9), based on computation using Equation (1). In this study, we used the FSLD, as it yields more conservative probabilities. Using the FSLD, we implicitly excluded accounts with zero features (e.g. zero status_count). Because the abovementioned accounts are inactive accounts, one cannot determine whether they are bots or human accounts.

IV. DATASETS AND METHODS

Data preparation

Big data are characterised by their volume, velocity, and variety [36]. Volume describes the amount of data that is being generated, velocity describes the speed at which the data is generated, and variety describes the different types of data [36]. A Twitter dataset is a classic example of big data, as users continuously interact with one another via text messages and live streams to share pictures and videos (among other things). To test the effectiveness of Benford's law in differentiating between human and malicious bot datasets, we used the same features used in previous studies (see summary in Table 1). In addition, we used data from the Bot Repository (https://botometer.osome.iu.edu/bot-repository/index.html), which is a centralised repository for Twitter social bots. Metadata such as status_count are provided for bots and human datasets. The data are summarised in Table 2.

Dataset	Bot type	#Human	#Bot
Cresci-2017 [31]	Traditional spambots, fake followers, retweet spam, and hashtag spam	3474	9391
Gilani-2017 [41]	Spam bots	1413	1090
Botometer-feedback-2019 [42]	Fake followers and spambots	372	134
Pronbots-2019 [42] and verified-2019 [43]	Spam bots	1987	17809

Table 2. Summary of number of malicious bots and humans from Bot Repository

Human vs. malicious bots Benford's law experiments

Each dataset listed in Table 1 contains a minimum of 1000 data points (observed distribution) for each feature. Equation (1) is used to compute the FSLD distribution. A chi-squared test was used to test whether a significant difference existed between the observed and FSLD distributions (see [37] for details of chi-squared test). The goodness-of-fit test was formulated as follows:

Null hypothesis (H_0) = a feature obeys FSLD.
Alternative hypothesis (H_1) = a feature violates FSLD.
If p-value < 0.05, we reject H_0, else we cannot reject H_0.

Chi-squared results of Benford's law

Feature	Human	Retweet spam	Hashtag spam	Fake advertisement	Traditional spambots	Fake followers
Screen_name length	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Status_count	Cannot reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Cannot reject H_0.	Reject H_0.
Followers_count	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Cannot reject H_0.	Reject H_0.
Friends_count	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Favourite_count	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Listed_count	Cannot reject H_0.	Reject H_0.	Reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.
Re-tweet_count	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Reply_count	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Hashtag_count	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
URL_count	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Mention_count	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.

Table 3. Chi-squared results of FSLD test on human and malicious bot datasets [31]

The features in Table 1 were examined to determine whether they can differentiate between malicious bot and human datasets. In our analysis, we compared the FSLD with the observed distribution of each feature using the chi-squared distribution. As the screen_name length feature violated the FSLD on both the human and malicious bot datasets, it was regarded as an inferior differentiator. The screen_name feature of an account does not provide much information or insights regarding whether the account is controlled by a malicious bot. Furthermore, the screen name of an account can have only a limited number of characters (i.e. a limit of 15 characters on Twitter). Similarly, the reply_count, hashtag_count, URL_count, and mention_count features could not differentiate well between malicious bot and human datasets, as they violated the FSLD for both the human and bot datasets. The status_count feature was useful for differentiating between human and malicious bot [31] datasets well, although some bots (retweet and traditional spambots) obeyed the FSLD, similar to humans. This finding was as expected, since some bots can closely mimic their posting frequency to that of humans. Favourite_count, followers_count, friends_count, listed_count, and re-tweet_count were good differentiators of human vs. malicious bot [31] datasets. For these features, we observed that the human datasets closely reflected the FSLD, whereas the malicious bots violated the FSLD. Based on our experiments, Table 4 provides a summary of features that were good indicators, useful indicators, and inferior indicators of human vs. malicious bots based on Benford's law. Spambots tended to have fewer friends, followers, and favourites than humans. Furthermore, spambots tended to have more tweets and retweets than humans (see Appendix). A malicious bot can automate these features.

Good indicator	Bad indicator	Useful indicator
Followers_count	Screen_name length	Status_count
Friends_count	Reply_count	
Favourite_count	Hashtag_count	
Listed_count	URL_count	
Re-tweet_count	Mention_count	

Table 4. Summary results of FSLD using dataset from [31]

Table 4 summarises meaningful features that successfully differentiated human datasets from malicious bot datasets using Benford's law on dataset presented in [31]. Furthermore, we applied Benford's FSLD tests using the various malicious bot datasets listed in Table 2 to demonstrate that Benford's law can consistently identify significant distinguishing features between human and malicious bot datasets. The chi-squared results for the datasets [41–43] are provided in the Appendix.

Outcome	[41] Dataset	[42] Dataset	[42] and [43] Dataset
Good indicator	Friends_count Favourite_count Listed_count Retweet_count Followers_count	Friends_count Favourite_count Listed_count Retweet_count Status_count	Friends_count Favourite_count Listed_count Retweet_count Status_count
Bad indicator	Screen_name length Reply_count URL_count Mention_count Reply_count	Screen_name length Reply_count URL_count Mention_count Reply_count	Screen_name length Reply_count URL_count Mention_count Reply_count
Useful indicator	Status_count	Followers_count	Followers_count

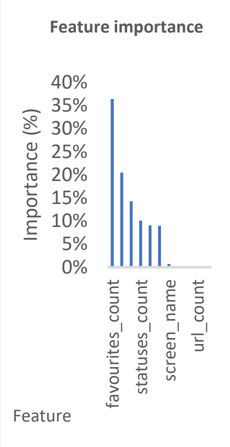
Table 5. Summary of Benford's FSLD test results for different datasets

Next, we applied the ensemble random forest on the same dataset presented in [31] to identify features that can differentiate human and malicious bot datasets and then compared them with Benford's results shown in Table 4.

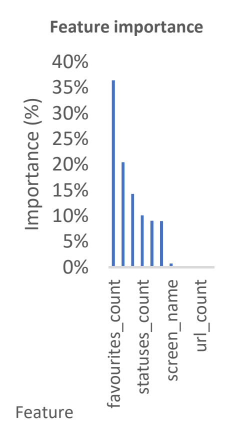
Ensemble random forest

The ensemble random forest feature selection method uses bagging (sample with replacement) or pasting (sampling without replacement) to identify important features [45]. This method measures the importance of a feature by measuring the amount by which a tree node that uses that feature decreases the impurity (*Gini* importance). Ensemble random forests have been shown effective in solving high-dimensional imbalanced data problems such as OSNs, provided that the hierarchical structure allows them to learn from the majority and minority classes [45]. This algorithm, as adapted from Chapter 7 of [45], can be summarised as follows for a training set $X = \{x_1, x_2, .., x_n\}$ and output $Y = \{y_1, y_2, .., y_n\}$. Bagging or pasting sampling is performed, and various decision tree methods are applied to aggregate the outputs and identify important features. The Python code for this model is described in Chapter 7 of [45]. The ensemble random forest results indicate that favourite_count, followers_count, friends_count, listed_count, retweet_count, and status_count are important features, which is consistent with Benford's results shown in Table 4 further results for other datasets are provided in the Appendix.

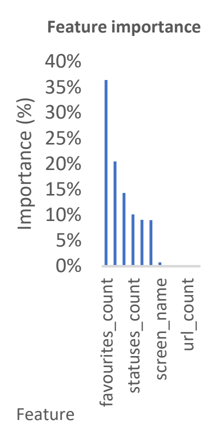
Human vs. Retweet spam



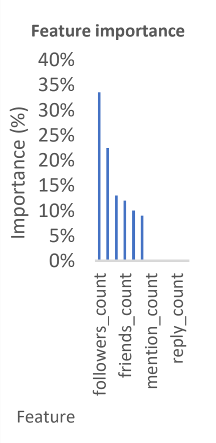
Human vs. Hashtag spam



Human vs Fake adv spam



Human vs. Traditional spam



Human vs. Fake followers

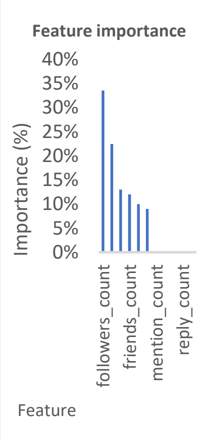


Table 6. Feature importance using ensemble random forest method on dataset from [31]

PCA

PCA is a technique that uses orthogonal transformation to convert a high-dimensional correlated variable into a low-dimensional uncorrelated variable [38]. It can effectively reduce the number of dimensions in a dataset with minimal information loss. Since we will omit the mathematical derivation of PCA herein, readers are referred to [38, 39]. The procedures pertaining to PCA can be summarised as follows: (i) Obtain data with high dimensions; (ii) centre and scale the data; (iii) calculate the covariance matrix; (iv) compute the eigenvectors and eigenvalues of the covariance matrix; (v) select the components that represent the majority of the variance of the data. The PCA code from [40] was adopted. The data used in the PCA are summarised in Table 7.

Feature	Human	Retweet spam	Hashtag spam	Fake advertisement	Traditional spam	Fake followers
Followers	4840045	1768843	80818	1154924	637297	59448
Friends	2199884	1837085	187496	872451	1326542	1240070
Favourite	16222261	156748	32923	8190	4328	14408
Listed	67731	3919	1031	5871	7900	245
Retweet	529398514	4592526	8569207	740039	112	126340097
Status	58912857	1101991	868982	5399138	220829	240931
Hashtag	148299	91766	48975	47707	20351	46384
Mention	519480	21941	174380	27571	46807	69958
Reply	109	67	34	75	111	189063
ScreenName	37789	12568	42369	5485	11260	39722
URL	111568	49968	18156	485402	97939	40372

Table 7. Summary of data from [31] used in PCA

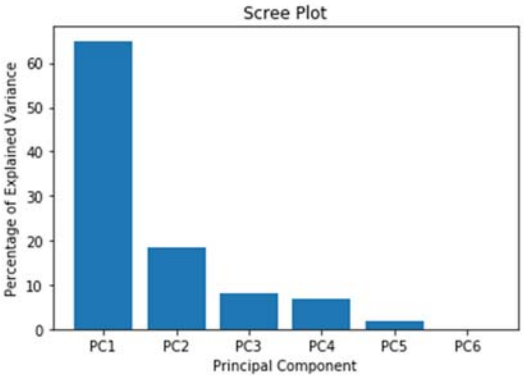


Figure 2. PCA scree plot

Figure 2 shows the principal components for separating human and malicious bot datasets. It is clear that PC1, PC2, and PC3 constituted most of the variations in the data (98.3%). The variation ratio is expressed as follows:

$$Variation\ ratio\ = \frac{local\ variation}{global\ variation}$$

$$VR\ PC1\ = \frac{65}{100}\ = \ 65\%.$$

$$VR\ PC2\ = \frac{18.3}{100}\ = \ 18.3\%.$$

More than 80% of the total variation in the data can be represented by PC1 and PC2 in a two-dimensional (2-D) graph.

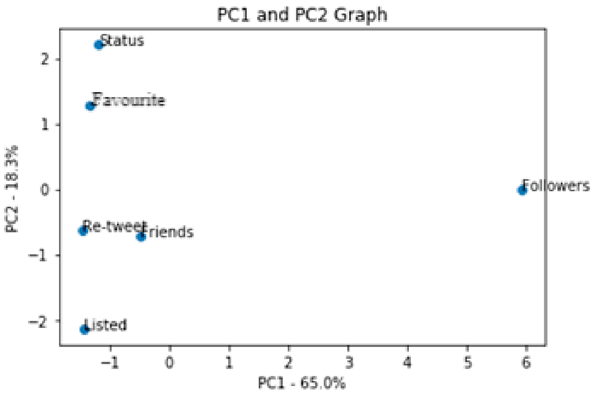
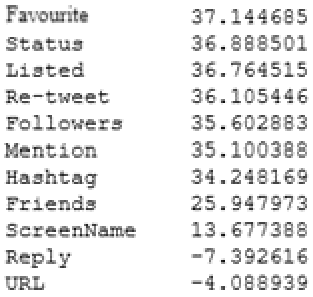


Figure 3. PC1 and PC2 in 2-D space

Figure 3 represents PC1 and PC2 in a 2-D space. The features highlighted in the figure were significant in differentiating between human and malicious bot datasets. The figure suggests that favourite_count, listed_count, status_count, re-tweet_count, and friends_count are positively correlated as compared with followers_count. Finally, we computed the loading scores to examine the features that contributed significantly to the differentiation between human and malicious bot datasets.



```
Favourite       37.144685
Status          36.888501
Listed          36.764515
Re-tweet        36.105446
Followers       35.602883
Mention         35.100388
Hashtag         34.248169
Friends         25.947973
ScreenName      13.677388
Reply           -7.392616
URL             -4.088939
```

Figure 4. PCA feature selection and loading scores.

Figure 4 shows PCA loading scores, which reflect ranked features that are significant to the differentiation between human and malicious bot datasets. Comparing the results in Figure 4 with results obtained Benford's law shown in Table 4, we observed that the meaningful features identified via Benford's law were consistent with those identified via PCA. However, both Benford's law and the PCA did not regard screenname length, reply_count, and URL_count as significant features for distinguishing between human and malicious bot datasets. These results appeared to be aligned with reality; for example, the screen_name length and URL_count of an account did not provide much insight in terms of the automation of an account or bot behaviour. The PCA test results for the other datasets are provided in the Appendix.

V. DISCUSSION OF RESULTS

Benford's law was applied to a real Twitter dataset from the Bot Repository, which comprises data of humans and various types of malicious bots. We experimentally identified features that can differentiate between malicious bot and human datasets using the FSLD test. It was discovered that Benford's law only focused on the FSLD distribution; therefore, it was not affected by imbalanced datasets. The results indicated that screen_name length, reply_count, hashtag_count, URL_count, and mention_count were not good features for distinguishing between humans and malicious bots, as they violated Benford's law for both the human and bot datasets (see Table 2). Furthermore, we discovered that humans and malicious bots exhibited similar behaviours on these features. The favourite_count, friends_count, re-tweet_count, and listed_count features consistently obeyed Benford's law for the human dataset, whereas they violated Benford's law for malicious bot datasets.

Hence, the features mentioned above were regarded as good differentiators of human and malicious bot accounts. Although the status_count and followers_count features were useful, some bots in different datasets obeyed Benford's law, similar to humans. Additionally, we compared the features identified using Benford's law with those identified via PCA and the ensemble random forest on the same datasets. The results indicated that the features identified using Benford's law were consistent with those identified via PCA and the ensemble random forest.

It is noteworthy that we do not claim that Benford's law outperformed these methods; however, we demonstrated that Benford's law was much simpler to implement as compared with these methods. Therefore, it will likely reduce the computational cost required to identify significant features on high-dimensional big data such as OSNs. Because we used metadata from Twitter attributes (e.g. friend_count and favourite_count), the findings of this study can be applied to other OSNs such as Facebook, LinkedIn, and YouTube. However, Benford's law is only applicable to numeric datasets; therefore, further studies should be conducted to address this limitation.

VI. CONCLUSIONS AND IMPLICATIONS

Machine-learning models can only benefit from intelligent feature selection techniques, particularly on high-velocity and dimensional imbalanced datasets. We believe that this study contributes to the further development of intelligent feature selection methods for designing effective machine-learning models that can detect malicious bots. Herein, a simplified feature selection method for a high-dimensional imbalanced Twitter dataset using Benford's law was proposed. The findings of this study contribute significantly to not only the field of feature selection, but also to other interdisciplinary fields such as cybersecurity. The ability to detect malicious bots effectively can minimise the dissemination of misinformation and improve the integrity of OSNs. Further research is required to design an automated machine-learning-based system that can detect malicious bots by leveraging the findings of this study.

ACKNOWLEDGEMENTS

Innocent Mbona would like to thank the University of Pretoria and Bank Seta for funding this study.

VII. REFERENCES

[1] M. Tsikerdekis and S. Zeadally, "Online deception in social media," *Communications of the ACM,* vol. 57, pp. 72-80, 2014.

[2] M. Tsikerdekis and S. Zeadally, "Detecting and preventing online identity deception in social networking services," *IEEE Internet Computing,* vol. 19, pp. 41-49, 2015.

[3] D. B. Kurka, A. Godoy, and F. J. Von Zuben, "Online social network analysis: A survey of research applications in computer science," *arXiv preprint arXiv:1504.05655,* 2015.

[4] X. Xu, L. Liu, and B. Li, "A survey of CAPTCHA technologies to distinguish between human and computer," *Neurocomputing*, vol. 408, pp. 292–307, 2020.

[5] P. Galán-García, J. G. d. l. Puerta, C. L. Gómez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying," *Logic Journal of the IGPL,* vol. 24, pp. 42-53, 2016.

[6] M. Al-Qurishi, M. Al-Rakhami, A. Alamri, M. Alrubaian, S. M. M. Rahman, and M. S. Hossain, "Sybil defense techniques in online social networks: a survey," *IEEE Access,* vol. 5, pp. 1200-1219, 2017.

[7] F. C. Akyon and M. E. Kalfaoglu, "Instagram Fake and Automated Account Detection," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, 2019, pp. 1-7.

[8] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?," *IEEE Transactions on Dependable and Secure Computing,* vol. 9, pp. 811-824, 2012.

[9] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: human, bot, or cyborg?," in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 21-30.

[10] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman*, et al.*, "The DARPA Twitter bot challenge," *Computer,* vol. 49, pp. 38-46, 2016.

[11] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition,* vol. 58, pp. 121-134, 2016.

[12] L. Yin, Y. Ge, K. Xiao, X. Wang, and X. Quan, "Feature selection for high-dimensional imbalanced data," *Neurocomputing,* vol. 105, pp. 3-11, 2013.

[13] S. Maldonado, R. Weber, and F. Famili, "Feature selection for high-dimensional class-imbalanced data sets using Support Vector Machines," *Information sciences,* vol. 286, pp. 228-246, 2014.

[14] S. García, J. Luengo, and F. Herrera, "Feature selection," *Intell. Syst. Ref. Libr.*, vol. 72, no. 6, pp. 163–193, 2015.

[15] A. Berger and T. P. Hill, "An introduction to Benford's law": *Princeton University Press, 2015.*

[16] J. Golbeck, "Benford's law applies to online social networks," *PloS one,* vol. 10, p. e0135169, 2015.

[17] M. J. Nigrini, "Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations". *John Wiley & Sons*, 2020.

[18] V. Chauhan, A. Pilaniya, V. Middha, A. Gupta, U. Bana, B. R. Prasad*, et al.*, "Anomalous behavior detection in social networking," in *2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2017, pp. 1-5.

[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, pp. 16-28, 2014.

[20] X. Tao, Q. Li, W. Guo, C. Ren, C. Li, R. Liu*, et al.*, "Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification," *Information Sciences,* vol. 487, pp. 31-56, 2019.

[21] R.Sudharsan. "Hands-on reinforcement learning with Python: master reinforcement and deep reinforcement learning using OpenAI gym and tensorFlow". *Packt Publishing Ltd*, 2018.

[22] M. Morchid, R. Dufour, P.-M. Bousquet, G. Linares, and J.-M. Torres-Moreno, "Feature selection using principal component analysis for massive retweet detection," *Pattern Recognition Letters,* vol. 49, pp. 33-39, 2014.

[23] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang, "Anomaly detection in online social networks," *Social Networks,* vol. 39, pp. 62-70, 2014.

[24] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery,* vol. 29, pp. 626-688, 2015.

[25] R. Yu, H. Qiu, Z. Wen, C. Lin, and Y. Liu, "A Survey on Social Media Anomaly Detection," *ACM SIGKDD Explor. Newsl.*, vol. 18, no. 1, pp. 1–14, 2016.

[26] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing,* vol. 159, no. 1, pp. 27–34, 2015.

[27] A. Talha and R. Kara, "A Survey of Spam Detection Methods on Twitter," *Int. J. Adv. Comput. Sci. Appl.,* vol. 8, no. 3, pp. 29–38, 2017.

[28] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in *Eleventh international AAAI conference on web and social media*, 2017.

[29] J. P. Dickerson, V. Kagan, and V. Subrahmanian, "Using sentiment to detect bots on Twitter: Are humans more opinionated than bots?," in *2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*, 2014, pp. 620-627.

[30] 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.

[31] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 963-972.

[32] E. Van Der Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access,* vol. 6, pp. 6540-6549, 2018.

[33] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59.7, pp. 96–104, 2016.

[34] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach," in *Proceedings of the 2015 international conference on social media & society*, 2015, pp. 1-7.

[35] E. Druică, B. Oancea, and C. Vâlsan, "Benford's law and the limits of digit analysis," *International Journal of Accounting Information Systems,* vol. 31, pp. 75-82, 2018.

[36] M. Kumar and A. Bala, "Analyzing Twitter sentiments through big data," *Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain.*

Glob. Dev. INDIACom 2016, pp. 2628–2631, 2016.

[37] S. Afanasiev and A. Smirnova, "Predictive fraud analytics: B-tests," *Journal of Operational Risk, Forthcoming,* 2018.

[38] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences,* vol. 374, p. 20150202, 2016.

[39] M. Narasimha, and V. Susheela Devi. "Introduction to pattern recognition and machine learning". Vol. 5. *World Scientific*, 2015.

[40] SKLearn. (2020, 01June2020). *Sklearn.decomposition.PCA*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[41] G.Zafar, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. "Of bots and humans (on twitter)." *In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pp. 349-354. 2017.

[42] Y.Kai-Cheng, et al., "Arming the public with artificial intelligence to counter social bots," *Human Behavior and Emerging Technologies 1.1 (2019): 48-61.*

[43] Y.Kai-Cheng, et al., " Scalable and generalizable social bot detection through data selection," *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 34. No. 01.* 2020

[44] J.Golbeck, "Benford's Law can detect malicious social bots," *First Monday (2019). https://journals.uic.edu/ojs/index.php/fm/article/view/10163.* Last accessed : 30 March 2021.

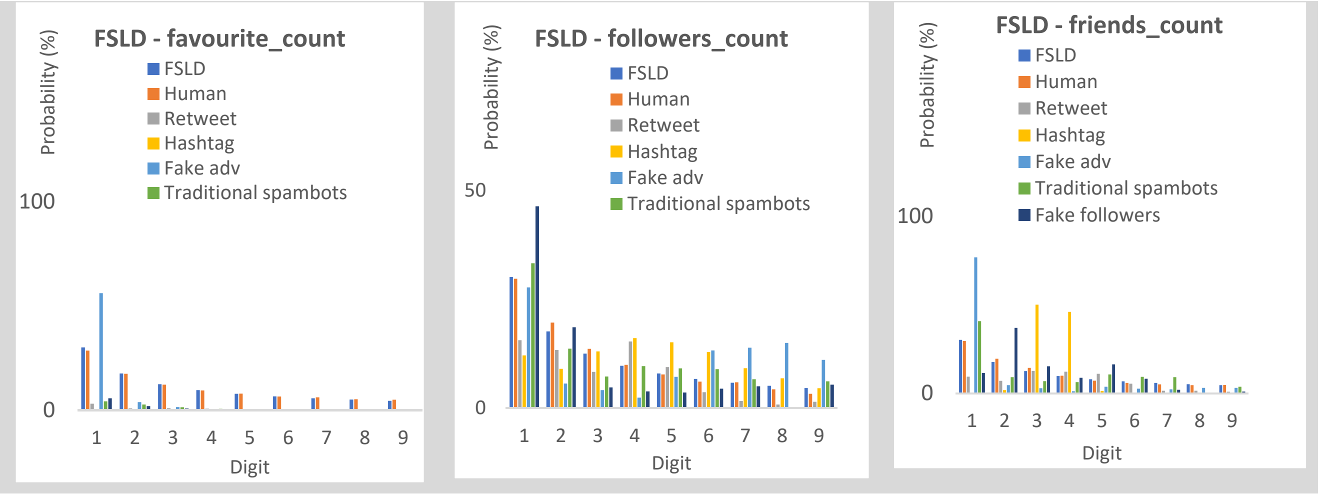
[45] G.Aurélien, "Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems," *O'Reilly Media,* 2019.

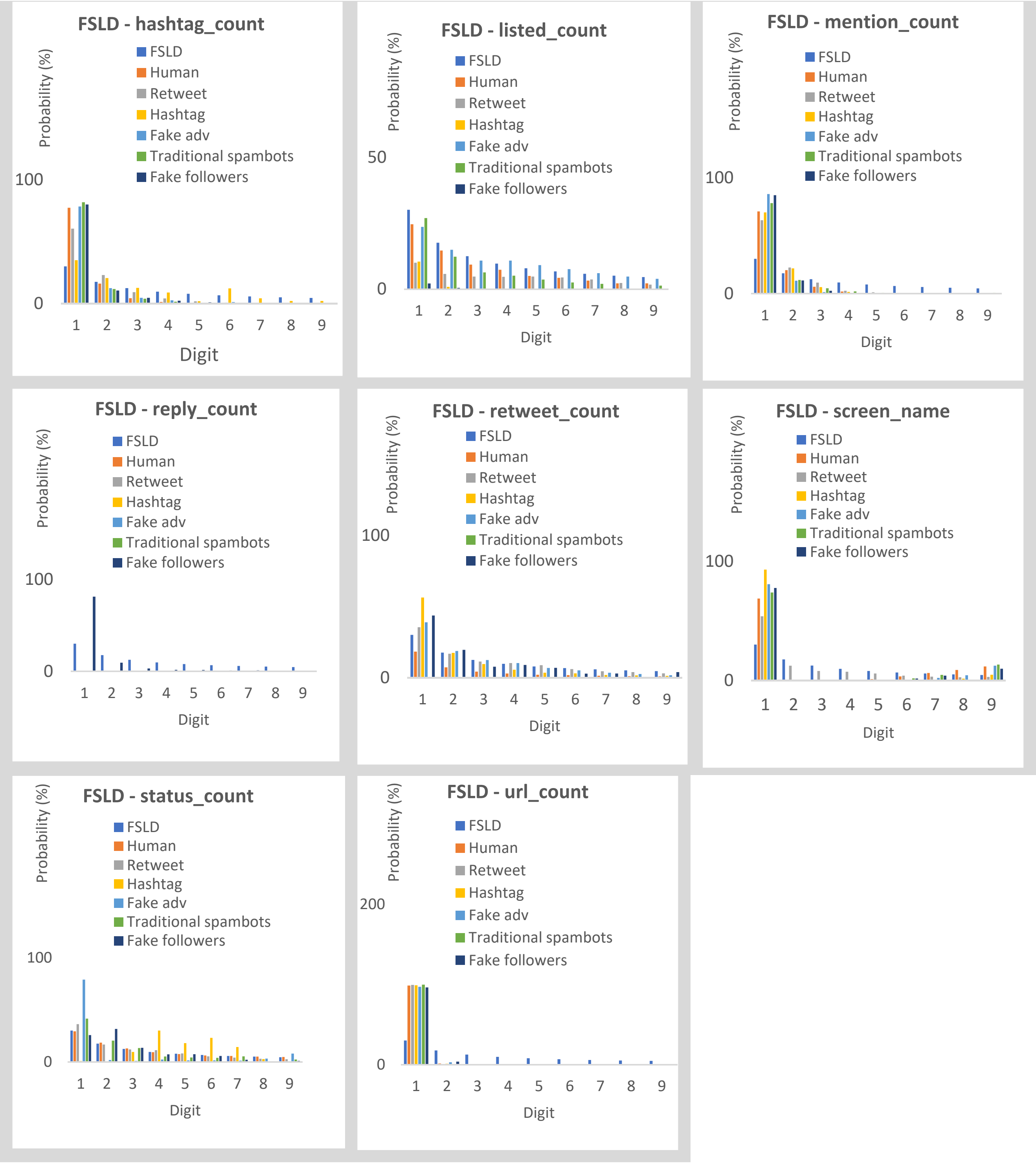
[46] J. Gui, Z. Sun, S. Ji, D. Tao, and T. Tan, "Feature selection based on structured sparsity: a comprehensive study," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 7, pp. 1490–1507, 2017.

[47] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," *Inf. Sci. (Ny).*, vol. 497, pp. 38–55, 2019.

[48] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Syst. Appl.,* vol. 151, p. 113383, 2020

VIII. APPENDIX: FSLD Graphical Representation – Cresci et al. [31]




FSLD - hashtag_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - listed_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
50
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - mention_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - reply_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - retweet_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - screen_name
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - status_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
100
0
1 2 3 4 5 6 7 8 9
Digit
FSLD - url_count
Probability (%)
FSLD
Human
Retweet
Hashtag
Fake adv
Traditional spambots
Fake followers
200
0
1 2 3 4 5 6 7 8 9
Digit

Benford’s FSLD Test results

Table 8. Benford’s law FSLD Experiment Results using [41] dataset

	Screen_name length	Status	Followers	Friends	Favourite	Listed	Retweet	Reply	Hashtag	URL	Mention
Human	Reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Bot	Reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.

Table 9. Benford’s law FSLD Experiment Results using [42] dataset

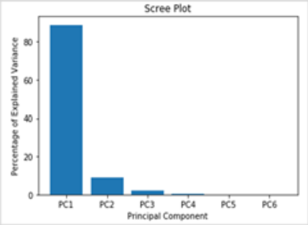
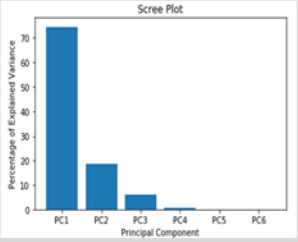
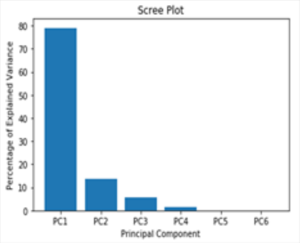
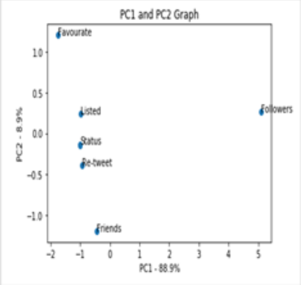
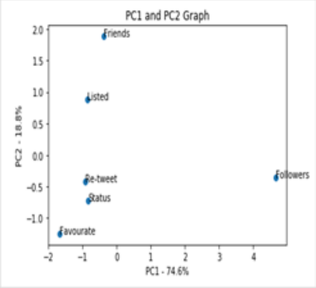
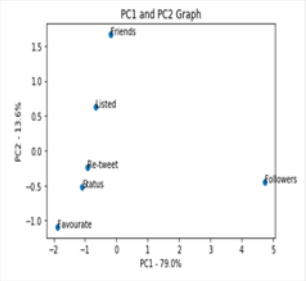
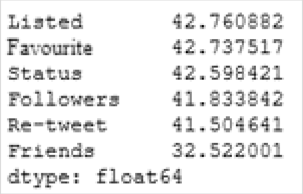
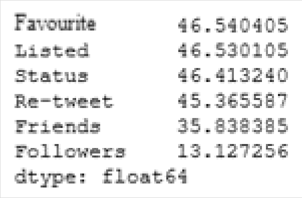
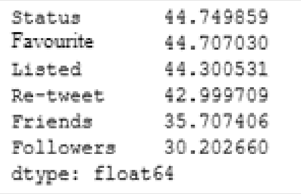
	Screen_name length	Status	Followers	Friends	Favourite	Listed	Retweet	Reply	Hashtag	URL	Mention
Human	Reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Bot	Reject H_0.	Reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.

Table 10. Benford’s law FSLD Experiment Results using [42] and [43] dataset

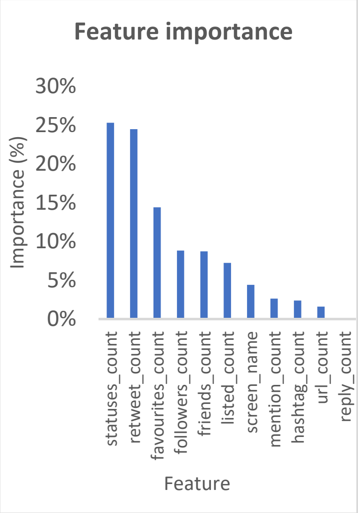
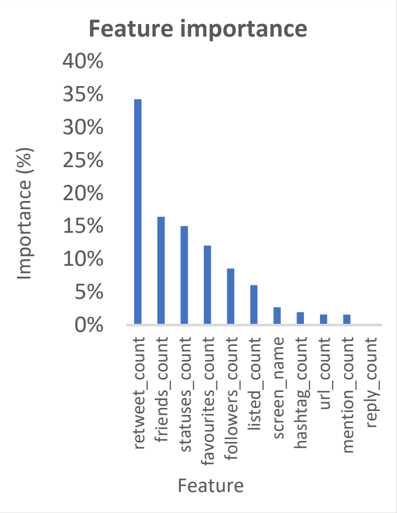
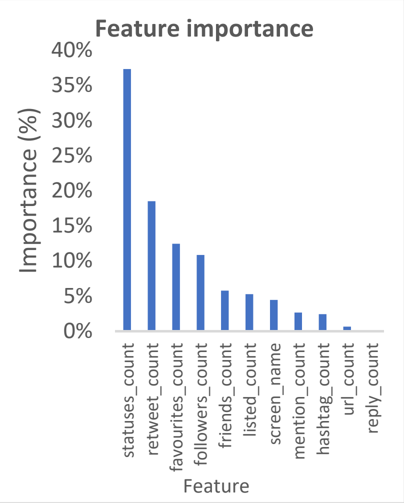
	Screen_name length	Status	Followers	Friends	Favourite	Listed	Retweet	Reply	Hashtag	URL	Mention
Human	Reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.
Bot	Reject H_0.	Reject H_0.	Cannot reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.	Reject H_0.

PCA Test results

Table 11. PCA results for different datasets.

	[41] dataset	[42] dataset	[42] and [43] dataset
Scree plot			
PC1 and PC2			
Loading scores	 Listed 42.760882 Favourite 42.737517 Status 42.598421 Followers 41.833842 Re-tweet 41.504641 Friends 32.522001 dtype: float64	 Favourite 46.540405 Listed 46.530105 Status 46.413240 Re-tweet 45.365587 Friends 35.838385 Followers 13.127256 dtype: float64	 Status 44.749859 Favourite 44.707030 Listed 44.300531 Re-tweet 42.999709 Friends 35.707406 Followers 30.202660 dtype: float64

Ensemble random forest results

[41] dataset – Important feature	[42] dataset – Important feature	[42] and [43] – Important features
 Feature importance Importance (%) 0% 5% 10% 15% 20% 25% 30% statuses_count retweet_count favourites_count followers_count friends_count listed_count screen_name mention_count hashtag_count url_count reply_count Feature	 Feature importance Importance (%) 0% 5% 10% 15% 20% 25% 30% 35% 40% retweet_count friends_count statuses_count favourites_count followers_count listed_count screen_name hashtag_count url_count mention_count reply_count Feature	 Feature importance Importance (%) 0% 5% 10% 15% 20% 25% 30% 35% 40% statuses_count retweet_count favourites_count followers_count friends_count listed_count screen_name mention_count hashtag_count url_count reply_count Feature