Interpretable Machine Learning in Natural Language Processing for Misinformation data

by

Yolanda Nkalashe

Submitted in partial fulfillment of the requirements for the degree
MIT Big Data Science
in the
Engineering, Built Environment and Information Technology
University of Pretoria, Pretoria

November 2022

Publication data:

Miss.Yolanda Nkalashe.Interpretable Natural Language Processing (Misinformation),University of Pretoria,(MIT Big data Science), Pretoria, November 2022.

Interpretable Machine Learning in Natural Language Processing

by

Yolanda Nkalashe
E-mail: u13193016@tuks.co.za

Abstract

The interpretability of models has been one of the focal research topics in the machine learning community due to a rise in the use of black box models and complex state-of-the-art models [6]. Most of these models are debugged through trial and error, based on end-to-end learning [7, 48]. This creates some uneasiness and distrust among the end-user consumers of the models, which has resulted in limited use of black box models in disciplines where explainability is required [33]. However, alternative models, ”white-box models,” come with a trade-off of accuracy and predictive power [7]. This research focuses on interpretability in natural language processing for misinformation data. First, we explore example-based techniques through prototype selection to determine if we can observe any key behavioural insights from a misinformation dataset. We use four prototype selection techniques: Clustering, Set Cover, MMD-critic, and Influential examples. We analyse the quality of each technique’s prototype set and use two prototype sets that have the optimal quality to further process for word analysis, linguistic characteristics, and together with the LIME technique for interpretability. Secondly, we compare if there are any critical insights in the South African disinformation context.

Keywords: Disinformation, Interpretability, South African, Prototypes, Example-based.

Supervisors	**:**	Dr. Vukosi Marivate
Department	**:**	Engineering, Built Environment and Information Technology
Degree	**:**	Masters in Technology (Big Data Science)

"Interpret means to explain or to present in understandable terms. In the context of ML systems, we define interpretability as the ability to explain or to present in understandable terms to a human. A formal definition of explanation remains elusive; in the field of psychology, Lombrozo states "explanations are the currency in which we exchanged beliefs" and notes that questions such as what constitutes an explanation, what makes some explanations better than others, how explanations are generated and when explanations are sought are just beginning to be addressed."

Doshi-Velez, F. and Kim, B. (2017) 'Towards A Rigorous Science of Interpretable Machine Learning'.

Contents

List of Figures

List of Algorithms

List of Tables

Chapter 1

Introduction

Machine learning models have become less explainable in the last couple of decades, such that they can solve complex problems while not compromising on predictive accuracy [3, 25, 16]. However, the more complex the machine learning model is, the more difficult it can be to explain how it made its decisions. Therefore, explaining or interpreting how a model makes its decisions has become paramount in recent years. Resulting in the model's learning biases, models not performing as expected in ”the real world,” end-users wanting to understand why a model selected or did not select them for a particular class/category, etc. Interpretability has thus become a crucial research field, especially in high-risk areas like self-driving cars, medical diagnosis, and criminal detection systems. Interpretable models build trust and transparency between the model and end-users. Various explainable models have been developed; however, some of the explainable models do not produce humanly interpretable outputs. Human interpretable models are essential as this ensures lay users can understand how a model makes its decision without having to be the model/system developer.

Natural language processing models interact with society frequently, like translation systems, next-word prediction, sentiment, and fake news detection. Therefore explainability needs to be humanly understandable in this field to ensure safety and trust.

This research aims to explore a humanly explainable model in natural language focusing on misinformation. We leverage known humanly explainable machine learning techniques on misinformation data to determine if there are no key characteristics of misinformation data that can aid in safety online when it comes to fake news.
This chapter provides an introduction to the study; we first discuss in detail the motivation and background, followed by research objectives & contributions, and finally, discuss the outline of this paper.

1.1 Motivation

Explainable AI refers to techniques designed to assist in interpreting how models make their decisions. The need to be able to interpret how a model makes its decision has risen in the past decade; due to regulation and ethical concerns for end-users to understand how a model came to its decisions [3]. For example, a recent law that has been put in place is the General Data Protection Regulation (GDPR) which states the "right to explanation," meaning that consumers of a model have a right to be provided with an understandable reason on how the model reached its decision [16, 27].
Interpretability is not required by all machine learning tasks or use-cases, it is required where an incorrect outcome has significant consequences or the problem has not been studied sufficiently leading to untrust in systems output [16, 3].
Explainable machine learning in natural language processing (NLP) tasks has recently become a focal topic. Historically NLP systems were based on hand-written rules and decision trees based on hard if-then rules. Thus, in earlier years, NLP tasks were explainable to a certain extent [13, 42]. However, with the introduction and rise in deep learning models and modern NLP tasks, the systems and models used for NLP tasks are more complex with deep learning networks or state-of-the-art transformer models such as BERT (Bidirectional Encoder Representations from Transformers) being used, making them less transparent and humanly explainable [3, 16].

The need for interpretability in NLP tasks is because most of the tasks in NLP interact with the society at large and tend to learn human biases [42]; For example, a sentiment classifier that Luminoso trained showed that the trained classifier was representing racist behaviour; as a neutral sentence ”Let’s go get Italian food” was classified more positive than a functionally similar statement ”Let’s go get Mexican food.” Another well-known example is Taytweets, the Microsoft tweet chatbot, which was supposed to resemble a 19-year-old girl but instead learned and tweeted racist and offensive tweets. This shows that interpretability in NLP is crucial to ensure the models can be trusted as some of the models have the potential to create instability in society. [13, 11].

In this research, we focus on the interpretability of NLP misinformation classification tasks. We have seen a rise in misinformation being distributed on various social media platforms. An example is the current COVID-19 pandemic; a lot of misinformation has been filtering through the media, causing significant panic in society. Governments took this spread of misinformation seriously and even tasked working groups to deal with the issue, as it may impact how effective a government can be in dealing with the pandemic. Thus, if we can explain misinformation and identify humanly interpretable characteristics, there would be more safety online.

Three main techniques can be used to explain models: model-agnostic, model-specific, and example-based techniques [17, 3, 16, 33]. Model agnostic are techniques that can be applied to any model, whilst model-specific are methods that can only be used for a specific family of models and lastly, example-based techniques focus on extracting examples from the data which is then used as an explanation for the models or summarisation of the data [17, 3, 16, 33].

In NLP tasks, extensive research has been done on model-agnostic and model-specific techniques, where the main aim is to determine which variables are the most influential to the decision made by the model. However, there is limited application of the use of example-based techniques where the objective is to explore a sparse sample of the dataset to achieve explainability. This paper explores using example-based techniques to achieve explainability in an NLP misinformation task.

Explainable machine learning literature has limited consensus on the difference between the term's interpretability and explainability. Some literature uses the terms interchangeably, while others have clear distinct differences between the two. In this paper, we will use the terms equally as done so in [29, 33] and define interpretability and explainability as "the degree to which a human can understand the cause of a decision" [16].

1.2 Objectives

The primary goal of the thesis is to use a human interpretable model for misinformation classification to identify and characterise misinformation effectively.We leverage example-based techniques, as psychological studies state humans understand or create reasoning through creating a mental model by using examples [30].

More specifically, our goal seeks to answer the following research questions:

1. What behavioural factors can distinguish a misinformation class?
   a. Can we use prototypes with a combination of another model and feature engineering technique to extract such features?
   b. Is there a contrast in these behavioural factors on different datasets from different countries or topics; what underlying commonalities exist?

2. How does prototype/example-based methodology compare to the most commonly used algorithms?
   a. Is an example-based approach significant to an approach that outputs an attribution map?

The secondary goal of this research is to determine the quality of prototypes selected by training a model on the sparse dataset and analysing how well the model performs. Based on the results, we can evaluate the quality of the selected prototypes (assist in use cases where there is insufficient data).

1.3 Contributions

The main contribution that this research will bring to the explainable research community is the use of example-based techniques on natural language misinformation tasks compared to the extensively researched model-agnostic methods with a focus on extracting key features and characteristics. Instead of focusing on the feature space, we focus on representative data samples and extract key behavioural insights to achieve interpretability in misinformation data. We further compare whether there is a contrast between South African fake news to Global news data characteristics.

1.4 Limitations

The main limitation we experience is the limited data for using South African fake news data. The dataset is comparable small to the rest of the data with only a sample size of 763 Fake news data and more than 40000 real news.

1.5 Dissertation Outline

The rest of this thesis will be organised as follows.:

- In **Chapter** 2 we look at and review the literature that has been researched on explainable machine learning.
- In **Chapter** 3 we discuss the data that was used in this research and how we performed exploratory data analysis[EDA] on our datasets.
- In **Chapter** 4 we provide a comparative application on the well known explainable techniques in natural language processing.
- In **Chapter** 5 we discuss the methodology that was applied in this research to determine interpretability in misinformation.

[0] Work presented on the 28th October 2021 at the Ethics and Explainability for Responsible Data Science (EE-RDS) and the paper is under Review

- In **Chapter 6** we provide the results of the experiments we performed.
- In **Chapter 7** provided the concluding remarks and discussion on future work.

Chapter 2

Literature Review

The research in this thesis aims to understand interpretability for specifically natural language processing tasks with a focus on misinformation tasks. We aim to identify characteristics of misinformation with the help of using interpretability techniques. Such that we can produce humanly explainable examples for misinformation tasks. This chapter starts by summarising the taxonomy of interpretability in section 2.1. We then discuss and review the literature on intrinsically interpretable models in section 2.3, followed by section 2.3, which focuses on post-hoc model interpretability, and finally, section 2.5 summarizes key takeaways from the literature reviewed.

2.1 Taxonomy of Interpretability

Interpretability is defined as the degree to which a human can understand an explanation and is required because of the incompleteness of the problem being solved. However, further to the issue of incompleteness various pieces of literature like [33] and [16] state that the importance and the why for having an explainable model is for the following reasons:

- Trust & Verification:

  The models must be trusted by the end-users so that when they interact with the model, they can fully trust the decisions made by it.

- Safety:

  Refers to understanding how a model works, which is valuable in high stake decisions.This allows for constant improvement of the system such that it has a near-to-zero error rate as we constantly learn from it.

- Debug & audit:

  Explainability allows users and developers to debug models when the model fails.

- Detect bias:

  Assists in picking and detecting if a model learns any biases from the data.

The survey by [14]: Opportunities and Challenges in Explainable Artificial Intelligence (XAI) summarises the above-stated need for interpretability into three groups being:

1. Accountability; to create a sense of accountability when a model fails.
2. Fairness; to ensure that models are fair and safe from bias and lastly
3. Transparency; to ensure trust and verification.

It is argued that accountability is the most important reason that we should aim to achieve with interpretability.

2.1.1 Evaluation of Interpretability

There has yet to be a consensus on how to validate interpretable models. However, there are three main suggested evaluation techniques that were first presented in the paper ”Towards A Rigorous Science of Interpretable Machine Learning” by [16]:

1. Application grounded method: The explainable model to be evaluated by the end-user in the environment in which it will be deployed.

2. Human grounded: How simply can a human evaluate the explanation.

3. Functionally grounded: How well does the explanation faithfully represent the model.

Further to the proposed evaluation methods [33] and [13] state that an explainable technique should have one or more of the following properties:

- Translucent: Relates to how reliant the model is to model parameters.The more translucent an explainable model is the better.
- Portability: Model can be used for different types of models.
- Accuracy: How well the explainable model predicts unseen data.
- Fidelity: How well the model mimics the model's prediction.The lower the fidelity,the more unreliable the explanations are.
- Consistency: Explanation must be consistent on different models for the same data.
- Stability: high stability is desired for explainable models; similar to consistency an explainable model is stable if there are no significant changes in data instances similar to the data point being explained.
- Comprehensibility: Refers to how well humans can understand the explanation.

2.1.2 Classification of Interpretability

Traditionally interpretable models are classified as either intrinsic or post-hoc which is determined based on when interpretability is achieved [11, 42].For example, suppose interpretability is achieved with the model's output or during the prediction phase.In that case, the model is classified as intrinsic, using figure 2.1 as a visual explanation, the model would end at the output phase and no additional processing would be required. Conversely, post-hoc interpretability is achieved through further processing after the model has predicted an outcome; this is seen in the adapted figure 2.2.

An explainable model $G(.)$ can either be 1. model-specific indicating that the technique can only be used for a specific class of machine learning models, or 2. Model-agnostic indicates that the technique can be used for all model types [16, 3, 27]. The output, g, of an explanation can be a:

- Feature summary: This indicates features that are the most relevant in making a decision for a particular class.
- Feature visualization: Output of an image highlighting the important features or pixels.
- Model internals: Internal weights or structure of models; usually the output of inherently interpretable models like decision trees, linear models, etc. and
- Data points: Returns sparse data points that best represent the full dataset X.

The scope of an explanation output, g, can either be local, explaining a single data point, or global, explaining the high-level overview of the model of class [14, 29, 17].

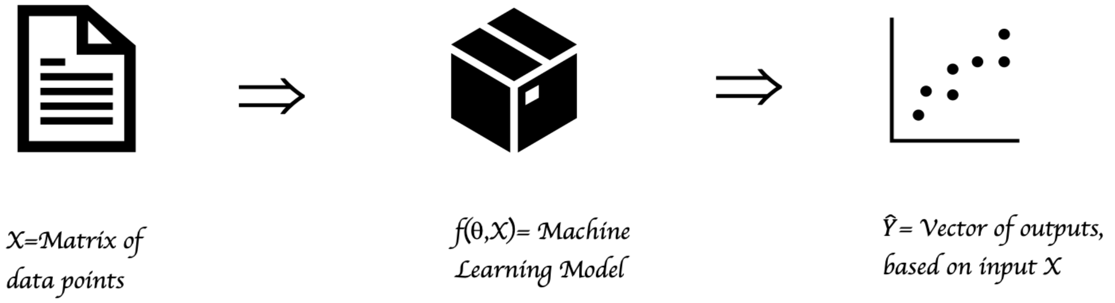

X=Matrix of data points
f(θ,X)= Machine Learning Model
Ŷ= Vector of outputs, based on input X

Figure 2.1: High-level overview of a typical machine learning system. Which takes in an input of X data points, which is fed into a learned machine learning algorithm f with parameters and outputs a vector of decisions $\hat{Y}$.

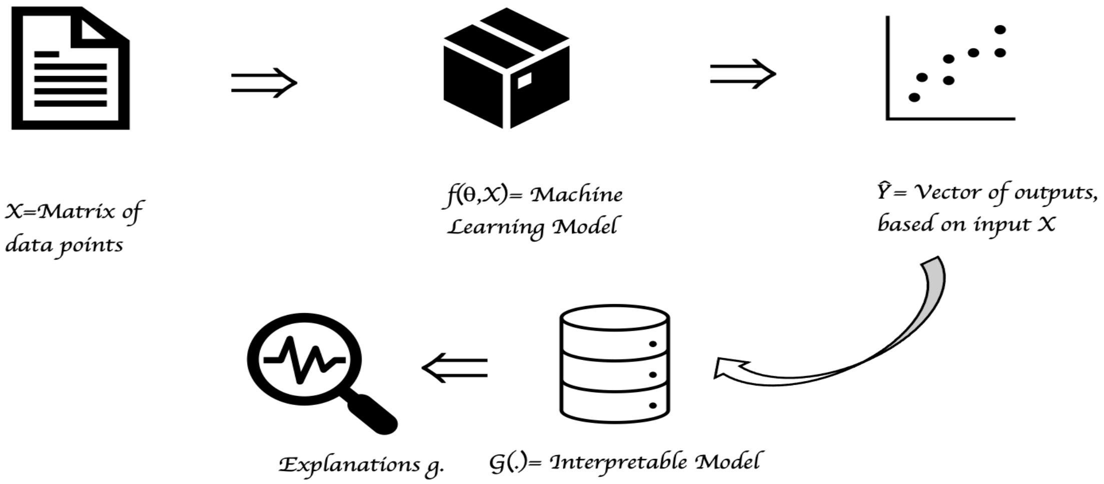

X=Matrix of data points
f(θ,X)= Machine Learning Model
Ŷ= Vector of outputs, based on input X
Explanations g.
G(.)= Interpretable Model

Figure 2.2: High-level overview of a typical post-hoc interpretable model. Which uses the output of machine learning and applies to an interpretable technique G to output an explanation.

2.2 Intrinsically Interpretable models

Most of the traditional intrinsic methods, like decision trees or logistic regression, do not offer the complex learning that comes with dealing with language data resulting in a significant trade-off of in accuracy. This motivated researchers to explore the ability to achieve intrinsic explainability in complex deep learning models by leveraging the internal structure and knowledge learned by the model. The prominently used architecture to achieve intrinsic interpretability using deep learning models is the attention mechanism that can be added to the architecture of a deep learning model [24].

The attention mechanism was proposed to improve the performance of language models. Attention uses a contextual vector, which is the output of the encoder layer; attention allows the input to focus on the previous output or, similarly, for the input features to interact with each other, known as self-attention [24, 44]. Because the attention mechanism produces weights, the weights can be used to output an attention map that highlights the most prevalent features that were used in each layer [12, 6]. Although attention may seem to provide a sense of transparency for deep learning models; [37] and

[6] states that the attention mechanism should be used with caution as it is a mixture of various representations and can be misleading.
[44] proposes to iterate through the attention query using an iterative, recursive attention model which attempts to trace back the contextualized word representation from the encoder with the aim of learning better representations and attention weights that can be used as an explanation of the model rather than normal attention.
Alternative to attention as an intrinsic interpretable model [10, 36, 19, 1] leverage off the convolution neural network (CNN) structure with adaptions in the last layers of the network to achieve explainability for the model's decisions. [49]'s approach uses the CNN with a ReLu(rectified linear unit) in the last two hidden layers, such that they can leverage the local linear models (LLM) produced by ReLu. The explanations are obtained by determining the importance of the text inputs by using the weights of LLM. Similarly, [19] uses a TC-CNN which, instead of ReLu, in the last hidden layer global average pool is applied to the spatial feature vectors of the CNN to obtain a single feature vector F. The obtained feature vector is used on class activation map extraction to determine the contribution of each feature.
Uniquely [36] uses latent features which are explainable to train a deep learning model. The proposed latent variables fall into six categories of the complexity contour generator (CoCoGen), which are part of the Stanford CoreNLP toolkit. The latent variable list and description can be seen below in table 2.1. The approach first analyses text using CoCoGen to obtain an output sequence of 154 feature vectors that relate to the features described in table 2.1. The feature vector obtained can then be fed into a network to determine output and explanation based on the relative importance of the language feature groups. The relative importance of a feature group is determined through delta scores which are standard scores obtained by performing normal standardization on all indicators (feature groups); with the scores, one can distinguish interesting patterns between classes.

Feature group	Size of Sub-types	Example/Description
Syntactic complexity	18	mean-length,clauses per sentences, Coordinate phrases per clause and Particular structures
Lexical Richness	12	contents words, Lexical diversity,type token ratio,Lexical sophistication
Register-based n-gram frequency	25	measures of frequencies, n-gram frequency of order 1-5 from five language registers
LIWC-style	60	function, grammar perceptual, cognitive and biological processes, personal concerns, affect, social, basic drives.
Word-Prevalence	36	information on word frequency,contextual diversity and semantic distinctiveness.

Table 2.1: Table of CoCoGen feature groups and descriptions adapted from the literature [36]

2.3 Post-hoc Techniques

The most common techniques for explainability for deep learning models is post-hoc methods. The techniques aim to achieve interpretability after the model has been trained, by using the information and output of the trained model to explain why a model made its decision as shown visually in figure 2.2. The techniques can be split into gradient/saliency based, surrogate based, probing and example based techniques.

2.3.1 Gradient based techniques

Gradient-based interpretable methods use gradient signals to assign relevancy of feature inputs with the goal of creating an attribution map[35, 43]. The attribution map captures the importance of each feature. Attribution maps have to be sensitive to changes in the input space to ensure they do not focus on unimportant features and are invariant such that they are consistent in the output result when applied to different models [43, 29]. Further, gradient methods commonly achieve local interpretability for a single instance x_i looking back at figure 2.2; figure 2.3 shows an adaptation of the explanation function G and output to show what the methodology outputs and the baseline explanation function.

One of the well-known attribution techniques is saliency, termed the vanilla gradient in some literature. Explainability is achieved by taking the gradient of the output with respect to the input. However, saliency can be very noisy. To improve the noise [29, 2] proposed grad * input to smoothen the attribution maps by multiplying the gradient with the instance value, see equation 2.1 and 2.2 respectively.

$$AttributionMap = A^C = \left| \frac{\partial \hat{Y}}{\partial \ x_i} \right| \tag{2.1}$$

$$A^C = \left| \frac{\partial \hat{Y}}{\partial \ x_i} \right| \circ \ x_i \tag{2.2}$$

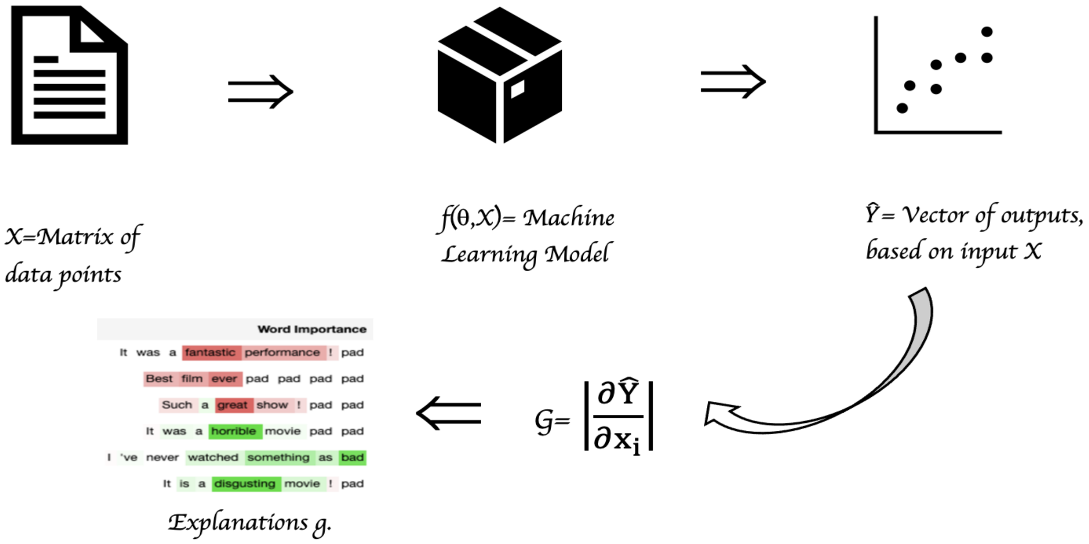

X=Matrix of data points
f(θ,X)= Machine Learning Model
Ŷ= Vector of outputs, based on input X
G= |∂Ŷ / ∂xᵢ|
Word Importance
It was a fantastic performance ! pad
Best film ever pad pad pad pad
Such a great show ! pad pad
It was a horrible movie pad pad
I 've never watched something as bad
It is a disgusting movie ! pad
Explanations g.

Figure 2.3: High-level overview of a salient-based explainable model. Which uses the output of machine learning and calculates the gradient of output with respect to input to determine important features for making the decision.

Vanilla gradient explanations are usually best suited in models where a linear relationship exists between output and input; given that deep learning models are complex and linearity does not exist, the technique may not be as faithful to deep learning models [34, 35]. [43] proposed integrated gradients, a modification of the gradient approach. Integrated gradients averages all gradients along a straight line between a baseline value $\bar{x}$ and the input of interest x_i. The baseline is usually chosen to be an empty sequence.

Attribution methods only focus on the change in the output layer with respect to the input layer. They do not consider the intermediate layers and the changes of the feature gradient [43, 35]. Attribution propagation is proposed, which looks at each layer from the output to the input. Propagation methods decompose the network work layer by layer assigning relevancy scores on each neuron[35, 34]. One of the most popular attribution propagation methods applied to NLP downstream tasks is layer-wise relevancy propagation(LRP). LRP starts from the outer layer $\hat{Y}$ and works backward, assigning relevancy score, R_k, to each neuron based on its sensitivity it is to its successor layer [43]. The score is calculated by decomposing the activation function in each neuron rather than

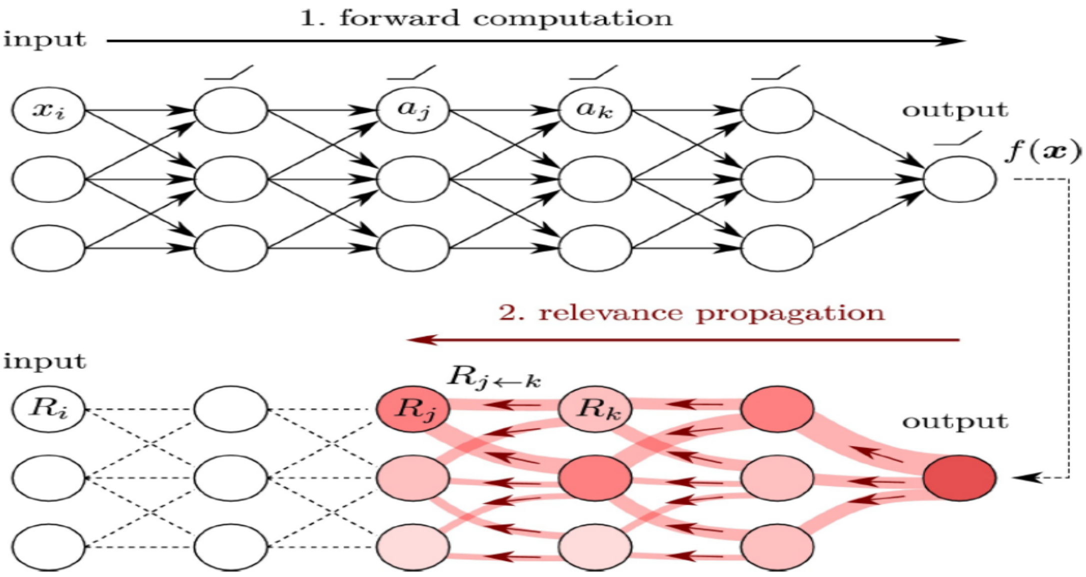


Figure 2.4: Overview of a Layer-wise propagation technique as shown in [42]. Shows the decomposition process from model prediction layer to input feature layer.

calculating gradients [42]. Figure 2.4 shows a visual example of the decomposition.

Alternative to LRP, [39] proposes a more improved propagation technique, DeepLift. Instead of directly redistribution the score from the upper layer to the lower layer, DeepLift uses a baseline to calculate the relevance score by looking at the effect of the activation neuron from the original data, x_i, compared to the baseline data $\hat{x}_i$. Contrary to attribution-based methods, [27, 3] Proposes the use of surrogate models over attribution methods as attribution methods can be fragile and highly unreliable for an explanation of complex-based models.

2.3.2 Surrogate Models

Surrogate models attempts to mimic the model and how it achieved its output by using the information from the model to train a more intrinsic model with the information captured [37, 23]. Surrogate models were first proposed by [46] to model latent variables with the aim of making the process more explainable. The benefit of using surrogate models for explainability is that they are portable and can be applied to any model for any dataset [33, 16]. The commonly used surrogate model in machine learning is LIME (local interpretable model-agnostic explainer). LIME was first proposed by [37], in the paper ”why should I trust you? Explaining the predictions of any classifier”. The technique selects an instance x_i for interpretation that has been predicted using a model F; LIME selects perturbed points close to the instance to create a new data set. The new data is trained using the model F and an intrinsic model like logistic regression to gain insight and explanation for the model's decision. LIME outputs the most significant variables for the model when making a classification decision.

LIME only produces local explanations; as such, there have been various adaptations of LIME to achieve global interpretability. [23] proposes K-LIME, which uses partitioning to partition the space of the predictors and fits an intrinsic model in each partition. Similarly, [37] proposed SP-LIME, which rather than partitioning, picks a set of instances that best represent the data that are used to fit the intrinsic model to gain a global understanding of the model.

A major limitation of LIME is that the weights produced do not necessarily indicate the importance of a feature and can create a false sense of relevance in the features. [49] and [33] propose to compute Shapley values for explanation compared to LIME. Shapley values use the ideology of game theory to create a linear combination or model for every permutation of features. The shapely value shows the average contribution of a feature across all possible permutations. For example given data which has two features $[x_1, x_2]$ shapely will average across the weights obtained from the permutation $[x_1]$, $[x_2]$, $[x_1, x_2]$. The computational time of obtaining Shapley values can exponentially increase with the more features one has.

Contrary to determining which features are relevant as explanation [22], proposes to use feature importance techniques with probing techniques that can inform us what linguistic properties are captured in the model. There are various techniques which probe for various properties such as syntactical structure, context, semantics, sentence length, word order, verb tense. Probes are shallow networks that are placed on the intermediate layers of a network and can investigate in each layer what is captured [22, 48, 18]. If the task being probed for performs poorly, in accuracy terms, then the information is not encoded or is not needed to make a decision. The main issue with probes is they can be complex and maybe be unfaithful; instead of learning the representation being probed it learns the task [22]. Given that linguistics are essential for NLP, probes are important in giving a holistic explanation and should instead be used as a combination with an alternate explanation technique.

2.3.3 Example based techniques

[45] argues that using importance value for interpretation may be abstract for a lay-user to gain a full understating for NLP task as they may not grasp the significance of the importance of the word “help” for instance. They thus propose using example-based techniques, given that humans work with mental models and contrasts for explanations; therefore, example-based techniques will help humans build mental models of the machine learning model [30]. The three most prominent interpretation example-based methods are:

Counterfactual examples

Counterfactual examples aim to achieve interpretability by producing a contradictory reality; It aims to describe a causal situation by stating that if X had not occurred , Y could not have occurred. In figure 2.5 from [38], we see an example of a counterfactual example generator using text-to-text transfer learning; for an instance where the explanation is a contrasting scenario selected from a sample of candidates. Counterfactual is interested in what changes in the input of an instance will cause a prediction to flip [33, 29]. In earlier years, to obtain counterfactuals, it was either through trial and error, where features where flipped until the desired outcome was obtained, or manual annotations were required by experts of the data [38]. Recently, there have been tool-kits created to obtain counterfactual explanations from a model by masking a percentage

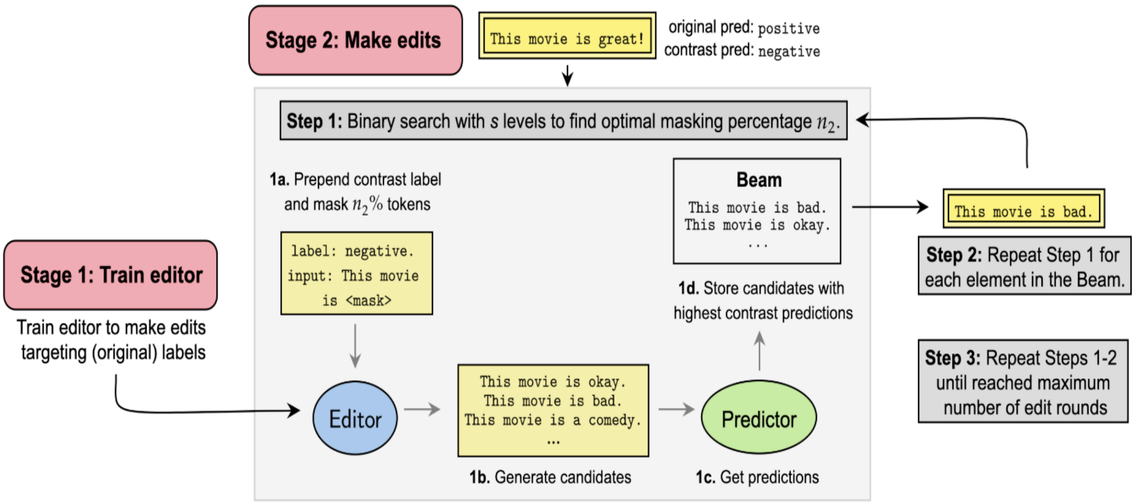

Stage 2: Make edits
This movie is great!
original pred: positive
contrast pred: negative
Step 1: Binary search with s levels to find optimal masking percentage n_2.
1a. Prepend contrast label and mask n_2% tokens
label: negative.
input: This movie is <mask>
Beam
This movie is bad.
This movie is okay.
...
1d. Store candidates with highest contrast predictions
This movie is bad.
Step 2: Repeat Step 1 for each element in the Beam.
Step 3: Repeat Steps 1-2 until reached maximum number of edit rounds
Stage 1: Train editor
Train editor to make edits targeting (original) labels
Editor
This movie is okay.
This movie is bad.
This movie is a comedy.
...
Predictor
1b. Generate candidates
1c. Get predictions

Figure 2.5: Overview of counterfactual generator as shown in [38]. Shows how from the positive classified input text of a sentiment classifier "This movie is great," a % of the words are masked; in this case, the word 'great' is masked, and editor searchers for appropriate counterfactual for the use case being "This movie is bad." Explaining that if the word "great" was not there, the sentiment decision would not have been positive

of data through text-to-text transfer learning and conducting a binary search to generate edits, and once the desired outcome is obtained, the process stops. Similarly, [45] proposes Polyjuice, which uses a library of annotated data to generate real-world counterfactuals. Counterfactuals may run into the issue of having too many explanations that can contradict each other [3]. Moreover, they need to be minimal in the sense that a minimal change to the input is inserted into the original data point, further, they must be grammatically, and semantically meaningful [45, 37, 16].

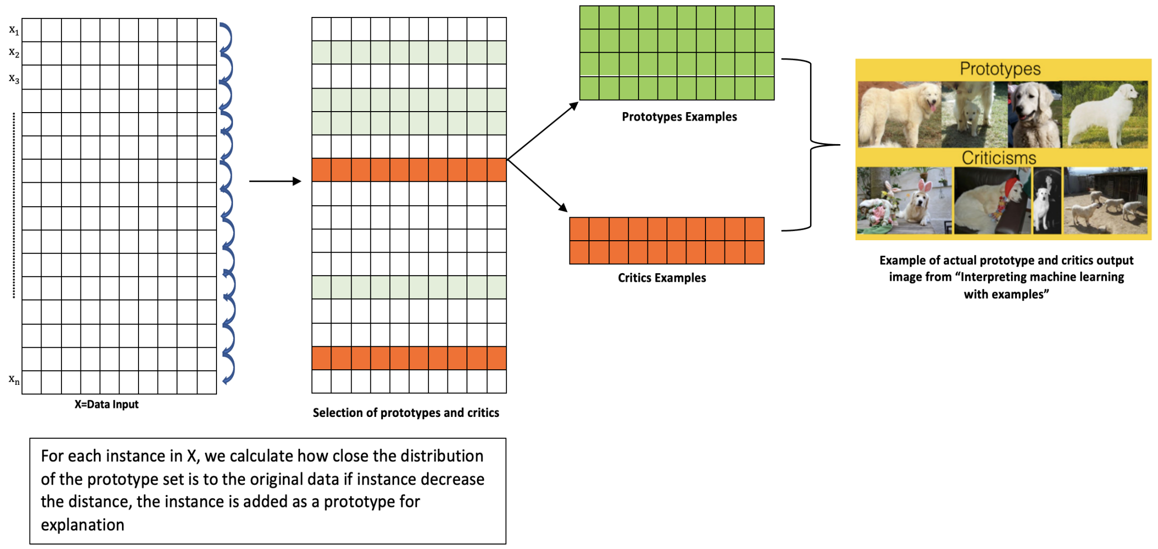
Adversarial Examples

Adversarial examples are very similar to counterfactual examples, but more with the aim of identifying the model's vulnerable points [33, 29]. Adversarial examples are looking for input points that would make the model make an incorrect classification; for instance, in an NLP misclassification case, the input example would be text that would classify the text as genuine when in fact, it is fake. Adversarial examples are commonly used for debugging a model more than interpretability [29].

Prototype Selection

The prototype algorithm goes back to 1968 when it was termed instance-based learning. The original aim of selecting prototypes was for data editing purposes with the aim of improving accuracy and efficiency in their data, rather than wanting to understand the data set [8]. Historically the techniques to select prototypes were using clustering techniques (K-medoids or K-means) and determining how far instances are from one another [8, 33, 25]. A recent approach, with the aim of explainability, is the set cover theorem proposed by [8]. This approach aims to find the smallest subset of data so that every point around the data is in close proximity.

All the techniques above are aimed at just identifying the dataset which best represents the full data; however, what about the outliers of the data? [7] states that if we just rely on prototypes, we may miss important information about our data and consequently over-generalize. They propose an approach that leverages a bayesian framework and aims to extract prototypes as well as critics which are data points that do not best represent the data to gain holistic explainability and understanding of complex data [7, 33]. The algorithm is commonly known as MMD-critic. It has been explored extensively in image classification tasks where the prototypes are visual examples of images that best represent the data. From the small set of prototypes and critics, one can understand the data and different attributes and features; a visual example of the method can be seen in figure 2.6. The outputs of the prototypes can be used to explain a particular class in a classification task and why the model made its decision. Further, prototypes can be used to train the nearest neighbour classifier for intrinsic explainability [3].



Example of actual prototype and critics output image from "Interpreting machine learning with examples"

For each instance in X, we calculate how close the distribution of the prototype set is to the original data if instance decrease the distance, the instance is added as a prototype for explanation

Figure 2.6: Overview of prototype and critics selection process; we start of with the full dataset X and from the data set add, x_i, data instances which minimise the distance between the distribution of prototypes and full dataset

2.4 Application of NLP Interpretable models on misinformation models.

In this research, we focus on one of the applications of NLP, misinformation. The rise and detrimental effects of fake news have attracted the use of deep learning models for detecting fake news. However, the model's performance alone is not sufficient [5]. Therefore in this section, we zoom into the literature related specifically to misinformation explainability. [31] in the paper "A Survey on Explainable Fake News Detection" categories the explainability of misinformation models into:

1. Explanation by social features:

Techniques that use social media platform data to achieve interpretability in fake news detection. These methods often use comments and user profile history of who shared the news article to explain why an article is classified as fake. For example, [40] proposes the dEFEND framework, which uses the co-attention mechanism by using both comments and news content to achieve interpretability. Similarly, [28] proposed using Graph Neural Networks, which looks at more attributes compared to the dEFEND framework. Over and above the comments and news content, the methodology also uses profile history features to explain why a text was classified as fake. However, these techniques are often limited to articles disseminated through social media platforms.

2. Explanation by feature importance:

The most common techniques used to achieve interpretability in misinformation models aim to extract key features. [5, 20, 4] suggest the use of surrogate models or a combination of surrogate models to extract key features. They propose to use techniques such as LIME, anchors, SHAP, and attention mechanism with misinformation models to achieve interpretability. Given the gaps identified with surrogate methods in 2.3.2; [5] further suggests using human-in-the-loop together with the techniques. They selected a sample of people; one group was provided the output of the model's prediction, while another group was provided with both the output of the model and feature importance outcome from SHAP. The observations noted were that when humans are included to achieve explanation, it is easier for the users to trust a model and build mental models of the different categories in the future. Uniquely, [36, 32, 41] suggest using latent variables and hierarchal structures to achieve explainability in misinformation use cases. [36] used latent variables as discussed in 2.2; the latent variables are generated using the complexity contour generator and fed into a misinformation model. The relative importance of the feature group is then used for an explanation by observing any interesting patterns in the classes. On the other hand, [32, 41] use hierarchal structures to achieve interpretability in misinformation use cases. [32] uses a hierarchal attention network; the attention network looks at salient sentences to determine the most important sentences in each article. The highlighted sentences are outputs and are used to explain the model's prediction. Alternatively, [41] builds a hierarchical propagation network for each class to extract structural, temporal, and linguistic features. The features extracted from each class are compared to note if any differences exist between fake news data and true news data.

3. Explanation by news content:

These methods output the most important words based on the linguistic and semantic features from the content of the article [31]. [47] developed the xFake system to achieve this; the system is designed to analyze news items from different perspectives. The system uses PERT, which studies the linguistic properties in the news context, and ATTN, which studies the semantic properties. As a result, the system highlights important words or features based on linguistic and semantic properties to achieve explainability in misinformation articles.

In most literature, when it comes to misinformation models and interpretability, the aim is to provide a reasonable explanation to the user to increase user trust and confidence. Given that the users are not experts, it is thus paramount for the explanation to be humanly understandable [5, 31, 4].

2.5 Summary

In the sample literature reviewed, we see that there has been significant progress in the research of interpretable black box models for natural language tasks. Most approaches offer post-hoc interpretability meaning explainability is achieved once the model is trained. We determined two categories of techniques of explainability into two model-agnostic and model-specific. We find that the approaches applied to NLP models aim to understand how features interact, determine which words are the most relevant, and what linguistic properties are captured in the model. We found that example-based techniques are applied for natural language tasks but with a primary focus on counterfactual and adversarial examples. We focus on example-based methodologies, which select a subset of the data or instance of the data for interpretability in an NLP misinformation classification task. We further apply the methodology to a South African fake news data set to determine if there are any significant insights between global vs. South African fake news.

Chapter 3

Data Collection

This chapter describes the data we used for this research and how it was collected. The data used is categorised into true and fake news classes. In section 3.2, we describe the pre-processing steps that we took on our datasets, and in section 3.3, we discuss the exploratory data analysis.

3.1 Datasets

3.1.1 ISOT DATA

We use the commonly known misinformation dataset, the ISOT dataset, as the baseline dataset collected by the University of Victoria's Information Security and Object Technology lab. The dataset consists mainly of a combination of world news and political news from the year 2015 to 2018. The data was collected from reuters.com, Politifact, and Wikipedia. In table 3.1 and 3.2 we show how the data is structured and the size of the data.

Data Attributes	Description
ID	Unique identification of document
Title	Title of document
Author	Arthur of document
Text	Full text of document
label	Class specification (1-misinformation, 0-true)
Date	Date the news article was posted(ranges between the year 2015-2018)

Table 3.1: Table showing attributes ISOT dataset

Misinformation	Size (number of documents)
Fake News	23481
True News	21417

Table 3.2: Table showing number of news articles in each class

3.1.2 COVID Dataset

The second dataset we use is a COVID dataset; we chose this dataset given the recent increase in misinformation given the COVID pandemic; the data is named COVID-FNIR and extracted from IEEE DataPort. The dataset contains news articles related to COVID news during the year 2020. It consisted of India, USA, and Europe data scraped from different sources; fake news was collected from Poynter, and true news from verified Twitter handles and news publishers. The fake and true news datasets were stored in different files as they had slightly different attributes in table 3.3. We show the attributes and also include the size of the data.

Data Attributes	Description	FakeNews	TrueNews
Date	Date the article was created(Period of 2020)	Yes	Yes
Link	Link to the article	Yes	Yes
Text	Full text of the news	Yes	Yes
Region	Region the article originated from	Yes	Yes
Country	Country the article originated from	Yes	Yes
Explanation	Explanation of why it was labeled fake news	Yes	No
Origin URL	Original link to article	Yes	No
Poynter Label	The class label provided by poynter	Yes	No
Label	Classification provided by the data collectors	0(Fake)	1(True)
size	Size of Article	3794	

Table 3.3: Table showing attributes COVID dataset

3.1.3 South African Dataset

South African dataset, which was curated with fake and real news data; the news articles were posted between the years 2017-2019. The fake news data was obtained from Zenodo Data Repository, which was data scrapped from disinformation websites in South Africa. The true news data was collected from various news media across South Africa news24, eyewitness, etc. The limitation of South African disinformation data is that it is minimal in size; we had 763 Fake news data and more than 40000 real news datasets as shown in table 3.4.

Data Attributes	Description
Date	Date article was released (ranges between the year 2017-2019)
Title	Title of document
Text	Full text of document
URL	URL to the news article
Medium	Medium news was shared

Table 3.4: Table showing attributes of South African dataset

3.2 Data Pre-processing

We perform pre-processing such that we can have quality data when applying explanation techniques. First, we reduce the number of attributes in each dataset. We want all the data sets to be standard and comprise just two columns, text and label. For the Covid-19 dataset, we had to change the class labels to align to ISOT where 0=True and 1= Fake news; we added an additional label column to the South African data to have similar class labels.

To ensure our data was ready for processing, the following steps were taken.

- Removed all empty text instances and instances with less than five words as these instances may not provide us with the necessary insight required.
- We detect non-English texts in the text data through language detection.
- We perform regular expression processing by removing HTML tags, digits, punctuation, stop-words, remove duplicate texts, and lemmatizing the text data.
- For each data set we perform further cleaning; for the ISOT data we had to remove the publication words in the text such that the models do not pick up biases; for the Covid dataset we ensure one consistent wording for the word covid and lastly for the South African dataset provided that it has imbalanced classes we performed under-sampling.

3.3 Exploratory data analysis

To analyse the dataset further, we use some exploratory text analysis techniques. We analyze prominent words and word length, perform sentiment analysis, and lastly, unsupervised topic modelling.

3.3.1 Word/length analysis

We analyse the length distribution of words and characters in each dataset by determining the average word length and vocab in our text. We compare the word count of class 1 (fake) and class 0 (True). We find that in both classes across the datasets, the average length of words of fake and real text is similar, with fake news having more outlying values as seen in figure 3.1 to figure 3.3.
To further analyse the words in each datasets we observe the top bi- and tri-gram words and word cloud to observe the most frequent words in each dataset. As observed in appendix 1, we find the most frequent words in the true dataset similar to that of fake news, which is expected as fake news text tries to disguise itself to be true and is usually around a common topic; we summarise the frequent words in table 3.5.

Data	ISOT	Covid	SA
Class 1(Fake)	President,Country,Trump	COVID19,People,lockdown	ANC,country,Police,government,Ramaphosa
Class 0(True)	President,Country,Trump,Republican	COVID19,Test, lockdown,hospital	ANC,country,newcase,government,ramaphosa

Table 3.5: Table showing prominent words in each datasets as observed in the word cloud in Appendix1

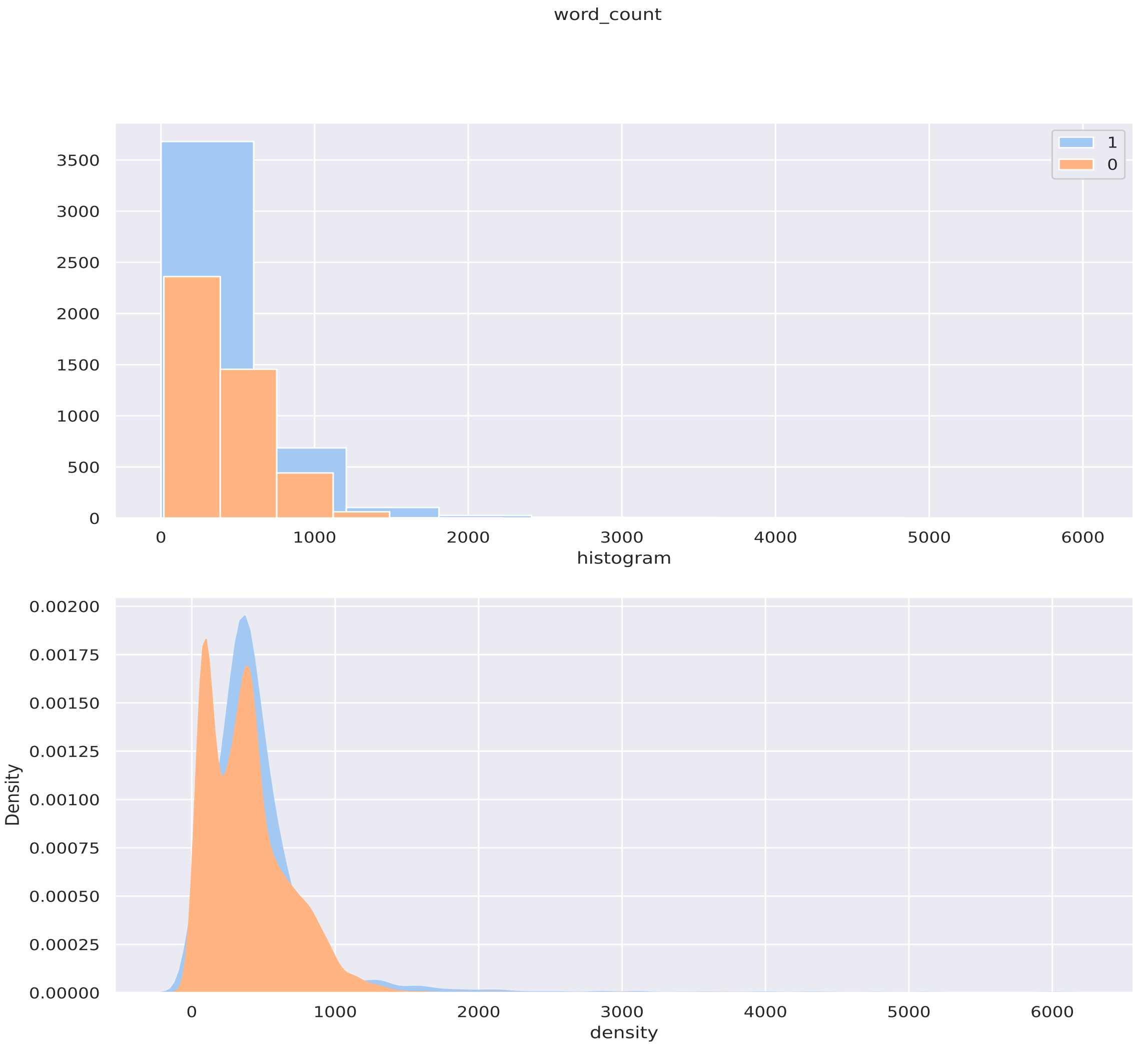


Figure 3.1: Density plot of word count for ISOT dataset

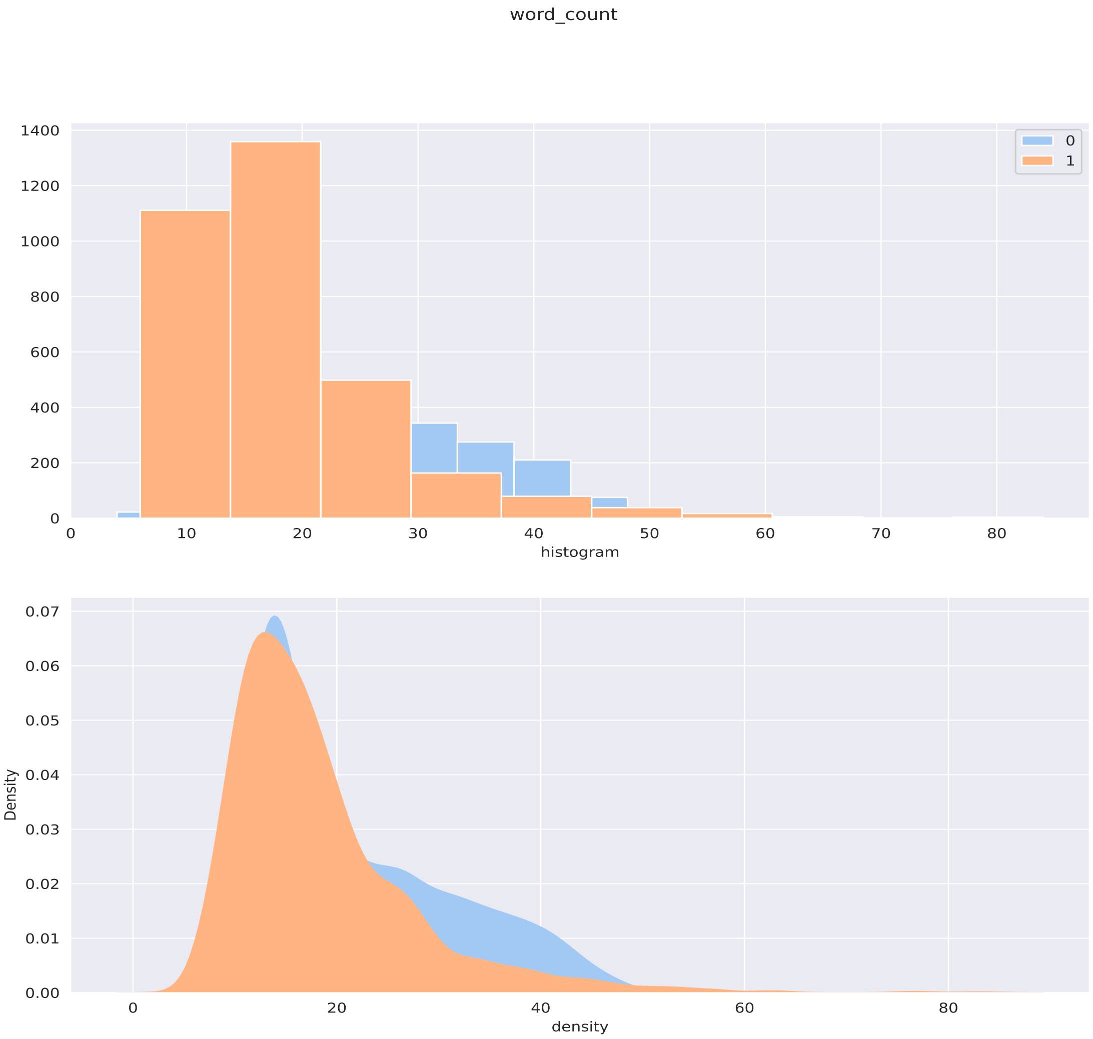


Figure 3.2: Density plot of word count for Covid dataset

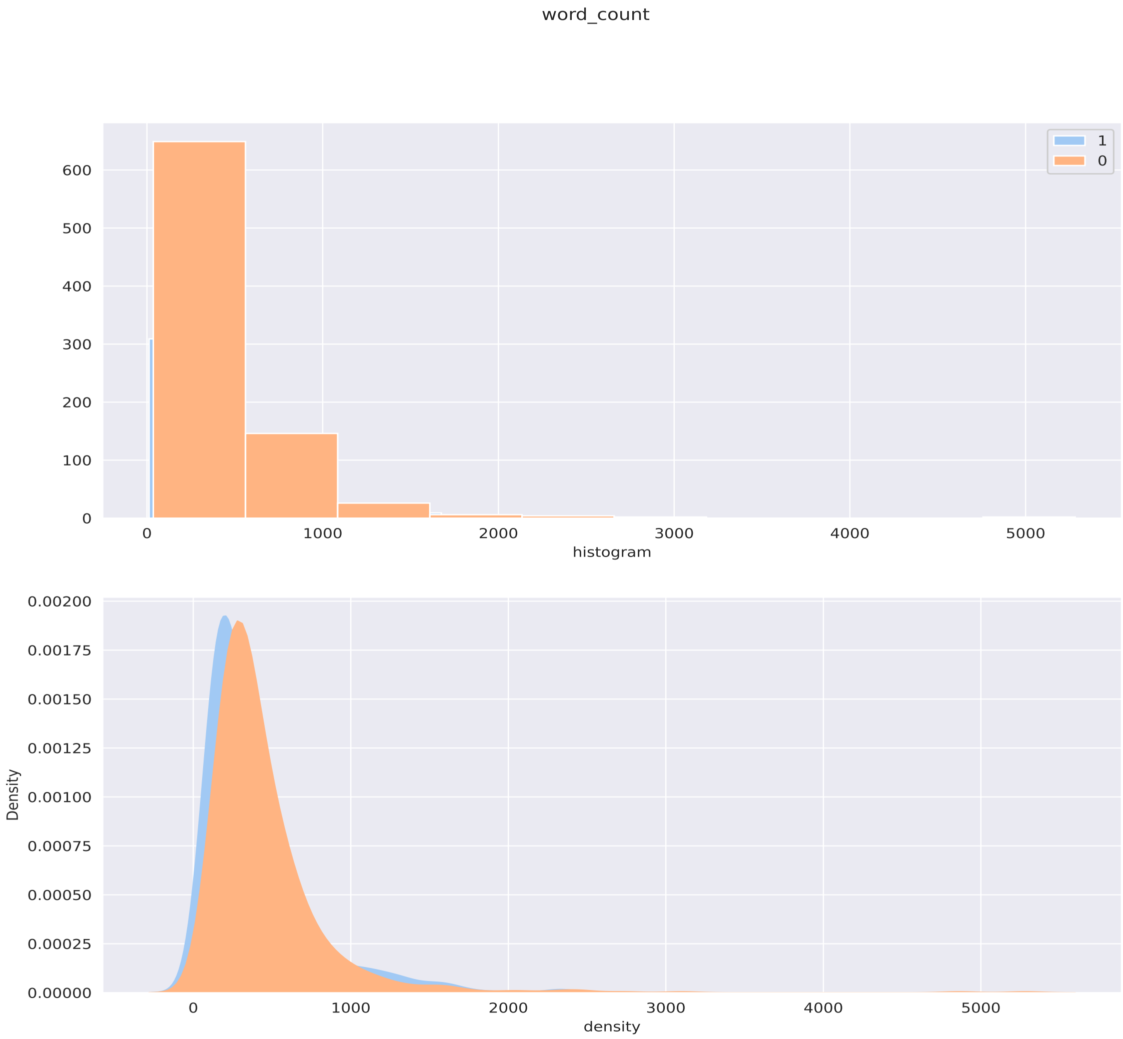


Figure 3.3: Density plot of word count for SA dataset

3.3.2 Sentiment analysis

Textblob is used to determine the sentiment of the text between the classes. Overall the sentiment of each class follows a similar distribution and is neutral. We further observe, as can be seen in figure 3.4 - 3.6 that the South African and ‘Covid data has more positive outlook compared to that of the ISOT data, which may also be influenced by the vocabulary and culture in the different regions the data was collected.

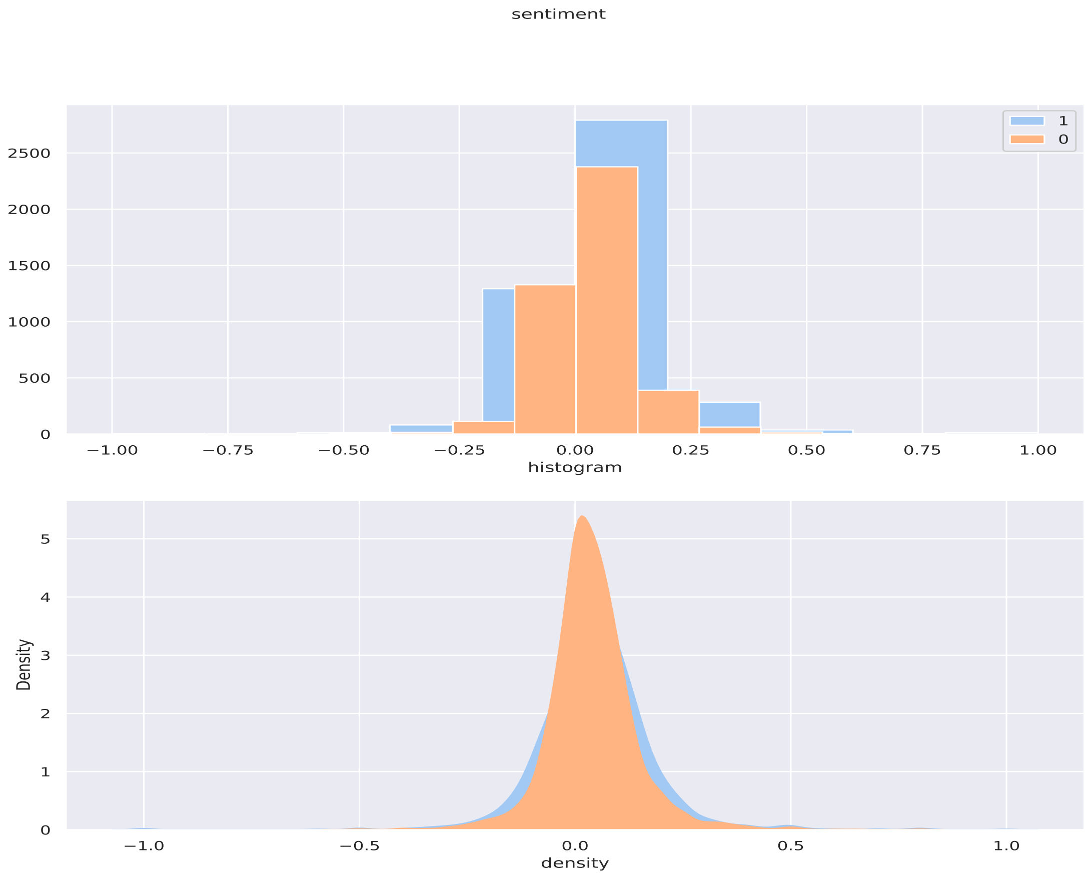


Figure 3.4: Density plot of Sentiment for ISOT Data

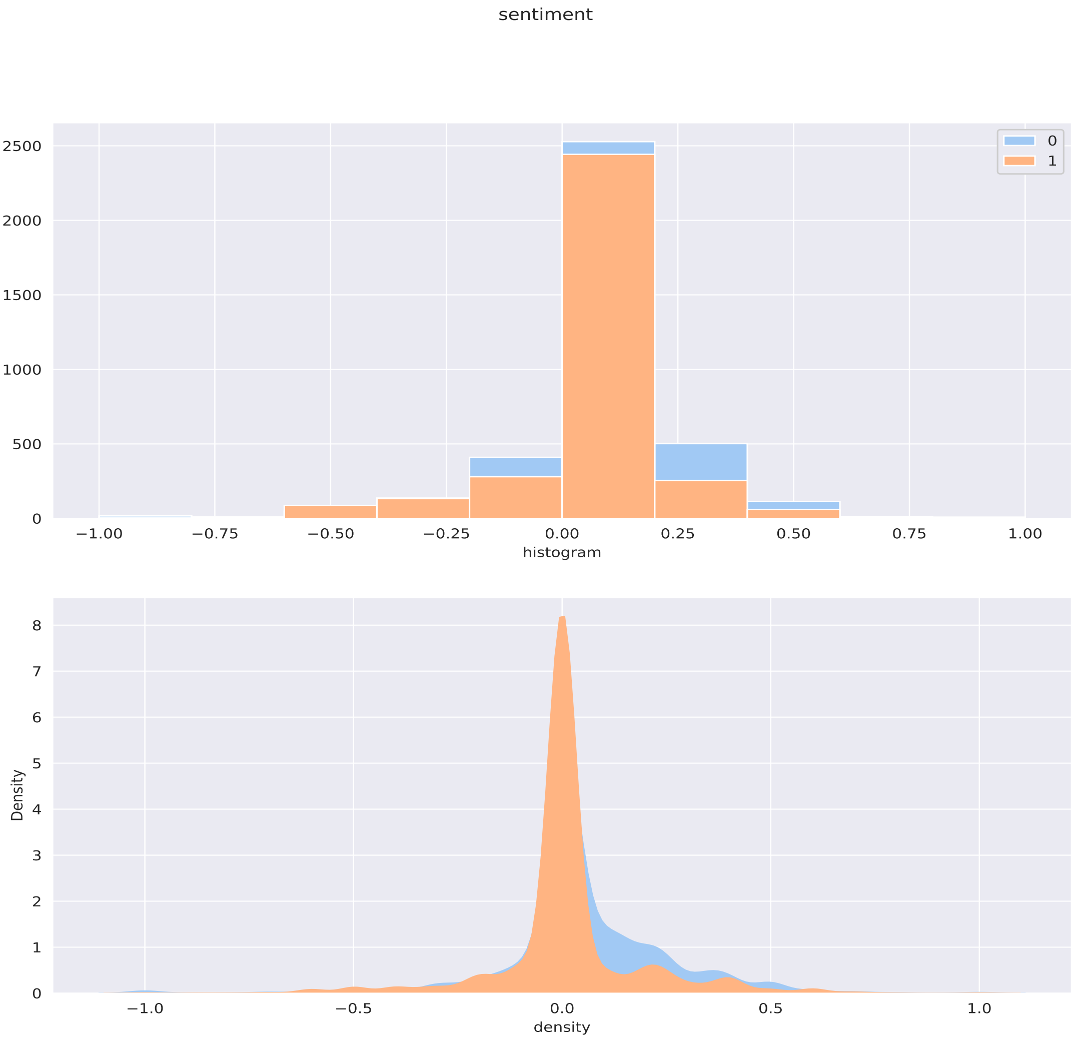


Figure 3.5: Density plot of Sentiment for Covid Data

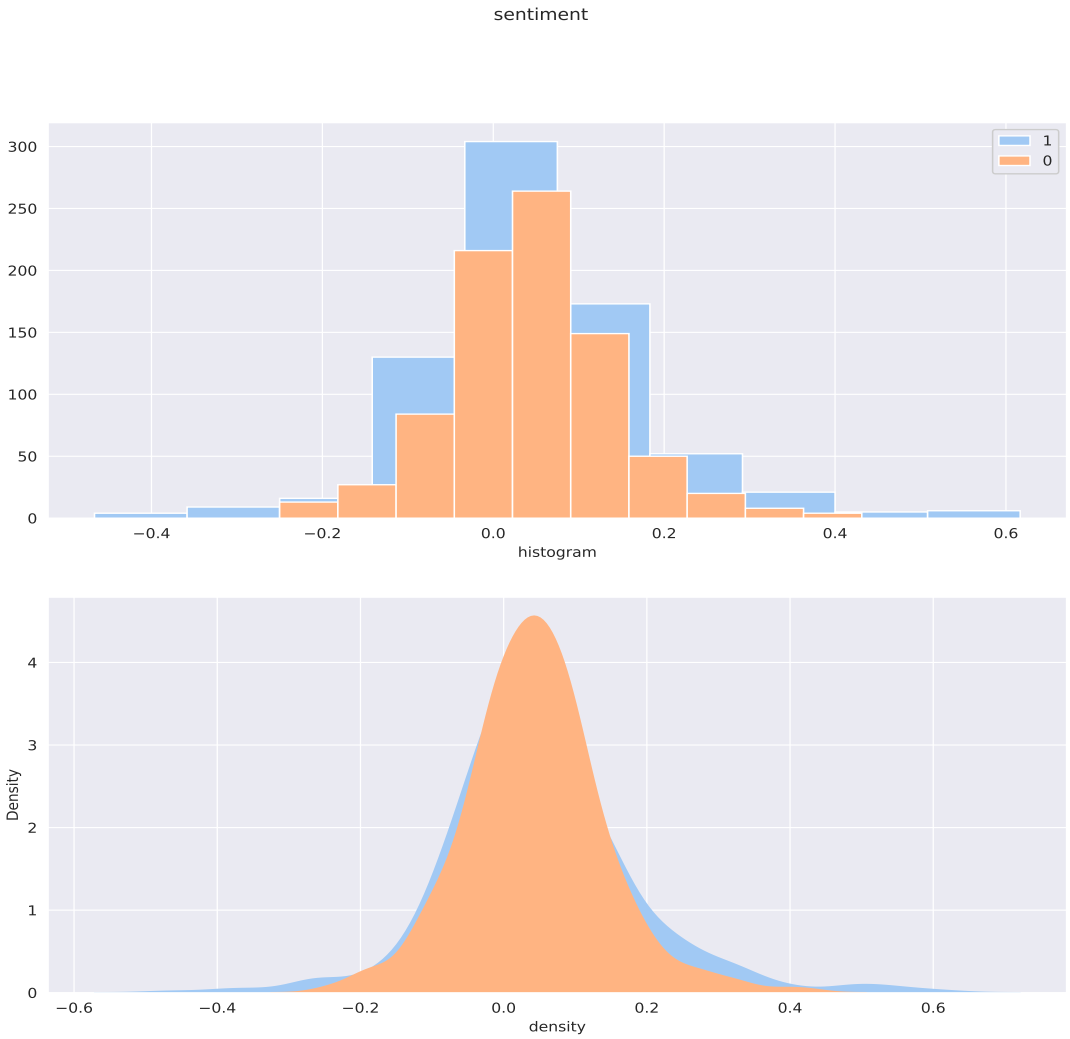


Figure 3.6: Density plot of Sentiment for SA Data

3.3.3 Unsupervised Topic Modelling

Lastly, we used the genism model to perform unsupervised bi-gram topic modelling on our dataset to determine the common themes in our data. The output can be observed in Appendix 2. We observe the themes in ISOT are centered around Presidency, American politics, and Ministers; the Covid data is centered around Covid cases, death rate, and lockdowns.

3.4 Summary

We observe that the prominent words, sentiments, and topics in fake news are similar to that of true news. Showing it is challenging to separate fake and true news by just using prominent words. We will explore this further in the coming chapters.

Chapter 4

Measuring and Comparing Attribution Type Interpretable Models

In this chapter, we explore some of the popular explainable techniques applied to NLP misinformation classification tasks. We use the baseline dataset ISOT to obtain results from some of the most popular techniques for interpretable machine learning. We discuss each technique and compare the output results.

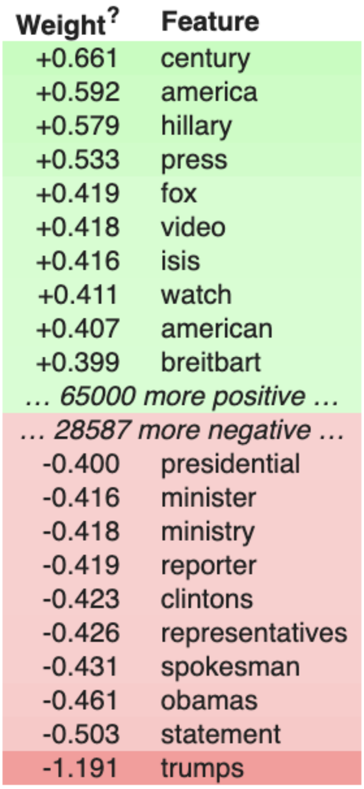
4.1 Logistic Regression & Decision Trees

We first explore intrinsic approaches, logistic regression, and decision trees that were traditionally used on NLP classification tasks to achieve interpretability [27]. We first tokenize our datasets before we train them using the models. We explore two tokenizing approaches, TfidfVectorizer, and word2vec embeddings. We further explore uni-gram and tri-gram tokenization to be able to extract phrases rather than single words.

Logistic regression

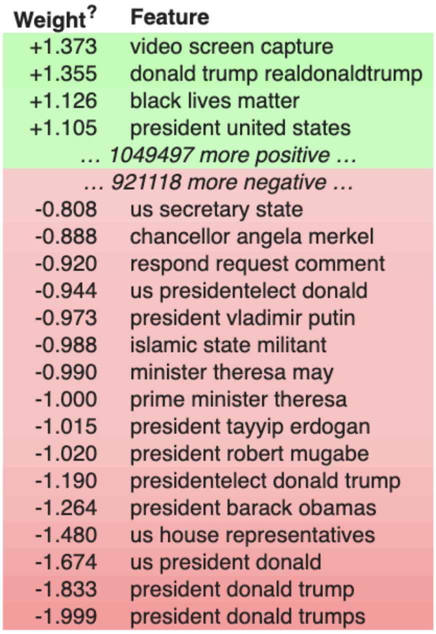
Once we have tokenized our text data, we use a logistic regression pipeline to train both uni-gram and tri-gram tokenized data. Given that once the model has been trained and tested, there is no further pre-processing required, we interpret the model's decision by extracting the learned model coefficients.

From the coefficients, we can determine which words and phrases were the most influential in making a model decision. In figure 4.1 and 4.2 we show the most positive and negative words for the model in the uni-gram and bi-gram case.



Weight?	Feature
+0.661	century
+0.592	america
+0.579	hillary
+0.533	press
+0.419	fox
+0.418	video
+0.416	isis
+0.411	watch
+0.407	american
+0.399	breitbart
… 65000 more positive …	
… 28587 more negative …	
-0.400	presidential
-0.416	minister
-0.418	ministry
-0.419	reporter
-0.423	clintons
-0.426	representatives
-0.431	spokesman
-0.461	obamas
-0.503	statement
-1.191	trumps

Figure 4.1: Top positive and negative features for uni-gram logistic regression

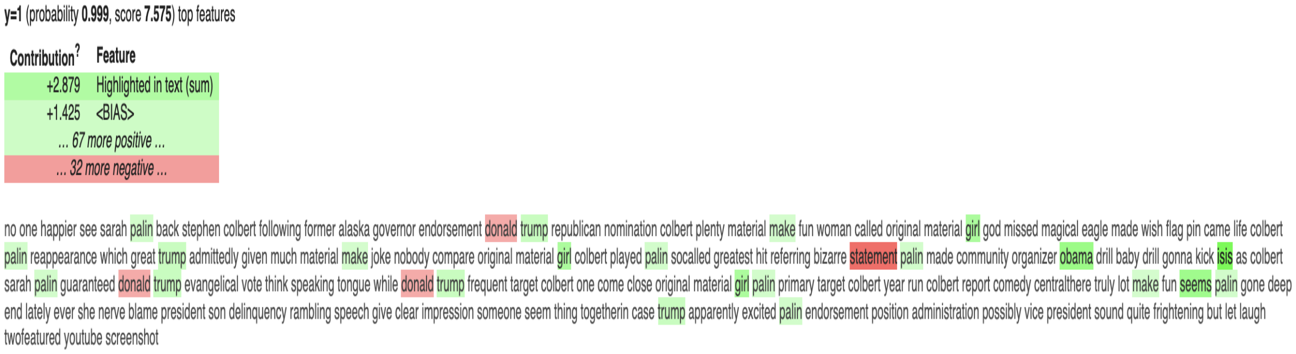


Weight?	Feature
+1.373	video screen capture
+1.355	donald trump realdonaldtrump
+1.126	black lives matter
+1.105	president united states
… 1049497 more positive …	
… 921118 more negative …	
-0.808	us secretary state
-0.888	chancellor angela merkel
-0.920	respond request comment
-0.944	us presidentelect donald
-0.973	president vladimir putin
-0.988	islamic state militant
-0.990	minister theresa may
-1.000	prime minister theresa
-1.015	president tayyip erdogan
-1.020	president robert mugabe
-1.190	presidentelect donald trump
-1.264	president barack obamas
-1.480	us house representatives
-1.674	us president donald
-1.833	president donald trump
-1.999	president donald trumps

Figure 4.2: Top positive and negative features for tri-gram logistic regression

We observe that words and phrases with president name's and ones that mention American politics negatively contribute to the model's decision.

We further perform local explanation for a single instance and compare the explanation between the tfidf and word2vec. In figure 4.3 and 4.4 we observe that the explanation for the instance is identical in both cases indicating that the logistic regression as an explainer is faithful; there is a slight difference in the probability and score and some words are highlighted darker in word2vec compared to Tfidf meaning they assigned more importance in the explanation.

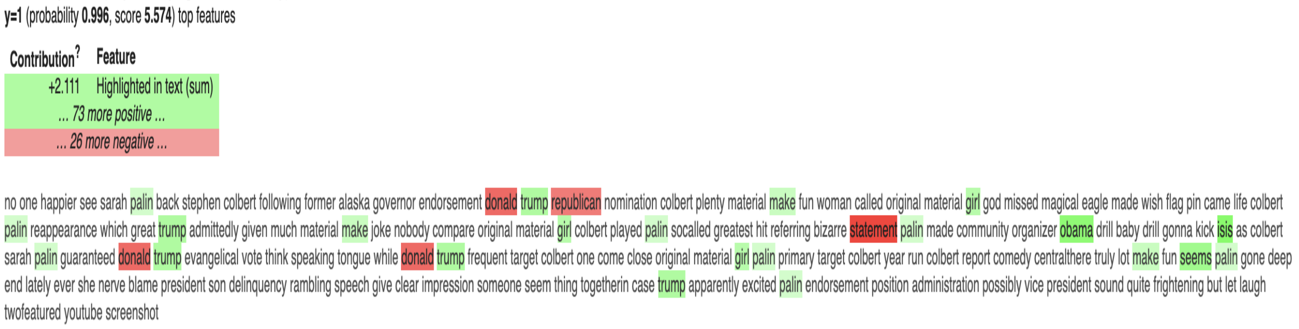


y=1 (probability **0.999**, score **7.575**) top features

Contribution?	Feature
+2.879	Highlighted in text (sum)
+1.425	<BIAS>
	... 67 more positive ...
	... 32 more negative ...

no one happier see sarah palin back stephen colbert following former alaska governor endorsement donald trump republican nomination colbert plenty material make fun woman called original material girl god missed magical eagle made wish flag pin came life colbert palin reappearance which great trump admittedly given much material make joke nobody compare original material girl colbert played palin socalled greatest hit referring bizarre statement palin made community organizer obama drill baby drill gonna kick isis as colbert sarah palin guaranteed donald trump evangelical vote think speaking tongue while donald trump frequent target colbert one come close original material girl palin primary target colbert year run colbert report comedy centralthere truly lot make fun seems palin gone deep end lately ever she nerve blame president son delinquency rambling speech give clear impression someone seem thing togetherin case trump apparently excited palin endorsement position administration possibly vice president sound quite frightening but let laugh twofeatured youtube screenshot

Figure 4.3: Local explanation for Tfidf logistic regression model



y=1 (probability **0.996**, score **5.574**) top features

Contribution?	Feature
+2.111	Highlighted in text (sum)
	... 73 more positive ...
	... 26 more negative ...

no one happier see sarah palin back stephen colbert following former alaska governor endorsement donald trump republican nomination colbert plenty material make fun woman called original material girl god missed magical eagle made wish flag pin came life colbert palin reappearance which great trump admittedly given much material make joke nobody compare original material girl colbert played palin socalled greatest hit referring bizarre statement palin made community organizer obama drill baby drill gonna kick isis as colbert sarah palin guaranteed donald trump evangelical vote think speaking tongue while donald trump frequent target colbert one come close original material girl palin primary target colbert year run colbert report comedy centralthere truly lot make fun seems palin gone deep end lately ever she nerve blame president son delinquency rambling speech give clear impression someone seem thing togetherin case trump apparently excited palin endorsement position administration possibly vice president sound quite frightening but let laugh twofeatured youtube screenshot

Figure 4.4: Local explanation for word2vec logistic regression model

Decision Tree

With the decision tree, the model's output is a rule-based global explanation. During the training phase, we limit our trees to a depth of 5. The results of the decision tree can be shown in figure 4.5.

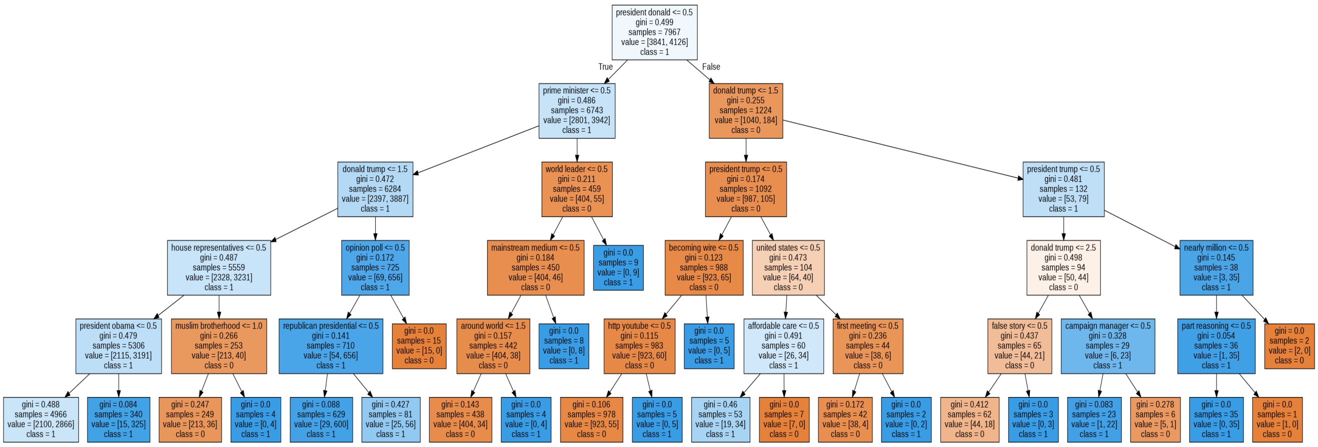


Figure 4.5: Local explanation for word2vec logistic regression model

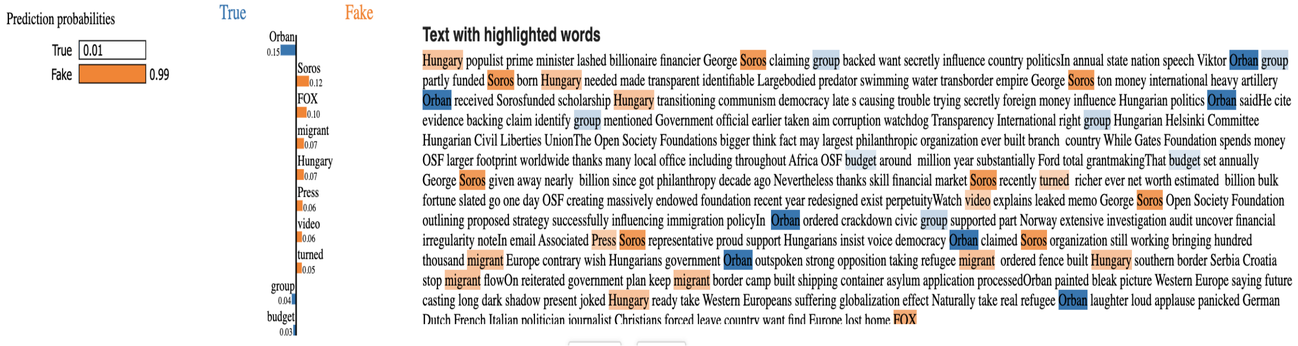
The phrase used to make decision rules in the model is also related to the presidency and politics from the USA, similar to the logistic regression. However, we also note that the decision rules for fake and true news classes are very similar and making it difficult to distinguish a clear decision rule between the classes..

4.2 LIME (Local interpretable model-agnostic explainer

We apply the LIME surrogate model as explained in chapter 2. We first train an LSTM (long short-term memory) network on our ISOT dataset to classify misinformation. Once our model has been trained and tested, we apply the LIME methodology to derive a local explanation for a single instance. The steps we followed are:

- Use word2vec embeddings on the cleaned dataset.
- Train LSTM
- We select an instance of interest x_{10} from our test dataset for an explanation, similar datapoint used across the techniques.
- Select perturbed points that are in the proximity of x_{10}
- We use LIME to generate explanations.

The output of the explanation can be seen in figure 4.6. The words with the highest weights for fake and real news are highlighted, with orange indicating importance for fake and blue real news. The darker the color, the more significant the feature is in explaining the outcome.



Text with highlighted words

Hungary populist prime minister lashed billionaire financier George Soros claiming group backed want secretly influence country politicsIn annual state nation speech Viktor Orban group partly funded Soros born Hungary needed made transparent identifiable Largebodied predator swimming water transborder empire George Soros ton money international heavy artillery Orban received Sorosfunded scholarship Hungary transitioning communism democracy late s causing trouble trying secretly foreign money influence Hungarian politics Orban saidHe cite evidence backing claim identify group mentioned Government official earlier taken aim corruption watchdog Transparency International right group Hungarian Helsinki Committee Hungarian Civil Liberties UnionThe Open Society Foundations bigger think fact may largest philanthropic organization ever built branch country While Gates Foundation spends money OSF larger footprint worldwide thanks many local office including throughout Africa OSF budget around million year substantially Ford total grantmakingThat budget set annually George Soros given away nearly billion since got philanthropy decade ago Nevertheless thanks skill financial market Soros recently turned richer ever net worth estimated billion bulk fortune slated go one day OSF creating massively endowed foundation recent year redesigned exist perpetuityWatch video explains leaked memo George Soros Open Society Foundation outlining proposed strategy successfully influencing immigration policyIn Orban ordered crackdown civic group supported part Norway extensive investigation audit uncover financial irregularity noteIn email Associated Press Soros representative proud support Hungarians insist voice democracy Orban claimed Soros organization still working bringing hundred thousand migrant Europe contrary wish Hungarians government Orban outspoken strong opposition taking refugee migrant ordered fence built Hungary southern border Serbia Croatia stop migrant flowOn reiterated government plan keep migrant border camp built shipping container asylum application processedOrban painted bleak picture Western Europe saying future casting long dark shadow present joked Hungary ready take Western Europeans suffering globalization effect Naturally take real refugee Orban laughter loud applause panicked German Dutch French Italian politician journalist Christians forced leave country want find Europe lost home FOX

Figure 4.6: Local explanation for word2vec LIME model

SP-LIME

We further use SP-LIME to derive a global explanation. We pick a representative sample in each class that applies LIME to each instance in the chosen samples to determine the most common explanation for each category. We observe that true news explanations are words with dates, facts, and neutral words; Fake news explanations are usually related to the government and are politically connected. The results of SP-LIME can be seen in Appendix C.

4.3 SHAP Values

Given that LIME weights cannot be trustworthy in determining the importance of a feature [33], we apply SHAP values as described in chapter 2. We observe the outcome in figure 4.7, which shows the top 20 features based on their importance. The weight of each variable based on the average of its combinations across the feature space is the x-axis and shows whether the variable positively or negatively affects the predicted outcome.

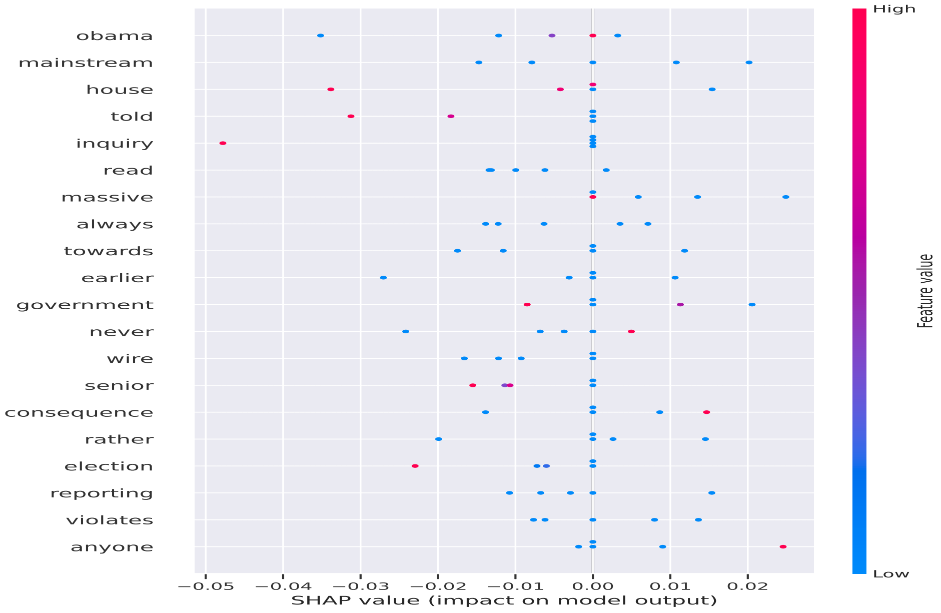
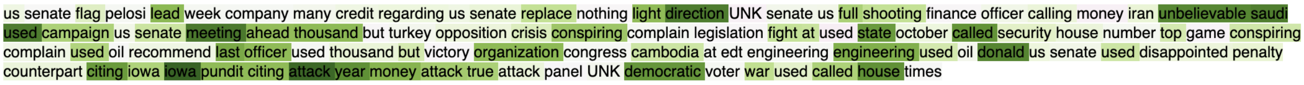


Figure 4.7: SHAP explainer on ISOT data

4.4 Attribution values: Integrated gradients

Lastly, we applied attribution methods to observe the outcome of the explanation technique. We used the same LSTM model applied in LIME; however, for integrated gradients, we create a baseline, empty embedding, and calculate the gradient between the baseline and the input. In figure 4.8 we observe the results of an explanation for class label 1 (fake news). It shows the words or features with the most substantial attribution score, with darker colors indicating more significant attribution scores compared to the lighter color.



us senate flag pelosi lead week company many credit regarding us senate replace nothing light direction UNK senate us full shooting finance officer calling money iran unbelievable saudi used campaign us senate meeting ahead thousand but turkey opposition crisis conspiring complain legislation fight at used state october called security house number top game conspiring complain used oil recommend last officer used thousand but victory organization congress cambodia at edt engineering engineering used oil donald us senate used disappointed penalty counterpart citing iowa iowa pundit citing attack year money attack true attack panel UNK democratic voter war used called house times

Figure 4.8: Attribution map using integrated gradients explanation for class label 0

4.5 Quantitative results of trained models

In table 4.1 we show the accuracy score of the models trained, indicating how well the models can predict unseen fake news data. As expected, we observe low accuracy scores for the Logistic regression and Decision tree model, models which are intrinsically interpretable as described in 2.2. When interpretability is achieved in black-box models through surrogate and model-agnostic techniques (LIME, SHAP, and Integrated gradients), we see the model's performance is not compromised. This shows the trade-off between accuracy and intrinsic explainable models.

Table 4.1: Table showing the accuracy values of the trained models to achieved interpretability in section 4.

Logistic Regression	**Decision Tree**	**LIME LSTM**	**SHAP LSTM**	**Integrated gradient LSTM**
89%	62%	96%	96%	96%

4.6 Summary

Majority of the results of the explanations discussed in this section output attribution/saliency maps highlighting the most important words(features) that are important in making the decision. However, although we found some explanation techniques to be faithful to a lay-user, the words may not be sufficient to distinguish and interpret between the two classes as we find fake news class to have similar words to true news. We thus try to use an example based technique in conjunction with one of the techniques to try and identify any key characteristics between the two classes.

We further observe the trade-off in accuracy and interpretability in table 4.1 as the accuracy values of intrinsic interpretable models are far less desirable than that of model-agnostic techniques.

Chapter 5

Methodology

In chapter 4 we discovered that most interpretable model outputs are attribution maps; we aim to use example-based techniques such that the explanations are more humanly understandable. We use four techniques that are commonly used on image data. First, we scale the techniques to our misinformation data to determine if we cannot extract any characteristics that could be used to interpret between the two classes. Further, we choose the best technique by analysing the quality of the examples produced. In this section, we discuss the methodology we used.

5.1 Experimental Design

The four techniques explored are Clustering, Set-Cover, Influential instances, and MMD-Critic. The techniques can be summarised as prototypical techniques. Prototype methodologies output a subset of the dataset. The subset selected is considered the best representation of the dataset. These subsets are then used to explain the data class before model prediction or after; explanation can be achieved by extracting the commonality and characteristics of each prototype class. Further, they can also be used to understand and investigate complex data [21]. For image data, prototypical examples are easy to digest and distinguish the characteristics that make up a class; however, with text data, it may not be and may be overwhelming to explain. Because of this, we further process the text prototypes extracted using text analytics techniques and employ one of the model-agnostic techniques described earlier. The design steps we follow are described high-level in figure 5.1.

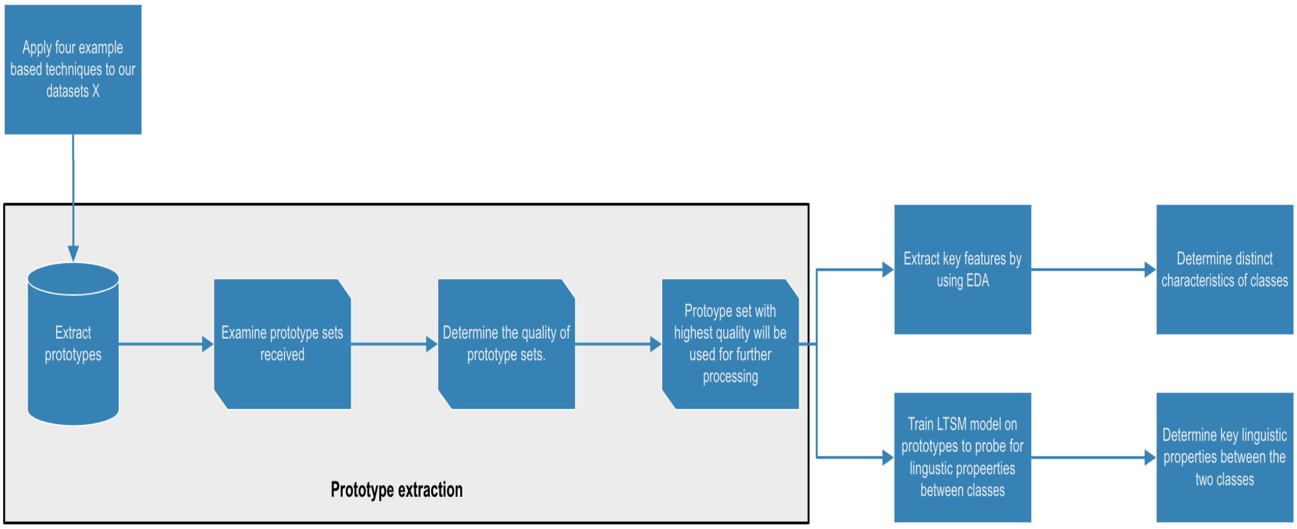

Apply four example based techniques to our datasets X
Extract prototypes
Examine prototype sets received
Determine the quality of prototype sets.
Protoype set with highest quality will be used for further processing
Prototype extraction
Extract key features by using EDA
Determine distinct characteristics of classes
Train LTSM model on prototypes to probe for lingustic propeerties between classes
Determine key linguistic properties between the two classes

Figure 5.1: High-level experimental design steps taken to achieve results

5.2 Algorithms

5.2.1 K-nearest neighbour Clustering Algorithm

Clustering is one of the earliest techniques, for example-based methods. The approach splits the data into clusters or classes based on their proximity to each other. A new instance x_i is assigned to the cluster to which it has the closest proximity based on K centroids.The advantage of the clustering method is that it can be used both as a method to extract prototypes and as an intrinsic interpreter in its own right; however, the disadvantage is that it can not achieve interpretability on a modular level(global explainability) just on a local level.

For text data, the similarity score between documents is used to determine the clusters of the dataset. We utilise the K-nearest neighbour clustering technique to determine the set of prototypes for each class. We show that running the algorithm on each class respectively, as stated in [9], yields prototype sets with the K instances chosen for an instance prediction being the set. Given that interpretability cannot be achieved on a modular level, the prototype sets are acquired through the prediction phase of test instances.

Given a dataset X with class label Y, where $y_i \in 1, 0$ we split the data X into X^{train} and X^{test}. For each $x \in X^{test}$ we determine which K documents in X^{train} are closest to the instance x; the K selected documents are the prototype sets for the selected instance. We repeat for all instances in X^{test} and obtain a list of documents K, which we make our set of prototypes. The algorithm can be seen in algorithm 5.1 below.

```
Start with an empty set $K = \emptyset$, given dataset $X^{train}$ and instance $x \in X^{test}$
while $K \leq 10$:
        Find $x_i^{train}$ which is the most similar to $x$
        $K \cup x_i^{train}$
        end while
```

Algorithm 5.1: Clustering Prototype selection method

5.2.2 Set Cover Algorithm

[9] proposed the set cover approach to select prototypes as an improvement to the K-clustering techniques. As the name suggests, the methodology leverages the set cover theorem, which states that given a set X what are the optimal subsets S which cover the universal set X by minimising the cost. Compared to the above clustering technique, the set cover approach automatically selects the number of prototypes for each class and can achieve both modular and local interpretability

The selection of prototypes using the technique can be made via two approaches the integer program and the greedy approach. We make use of the greedy approach.
An instant $x_i \in X$ is selected as a prototype if it encompasses the following characteristics:

- In its sphere, it should cover more training points that have the same class as itself.
- It should cover fewer instances of the opposite class.
- It must be sparse

To determine the prototype set P_l where P_l is the prototype set of class l. We require a sphere size dependent on a radius value; the sphere and radius are used to determine each instance's coverage. The size of the sphere is calculated by using a pairwise distance formula; we want to minimise the cost of covering a new instance.

Below in figure 5.2 we show an image of the selected prototypes adapted from [15] which shows the greedy algorithm for a 2-class toy problem using a radius of 1.5. The prototypes chosen for each class are bolded as their coverings include the most instances which belong to the same class and less of the other class.

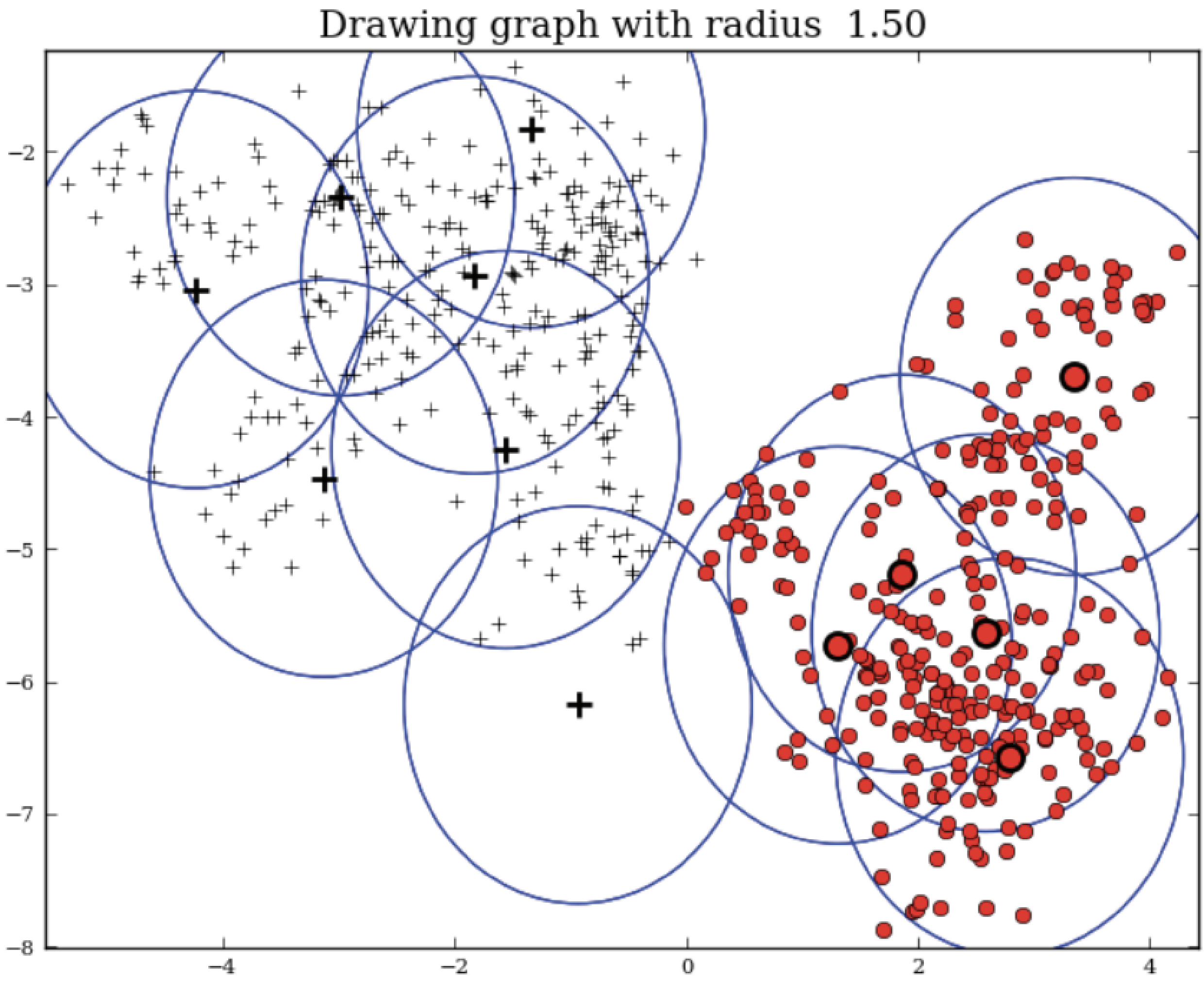


Figure 5.2: Image showing Set Cover example for a classification toy example

5.2.3 Influential Instances

Influential instances is an example-based technique that does not necessarily output a set of instances but calculates what influence each instant has on the model parameters. An influential instance changes the model parameters and decision when deleted; [26] states that the counterfactual question that is asked is: ”how would the model’s predictions change if we did not have this training point?”.The method is used mainly for debugging and to explain behaviour of the model.

There are two approaches one can take to determine influential instances:

- Deletion: In this approach, one needs to iterate through the training data and delete each instance; once deleted, analyse how the model parameters change by retraining the model.
- Calculation of influence function: The influence function determines which instances are the most influential to the model by up-weighting the instance by a small amount and determines the change in the model parameters.

We use the calculation of influence function as it is known to be more robust [3, 26]; further, the approach also has the benefit of highlighting which instances are the most harmful to the prediction of a particular test instance. To use the influential function, one needs to train a model with a loss function that is twice differentiable and a loss gradient that is accessible with respect to model parameters. The influence score calculation is based on the loss function.

Given a dataset X, which maps to a label set Y, we split the dataset into training and test points. With $x_1, x_2, ..., x_n \in X^{train}$ and $y_1, y_2, ..., y_n \in Y^{train}$ we state that $z_i = (x_i, y_i)$ as being the training point for instance i. The loss function is then defined as $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$, where θ is the model parameters. By up-weighting an instance z_i by a small amount ϵ we get a new loss function: $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) + \epsilon L(z, \theta)$.

To determine the influence measure, which measures the change of up-weighting a training instance, z_i by ϵ, the gradient of the loss and inverse of hessian is used as shown in the below equation 5.1:

$$I(z) = -H_{\hat{\theta}}^{-1} \nabla L(z, \hat{\theta}) \tag{5.1}$$

Where:
$H_\theta = \sum_{i=1}^{n} \nabla_\theta^2 L(z_i, \hat{\theta})$ is the Hessian of the loss function and twice differentiable.

We use the CNN network to train our data with cross-entropy loss. Then for each instant in the training data, we up-weight the loss function by a small amount($\frac{1}{n}$) and determine their influence score as described in equation 5.1. We then use the top 100 most influential instances for each class as our prototype sets.

5.2.4 MMD-Critic Algorithm

MMD-critic is a quantitative approach that uses maximum mean discrepancy, a greedy approach in selecting prototypes. It aims to select a subset of prototypes Z that best represent our dataset X such that the probability distribution of the data points in Z approximates that of X. This approach is one of the preferred approaches for selecting prototypes [33, 25] given that together with prototypes, it outputs critics, which are data points that least represent a class and has a high probability of being misclassified.

To get the prototype and critic sets, the following attributes are required:

1. **m* and n* number of prototypes and critics:**
   The number of prototypes and critics can be selected manually or alternatively use ϵ an error term, such that the algorithm stops searching once the MMD is above the threshold of the error.

2. **Kernel function:**

   We use the kernel function in the MMD formula to determine the distance between two points. The kernel ranges between 0 an 1. Values close to 0 indicate the points are infinitely apart, while a value closer to 1 indicates similar points. The Euclidean distance is used in our kernel, and we use the radial basis kernel, a popular kernel used across the literature.

3. **MMD:**

The maximum mean discrepancy function is generally used to determine the ”closeness” of two distributions. We use it to determine how well our prototype distribution approximates the data i.e., we will add data points that provide the lowest MMD compared to other samples in the data. The equation of the MMD is shown in equation 5.2:

$$MMD^2 = \frac{1}{m^2}\sum_{i,j=1}^{m} k(z_i, z_j) - \frac{z}{mn}\sum_{i,j=1}^{m,n} k(z_i, x_j) + \frac{1}{n^2}\sum_{i,j}^{n} k(x_i, xj) \tag{5.2}$$

Where:
z_i is the i^{th} data point in prototype sample set.
x_i is the i^{th} data point in data set
k=Kernal measure
m= number of prototypes and
n= number of data points in the dataset

The goal is to optimize the function so that the function’s middle term approximates to zero.

4. **Witness Function:**

   The witness function is only used once all prototypes are selected. The purpose of the function is to search for a subset of data points n, which are not represented well by the data (outliers). It measures the proximity of each remaining data point to the dataset as well as its proximity to the chosen prototype set. If the witness function value is close to zero, it means the instance is not a critic, but instances with a large positive or negative witness value are a critic. Having a large positive and negative witness function is an indication that the instance either underestimates or overestimates the data distribution; the equation is shown in equation 5.3

$$W = \frac{1}{n}\sum_{i}^{n} k(x, x_i) - \frac{1}{m}\sum_{j=1}^{m} k(x, z_j) \tag{5.3}$$

   where:
   z_j is the j^{th} data point in prototype set.
   x_i is the i^{th} data point in data set

The algorithm of using the MMD-critic approach for selecting prototypes and critics is described in algorithm 5.2 and 5.3.

```
Start with an empty set $Z = \emptyset$, given dataset $X$
while $Z$ $\leq$m:
          Find $x_i$ which reduces the $MMD^2$
         for all $x_i$ in the $X$ do
                  Calculate the $MMD^2$
                  Select $x_i$ which optimises $MMD^2$
                   $Z$ $\cup x_i$
         end while
```

Algorithm 5.2: MMD^2 Prototype selection algorithm

```
Start with an empty set $C = \emptyset$, given dataset $X \setminus Z$
while $C$ $\leq$n:
          Find $x_i$ with largest $|W|$
         for all $x_i$ in the $X \setminus Z$ do
                  Calculate the $W$
                  Select $x_i$ with largest value
                   $C$ $\cup x_i$
         end while
```

Algorithm 5.3: Critic selection algorithm

5.3 Prototype Analysis

In this section, we describe how we analyse the prototypes extracted from the data to determine characteristics that can be used to associate and explain the classes.

Prototype Quality

To determine the quality of the prototypes, we use the trained LSTM model on unseen data which was not selected as part of the prototypes. The accuracy score of how well the model predicted will help us understand the quality of chosen prototypes.

Determining class characteristics from prototypes

To determine class characteristics from prototypes, we apply various techniques. First, we extract top key phrases from each class using Tfidf and Word2Vec. We also determine the sentiment and perform opinion detection on the extracted prototypes. Further, we perform a comparative analysis on the part of speech used, word length, and special characters to determine any significant character differences between the two classes.

Using SP LIME in conjunction with prototypes

We employ the global explainer SP-LIME to determine how the model-agnostic technique would explain prototype classes, Such that we can have an explainer that is truthful, portable, and can be understood by a human to a degree.

5.4 Summary

This chapter discusses the approaches taken to extract, analyse and explain the prototypical examples. We highlighted the algorithm used and the steps that were taken to extract the necessary results, which we discuss in chapter 6.

Chapter 6

Results

In this chapter we discuss the results that we retrieve from the data and analysis performed as described in chapter 5.

6.1 Prototypes Selected

We apply the algorithms described in our methodology section to each of our datasets; ISOT, COVID, and SA data. We analyse the prototypes selected for each data using the algorithms and determine how common the sets are to each other.

6.1.1 Clustering Sets

We use our test data to acquire our prototype set for clustering. For each instance in the test data, we select the top 20 documents it is similar to; we then remove any duplicate indices chosen for more than one test indices to get a unique subset. After applying the approach to our datasets, we find that for all datasets, the fake news set is larger than the true news set by at least a margin of 12, as shown in table 6.1. This may indicate that fake news has more variety compared to true news data. In table 6.2 - 6.4, we also observe that the prototype sets chosen for ISOT and COVID match the Set Cover sets highly 43% and 47%, while SA has a match rate of 51% with MMD-critic.

When looking at the phrases extracted and word cloud for each set figure 6.1 - 6.6 , we observe that the sets are closely related to each other and may be difficult to distinguish between the two classes. This is especially the case for ISOT classes; however, for Covid and SA, we can see that fake news phrases are centered around controversial topics and show subjectivity.

Data Set	Size of True Set	Size of Fake Set
ISOT	29	43
COVID	24	33
SA	20	32

Table 6.1: Table showing the size of each prototype set using clustering technique for our dataset

	Clustering Set	Set Cover	Influential	MMD-Critic
Clustering Set	100%	43%	21%	19%
Set Cover	43%	100%	41%	31%
Influential	21%	41%	100%	51%
MMD-Critic	19%	31%	51%	100%

Table 6.2: Table showing the match rate of the prototype sets for ISOT data

	Clustering Set	Set Cover	Influential	MMD-Critic
Clustering Set	100%	47%	33%	22%
Set Cover	47%	100%	44%	29%
Influential	33%	44%	100%	56%
MMD-Critic	22%	29%	56%	100%

Table 6.3: Table showing the match rate of the prototype sets for Covid data

	Clustering Set	Set Cover	Influential	MMD-Critic
Clustering Set	100%	42%	41%	51%
Set Cover	42%	100%	49%	33%
Influential	41%	49%	100%	52%
MMD-Critic	51%	33%	52%	100%

Table 6.4: Table showing the match rate of the prototype sets for SA data

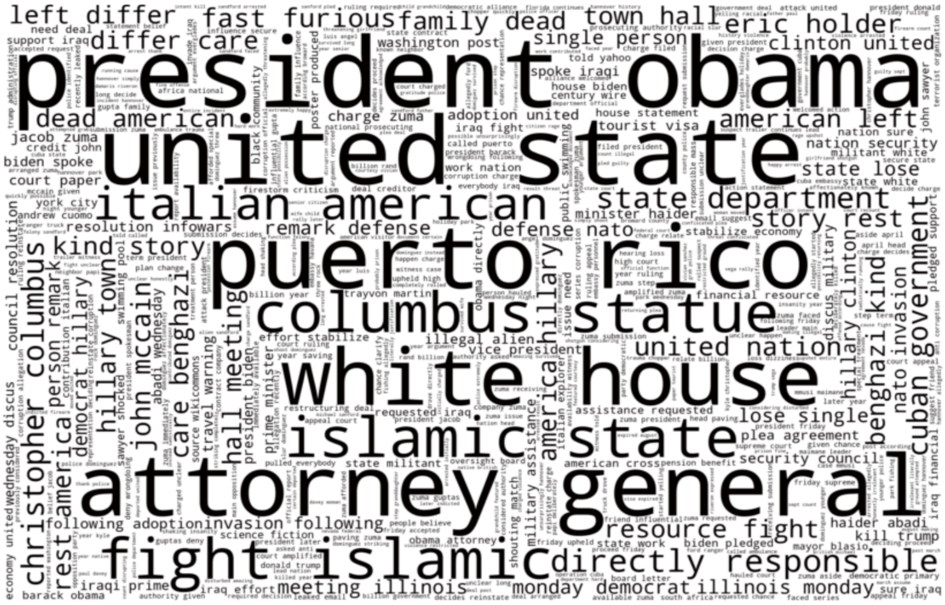


Figure 6.1: Word Cloud class 1(Fake)for ISOT

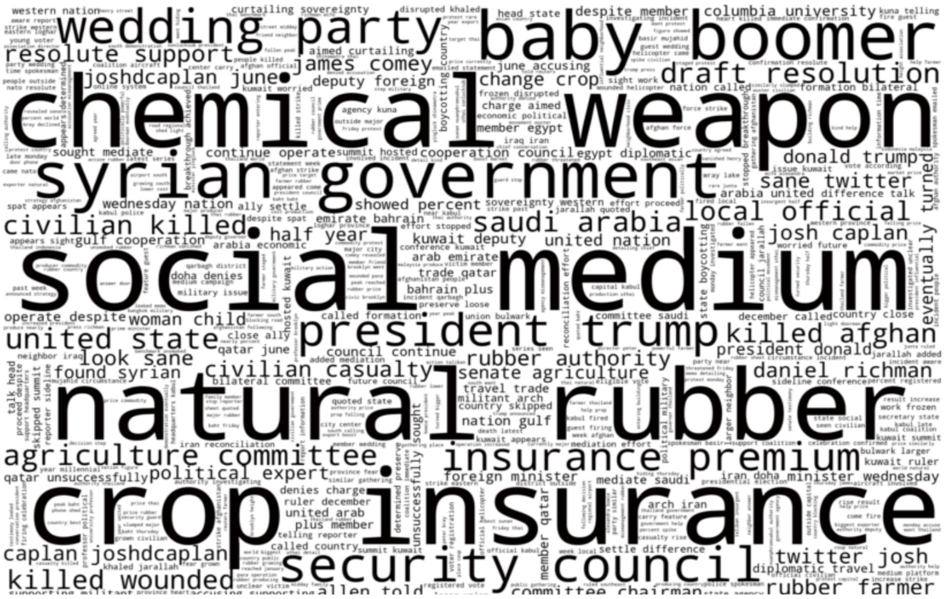


Figure 6.2: Word Cloud class 0(True) for ISOT.

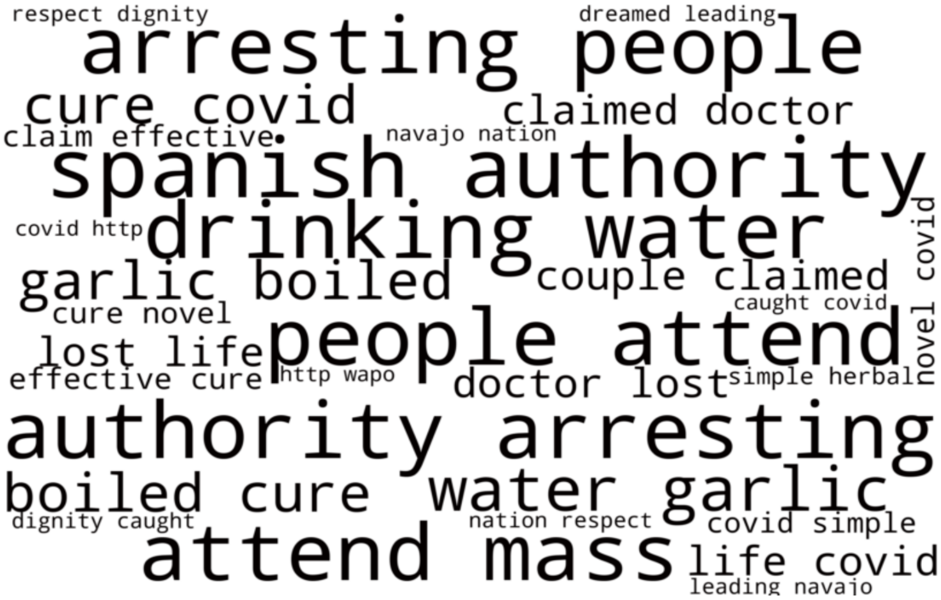


Figure 6.3: Word Cloud class 1(Fake)for COVID Data

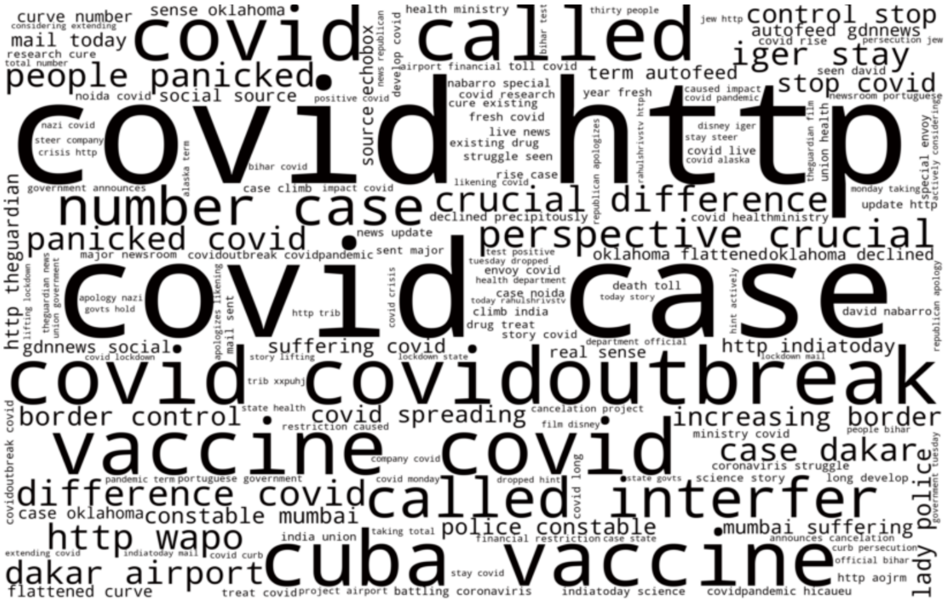


Figure 6.4: Word Cloud class 0(True) for COVID Data

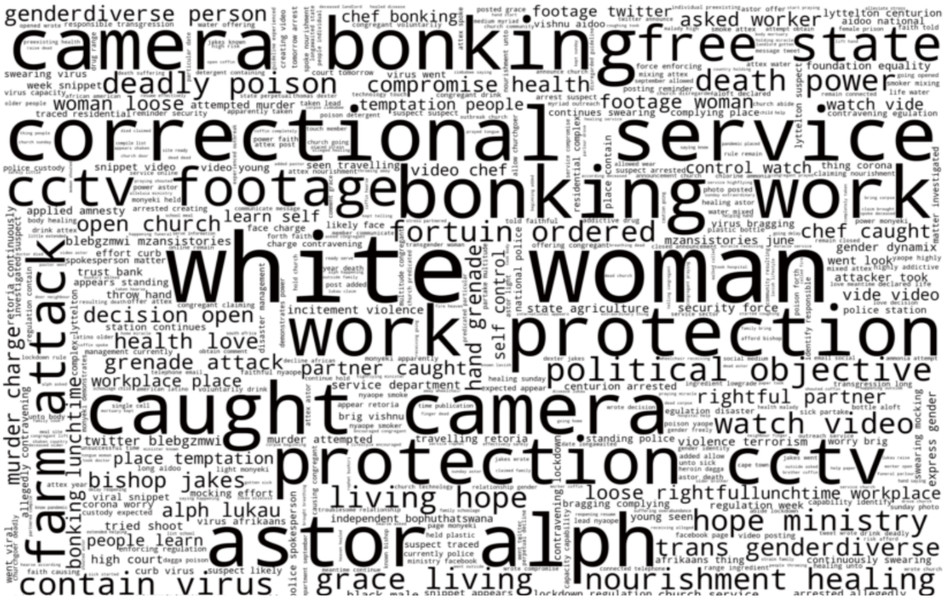


Figure 6.5: Word Cloud class 1(Fake)for SA Data

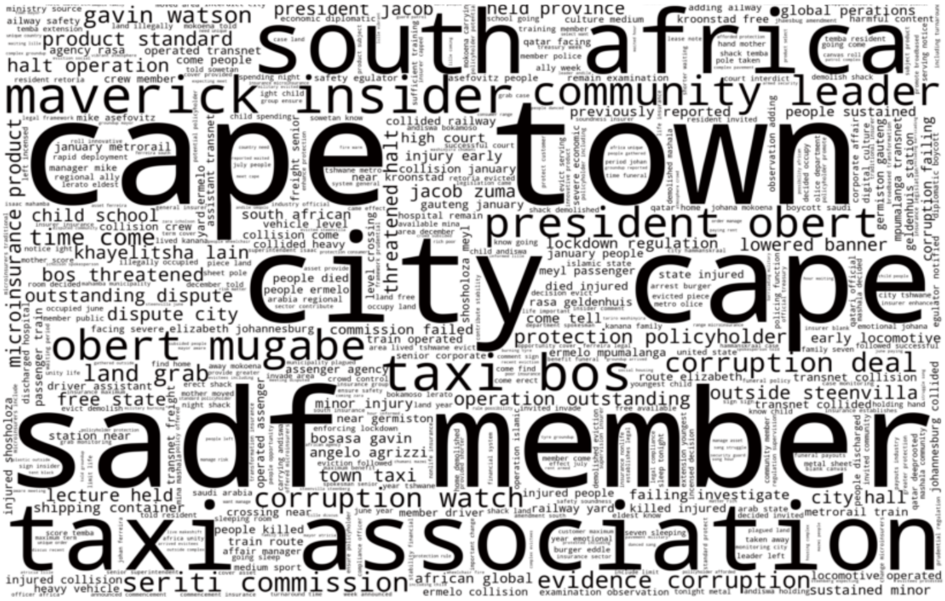


Figure 6.6: Word Cloud class 0(True) for SA Data

6.1.2 Set Cover Sets

Set Cover has a larger set of prototypes selected compared to clustering, as shown in table 6.5. Unlike clustering, we observe that the prototype sets for Set Cover are slightly larger for true news compared to fake, but given that the variance is smaller between the classes, it indicates similar variation.

When observing the matching rate, table 6.2-6.4, of each set we observe that ISOT and COVID sets have a higher match rate with clustering sets whilst SA set has higher rate with Influential instances set. Looking at the word cloud phrases in figure 6.7-6.12, similar to clustering, the top phrases can be difficult to distinguish between classes; however, for SA and COVID fake news word cloud we notice phrases such as farm attack and farm murderer and around covid cures which are controversial and subjective compared to true news phrases.

Data Set	Size of True Set	Size of Fake Set
ISOT	230	225
COVID	174	171
SA	123	119

Table 6.5: Table showing the size of each prototype set using Set Cover technique for our datasets

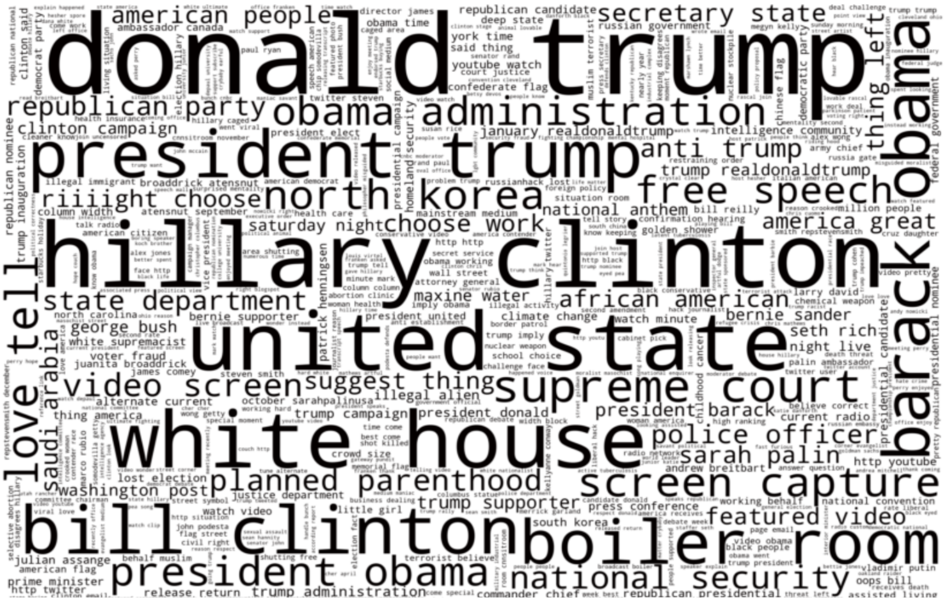


Figure 6.7: Word Cloud ISOT prototype output for class 1(Fake) using Set Cover

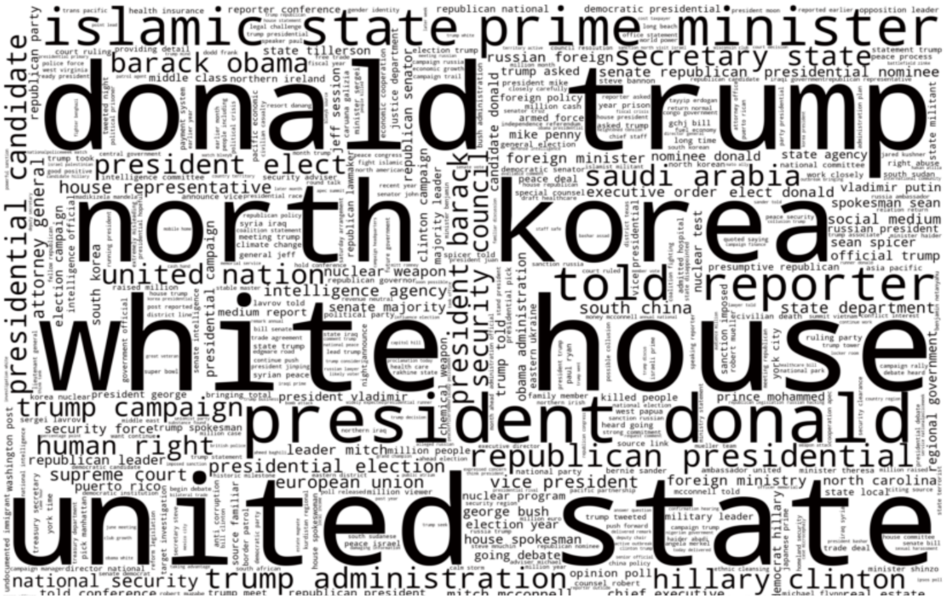


Figure 6.8: Word Cloud ISOT prototype output for class 0(True) using Set Cover

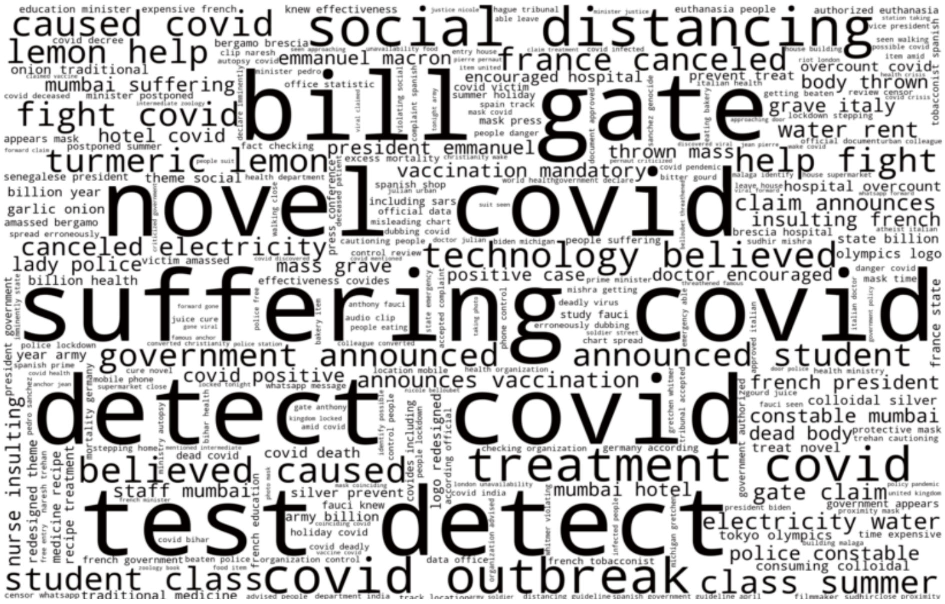


Figure 6.9: Word Cloud Covid prototype output for class 1(Fake) using Set Cover

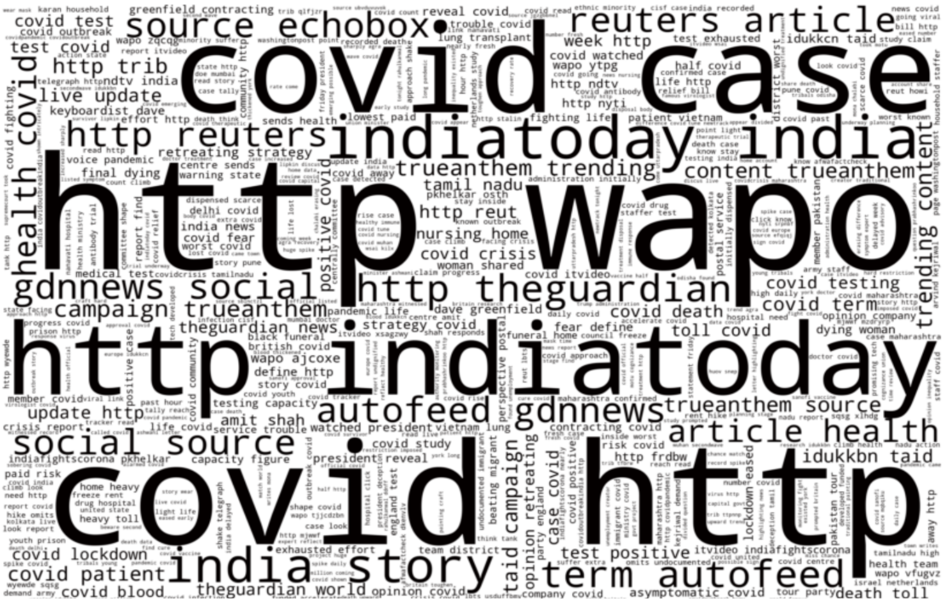


Figure 6.10: Word Cloud Covid prototype output for class 0(True) using Set Cover

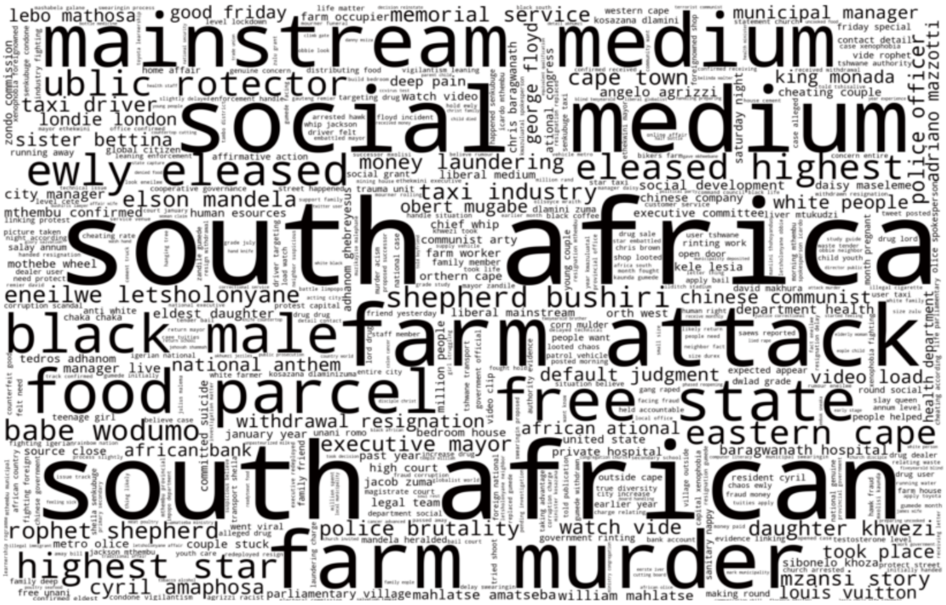


Figure 6.11: Word Cloud SA prototype output for class 1(Fake) using Set Cover

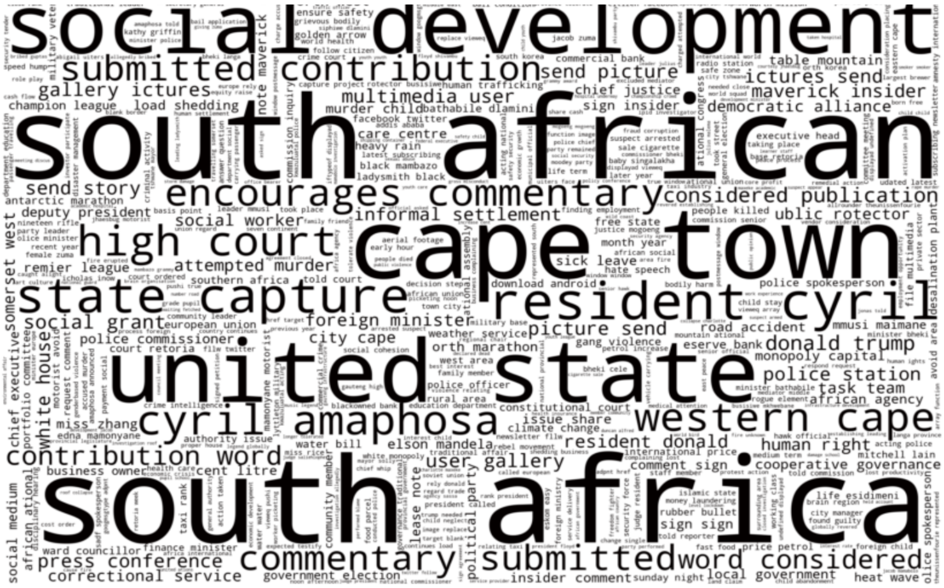


Figure 6.12: Word Cloud SA prototype output for class 0(True) using Set Cover

6.1.3 Influential function sets

For influential instances, the prototype sets are not automatically chosen. The algorithm will tell us which instances are the most influential and harmful. We then choose the top 100 instances, which are the most influential, and the top 20 harmful instances. From the top 100 influential examples, we show the split per class in table 6.6 and observe that the true news class makes up 47% of the set and fake news 53% for ISOT and SA data, similarly for COVID set we find true news is 44% and fake news 56%. We observe that the top 100 influential instances have a higher match rate with MMD-critic sets.

Data Set	Size of True Set	Size of Fake Set
ISOT	47	53
COVID	44	56
SA	47	53

Table 6.6: Table showing the size of each prototype set using Influential technique for our dataset

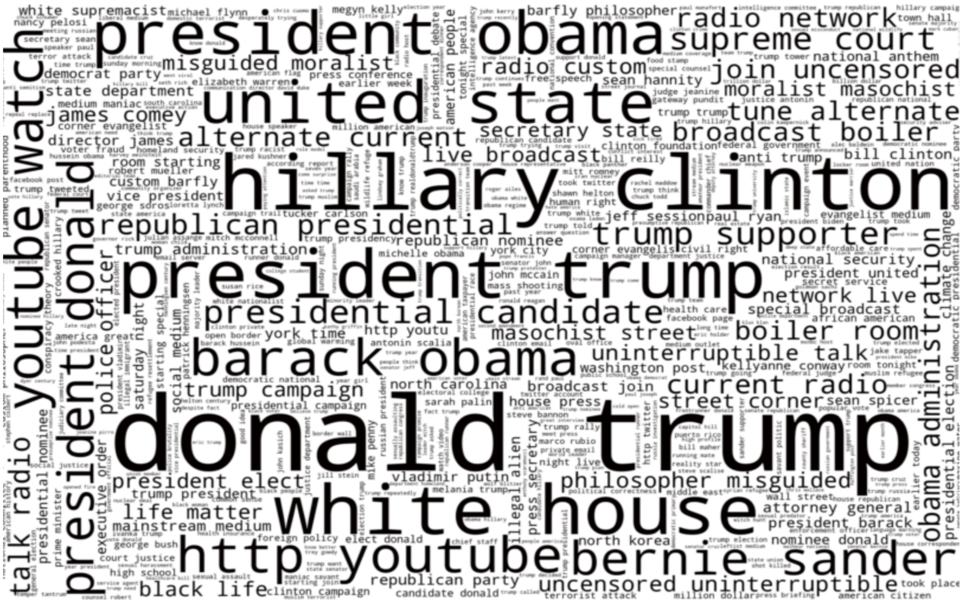


Figure 6.13: Word Cloud ISOT prototype output for class 1(Fake) using Influential instances

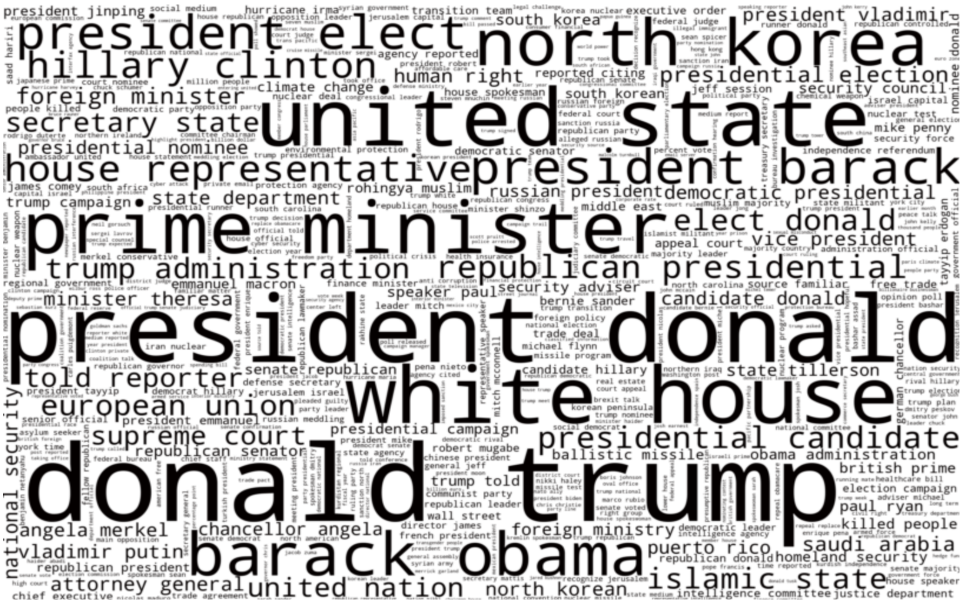


Figure 6.14: Word Cloud ISOT prototype output for class 0(True) using Influential instances

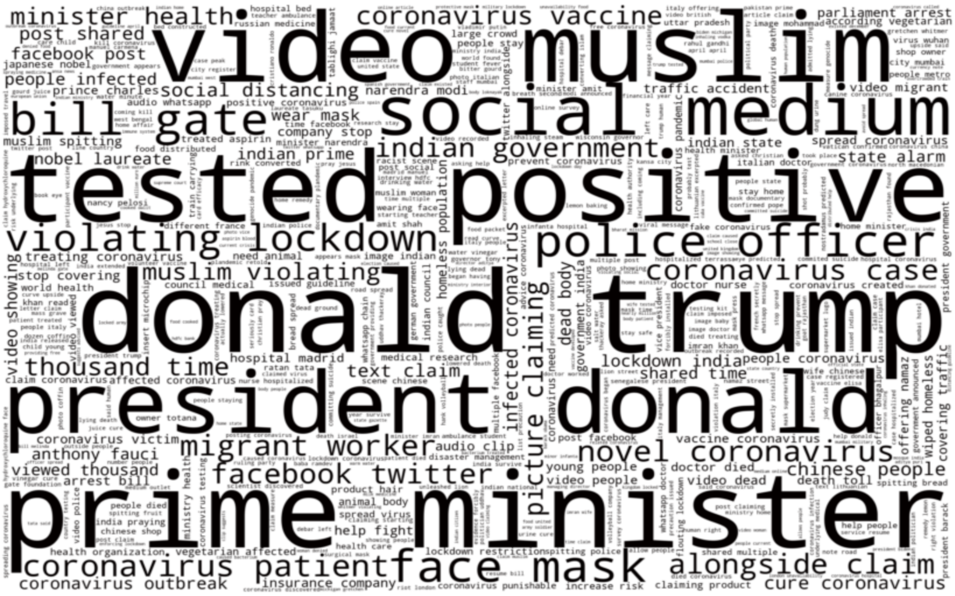


Figure 6.15: Word Cloud Covid prototype output for class 1(Fake) using Influential instances

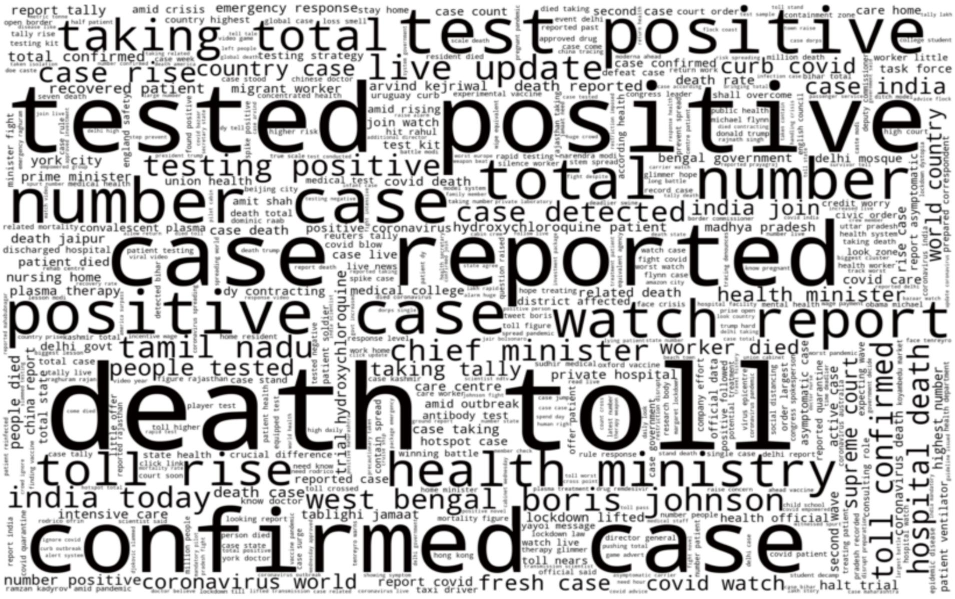


Figure 6.16: Word Cloud Covid prototype output for class 0(True) using Influential instances

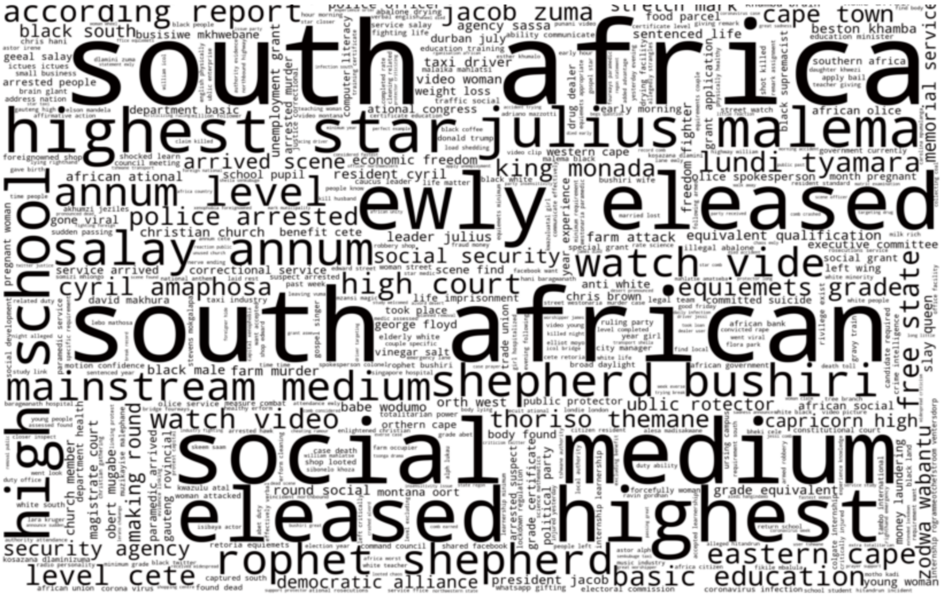


Figure 6.17: Word Cloud SA prototype output for class 1(Fake) using Influential instances

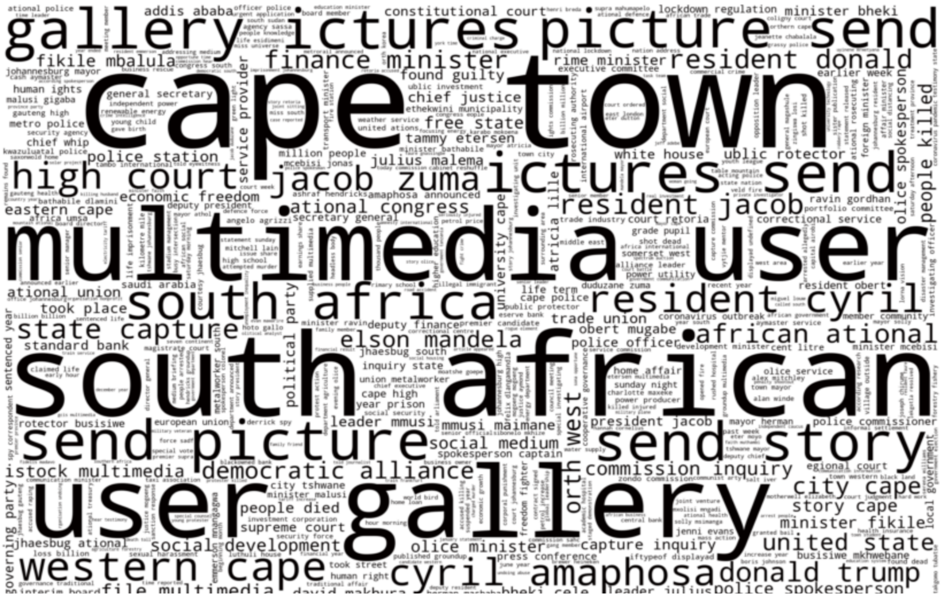


Figure 6.18: Word Cloud SA prototype output for class 0(True) using Influential instances

6.1.4 MMD-Critic Sets

The algorithm automatically selects MMD-Critic sets; we find that the size of the sets is larger for fake news compared to true news, shown in size split table 6.7. This indicates there is more variation in fake news compared to true news. We further select the top 10 data instances with a high witness score. Finally, we compare the critics to the harmful instances selected in influential instances and find that the chosen set is 57% match. It is observed in figure 6.19 to 6.24 that true phrases are more neutral and have a less aggressive tone to fake phrases as we can note that there are words such as farm attack, deliberately negligence, people dead in the fake news phrases.

Data Set	Size of True Set	Size of Fake Set
ISOT	54	113
COVID	49	92
SA	42	87

Table 6.7: Table showing the size of each prototype set using MMD-Critic technique for our dataset

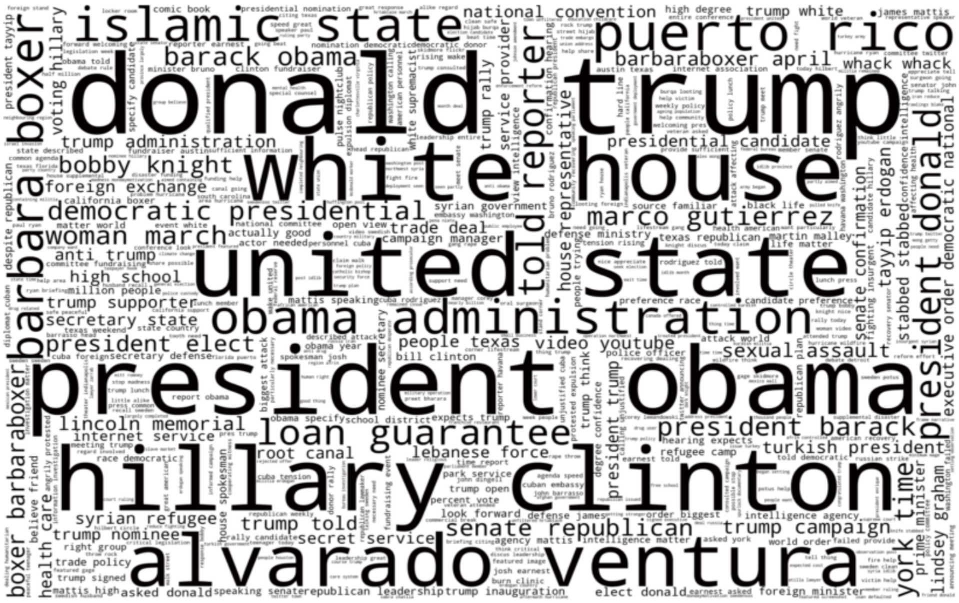


Figure 6.19: Word Cloud ISOT prototype output for class 1(Fake) using MMD-Critic technique

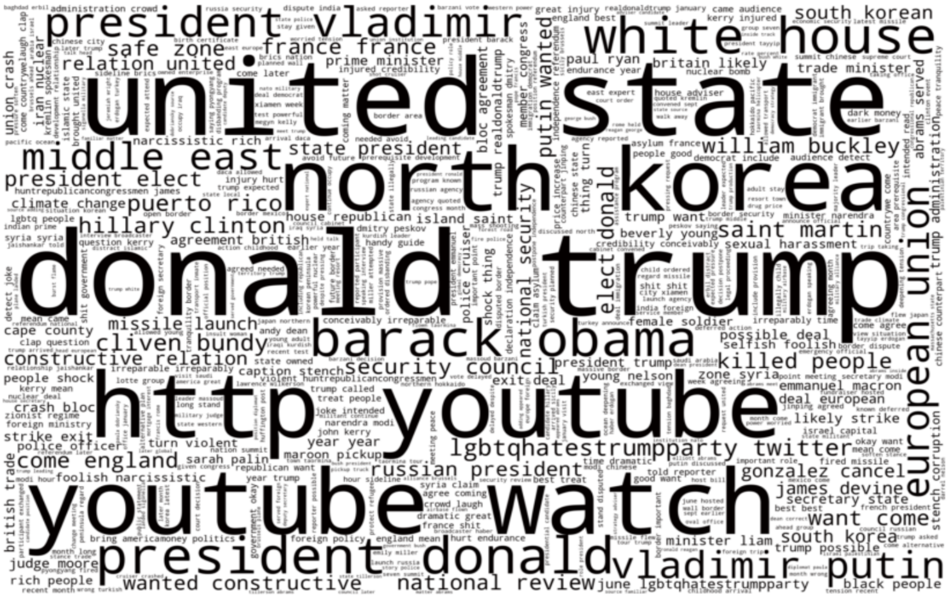


Figure 6.20: Word Cloud ISOT prototype output for class 0(True) using MMD-Critic technique

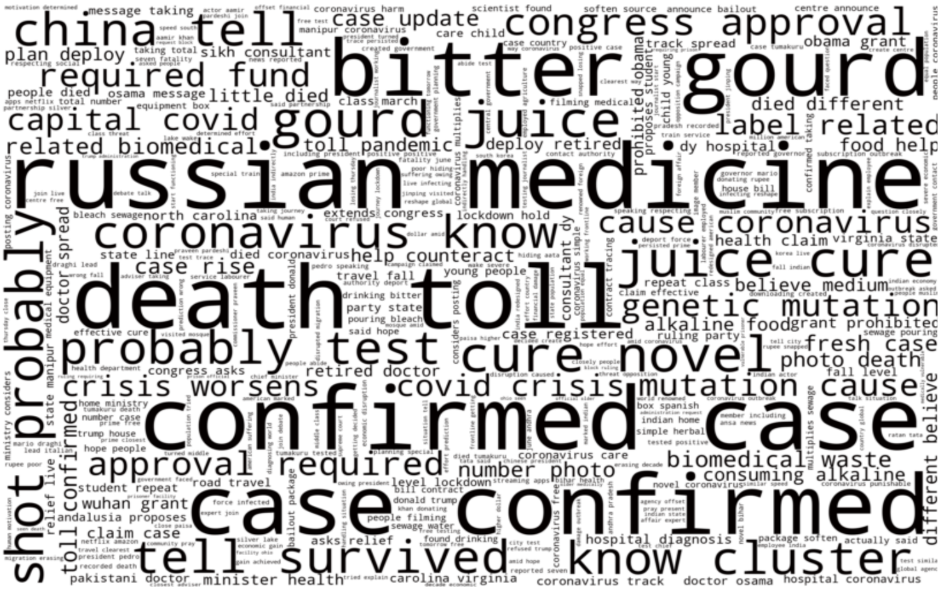


Figure 6.21: Word Cloud Covid prototype output for class 1(Fake) using MMD-Critic technique

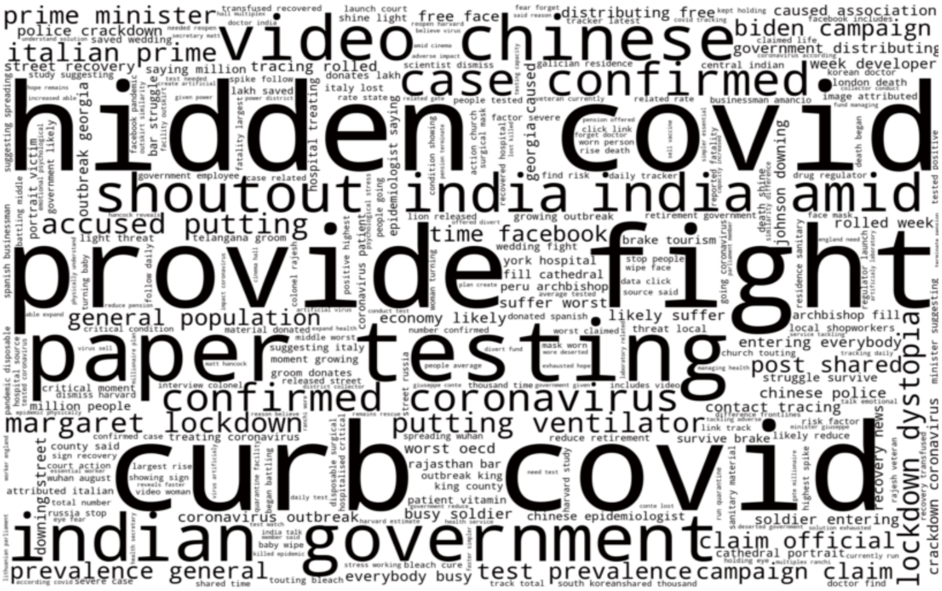


Figure 6.22: Word Cloud Covid prototype output for class 0(True) using MMD-Critic technique

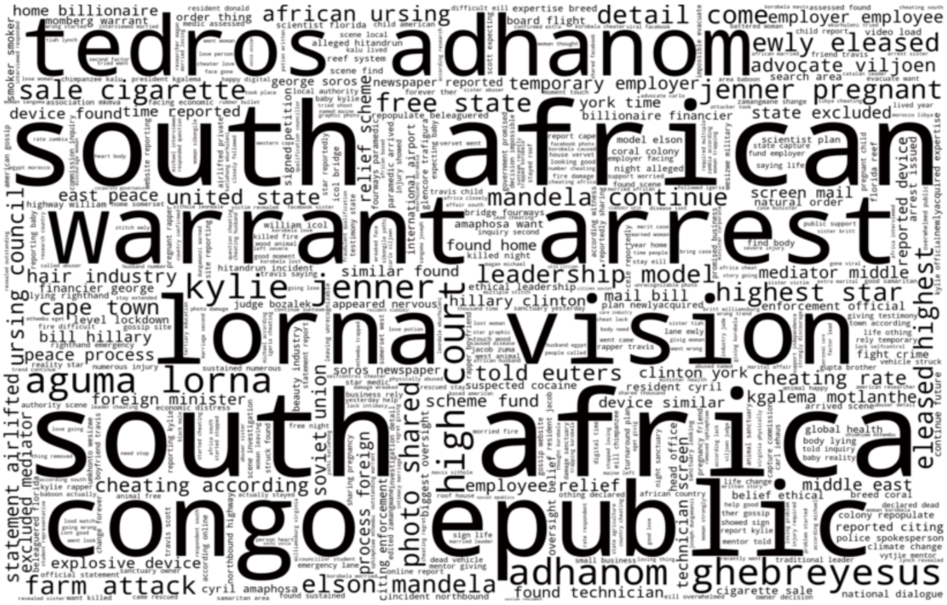


Figure 6.23: Word Cloud SA prototype output for class 1(Fake) using MMD-Critic technique

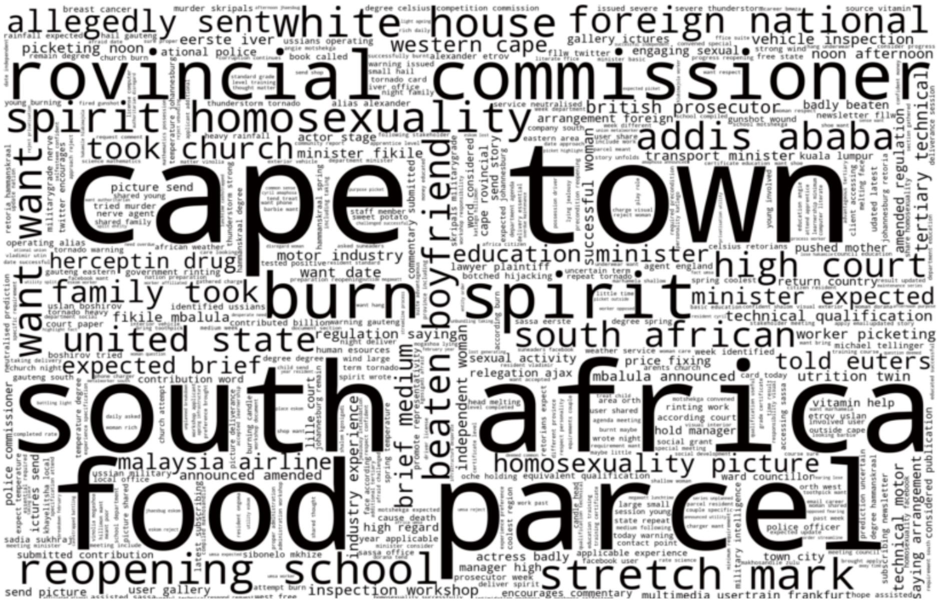


Figure 6.24: Word Cloud SA prototype output for class 0(True) using MMD-Critic technique

6.2 Quality of prototypes

We need to determine the quality of the prototypes to ensure that the information we extracted is truthful and portable to any use case related to misinformation classification. As described in [7, 8] to determine the quality of prototypes, we train a model with the prototype sets and determine the accuracy percentage when predicting unseen data. The prototype set with the highest accuracy has the highest quality. We train our prototype sets using an LTSM and observe the prediction accuracy in table 6.8.
Across the four example-based methods, we observe that the prototype sets with the highest accuracy are Influential and MMD-Critic; although the two methods have an accuracy of less than 70% we believe that if we manually select the prototype sets to be larger, the accuracy value would increase. Therefore, in the next section, we further process both the Influential and MMD-critic to determine key characteristics we can extract.

Data Set	Clustering	Set Cover	Influential	MMD-Critic
ISOT	36%	41%	67%	71%
COVID	33%	39%	56%	59%
SA	31%	34%	47%	53%

Table 6.8: Table showing the quality of each prototype set by observing prediction quality

6.3 Further Processing of Selected Sets.

In this section, we further analyse the two prototype sets with the highest quality measured by the prediction accuracy. First, we look at the top phrases of each set; we use them to perform sentiment, opinion detection, and word analysis. We further use a model agnostic explainer on the sets and, lastly, use a probing technique to determine linguistic properties.

6.3.1 Analysis of Set Phrases

As seen in Appendix D, the prototypical examples from text data are just a list of sentences where one can not distinguish the key characteristics between the classes. We take advantage of the SBERT functionality, keyBERT, which uses the embedding of the model to generate keywords and phrases from a document by calculating the cosine similarity of each generated key phrase to the document. We select the top 10 key phrases with the highest similarity score.

It is observed in figure 6.25 to 6.30 that true phrases are more neutral and have a less aggressive tone to fake phrases as we can note that there are words such as farm attack, deliberately negligence, people dead in the fake news phrases.

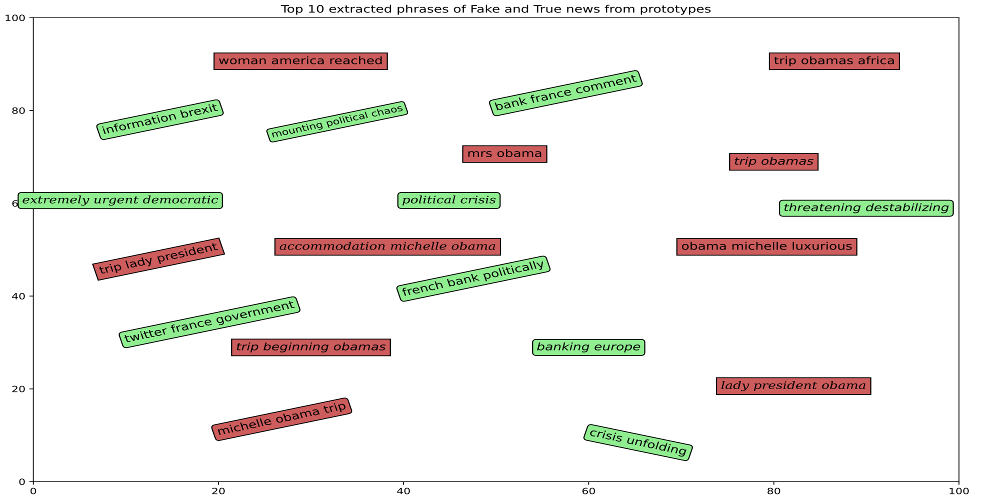


Figure 6.25: Figure showing top 10 phrases extracted from ISOT MMD prototypical set examples. Green phrases are true news phrases and red fake news phrases

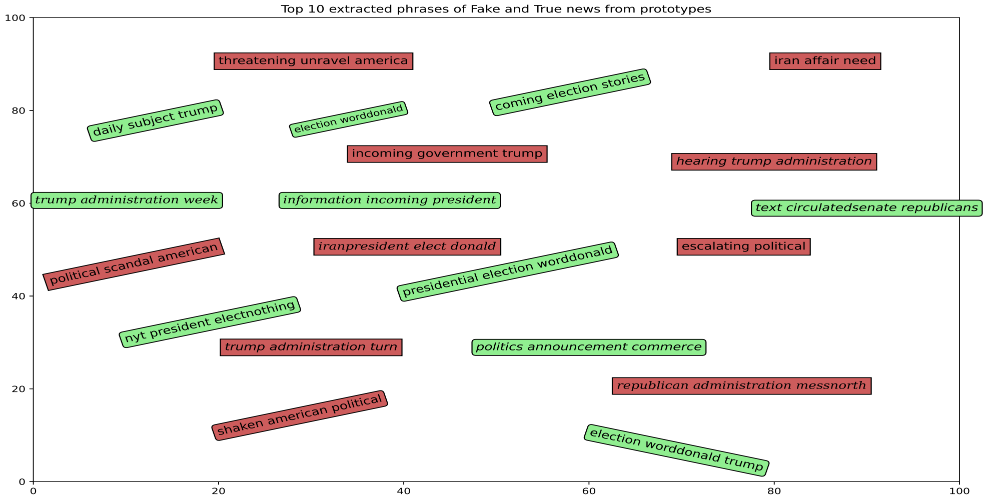


Figure 6.26: Figure showing top 10 phrases extracted from ISOT Influential prototypical set examples. Green phrases are true news phrases and red fake news phrases

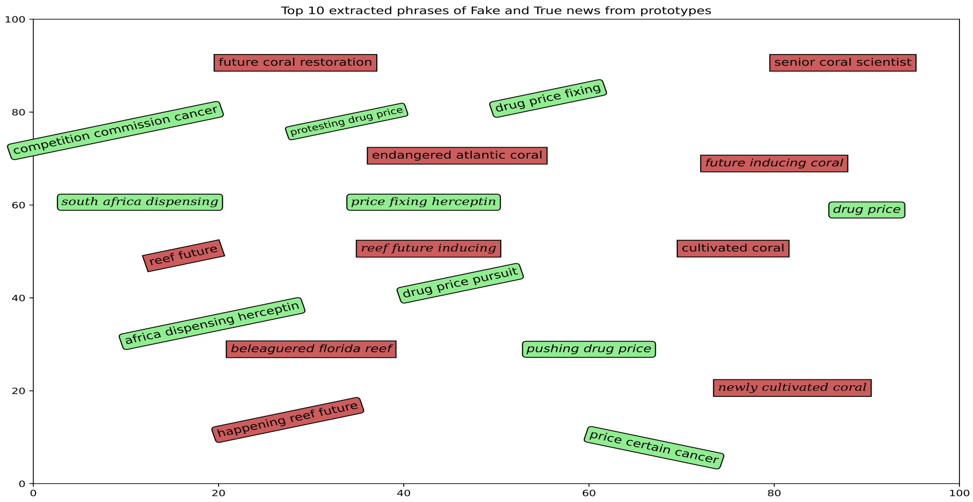


Figure 6.27: Figure showing top 10 phrases extracted from SA MMD prototypical set examples. Green phrases are true news phrases and red fake news phrases

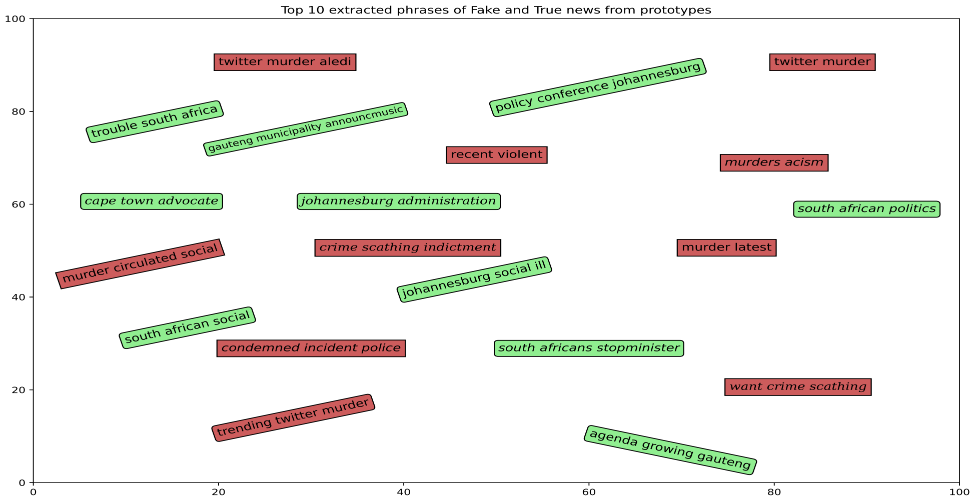


Figure 6.28: Figure showing top 10 phrases extracted from SA Influential prototypical set examples. Green phrases are true news phrases and red fake news phrases

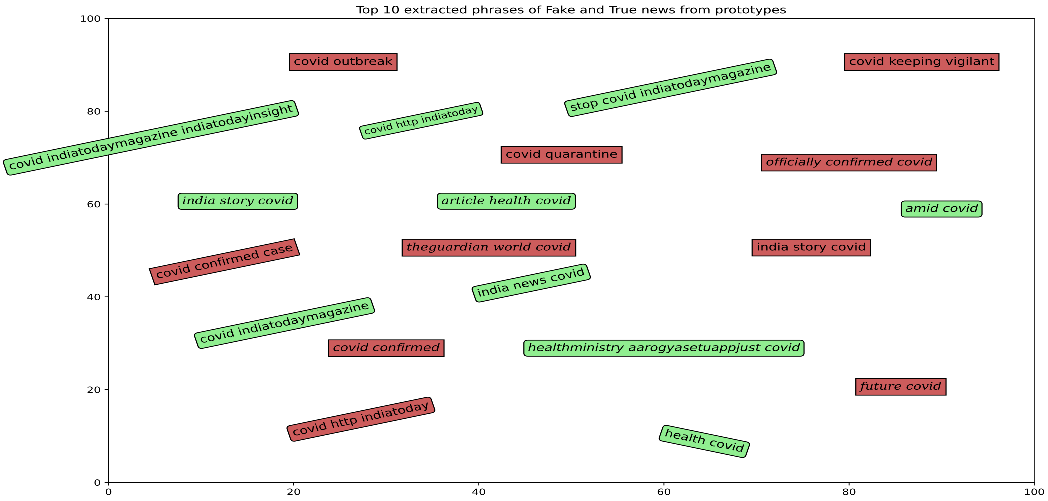


Figure 6.29: Figure showing top 10 phrases extracted from Covid MMD prototypical set examples. Green phrases are true news phrases and red fake news phrases

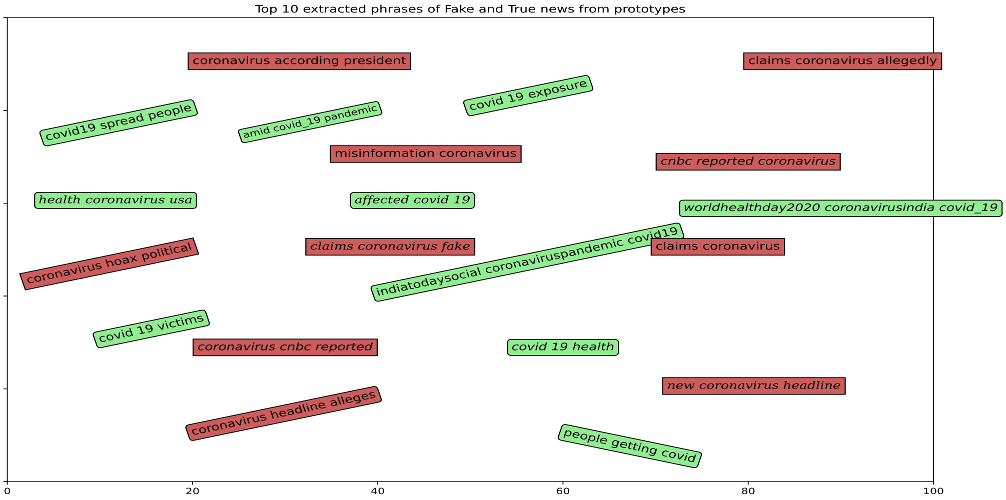


Figure 6.30: Figure showing top 10 phrases extracted from Covid Inflential prototypical set examples. Green phrases are true news phrases and red fake news

6.3.2 Sentiment and opinion detection

We examined both sentiment and subjectivity from the prototypical examples. We used the textblob measure of subjectivity and polarity to determine whether the text classes are more positive, negative, neutral, objective, or subjective. As can be observed in figure 6.31-6.35, we find that overall, fake news displayed in the orange bar is more subjective compared to true news displayed in the blue bar; however, when it comes to South African text, we find that true news is more subjective and fake is more objective. We further observe in figure 6.33-6.35 that true news is more positive than fake news. Another unique observation of the South African dataset is that it does not have neutral text when analysed with the MMD set.

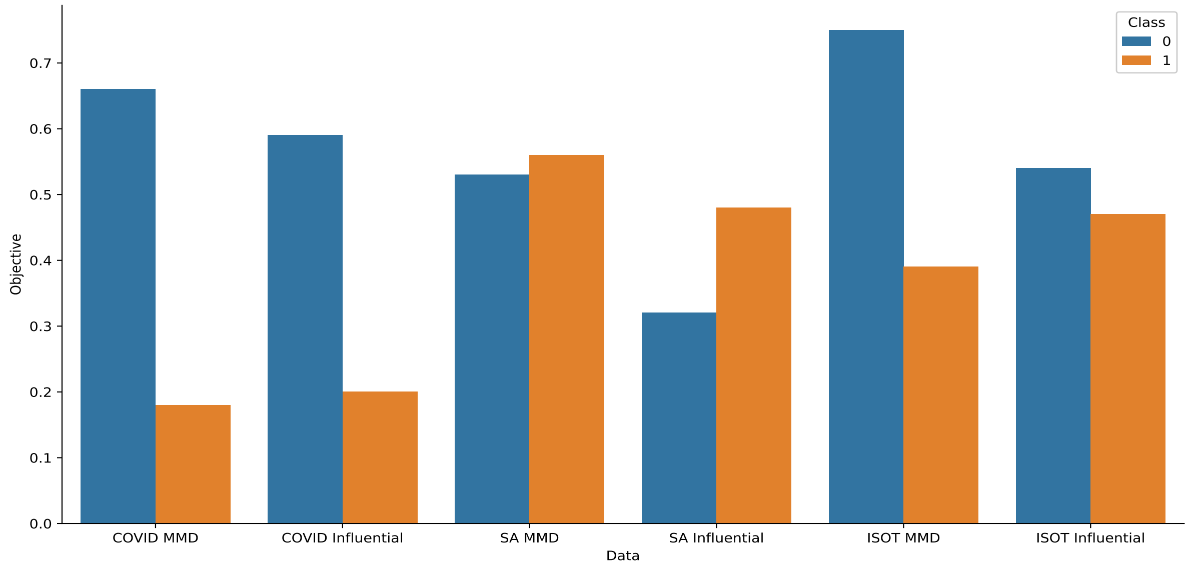


Figure 6.31: Figure showing opinion mining results based on the objective measure of textblob from all three datasets. It shows the percentage count of the text values that where subjective compared to objective

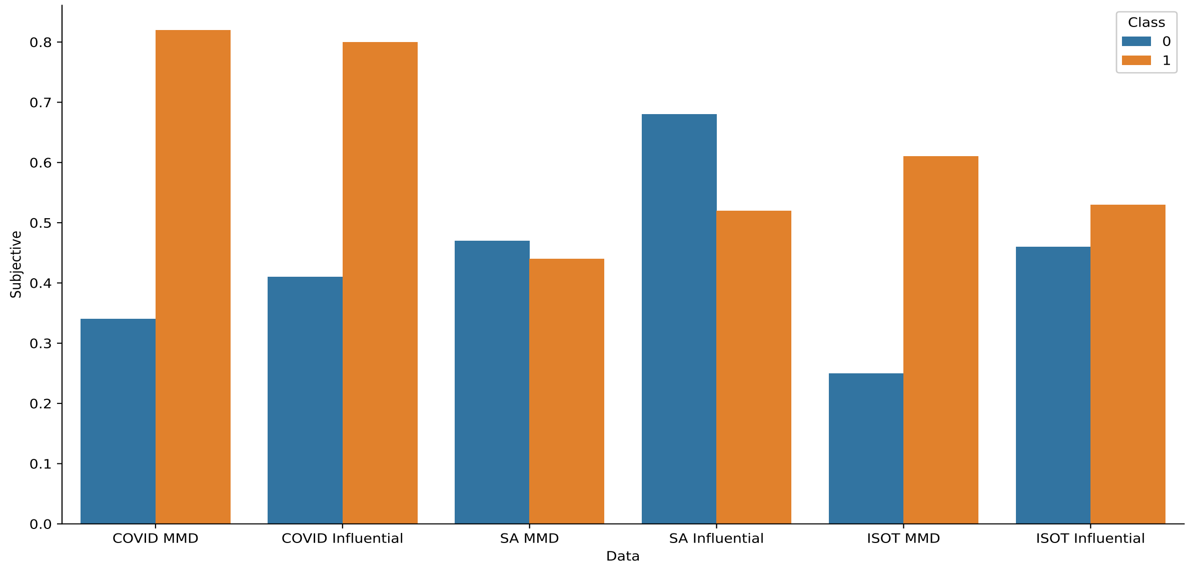


Figure 6.32: Figure showing opinion mining results based on subjectivity measure of textblob from all three datasets. It shows the percentage count of the text values that were subjective compared to objective

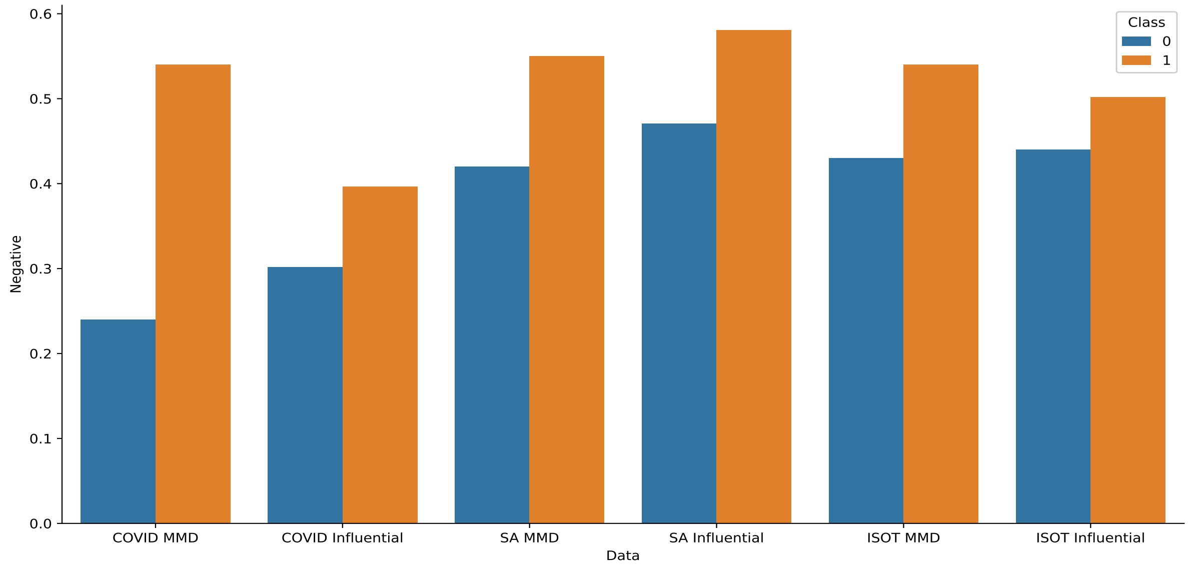


Figure 6.33: Figure showing sentiment results based on polarity measure of textblob from all three datasets. It shows the percentage count of the text values that were negative

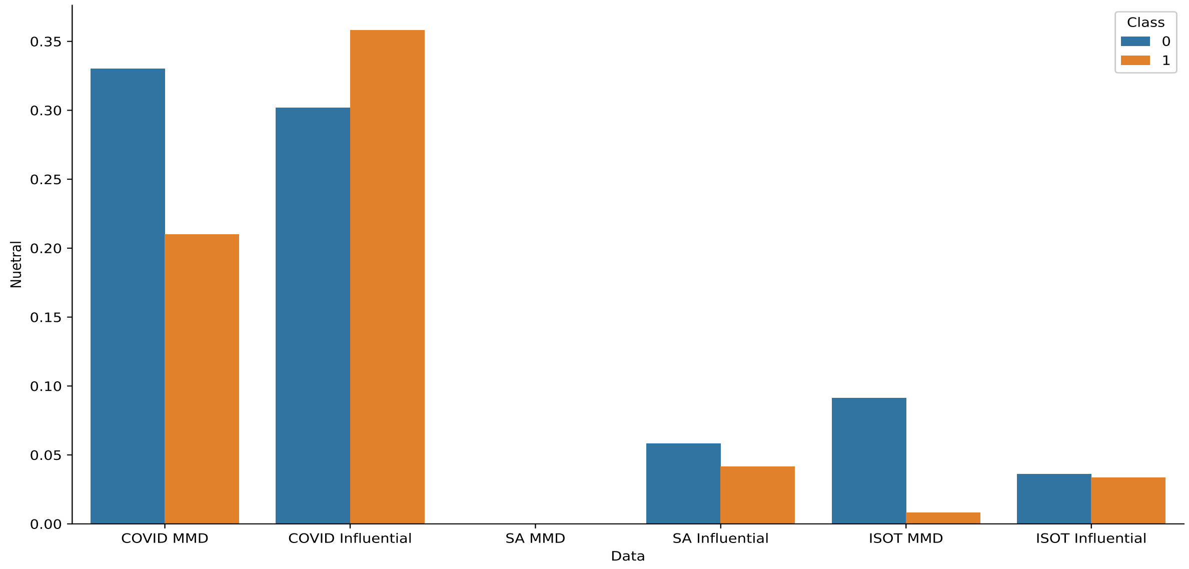


Figure 6.34: Figure showing sentiment results based on polarity measure of textblob from all three datasets. It shows the percentage count of the text values that were neutral

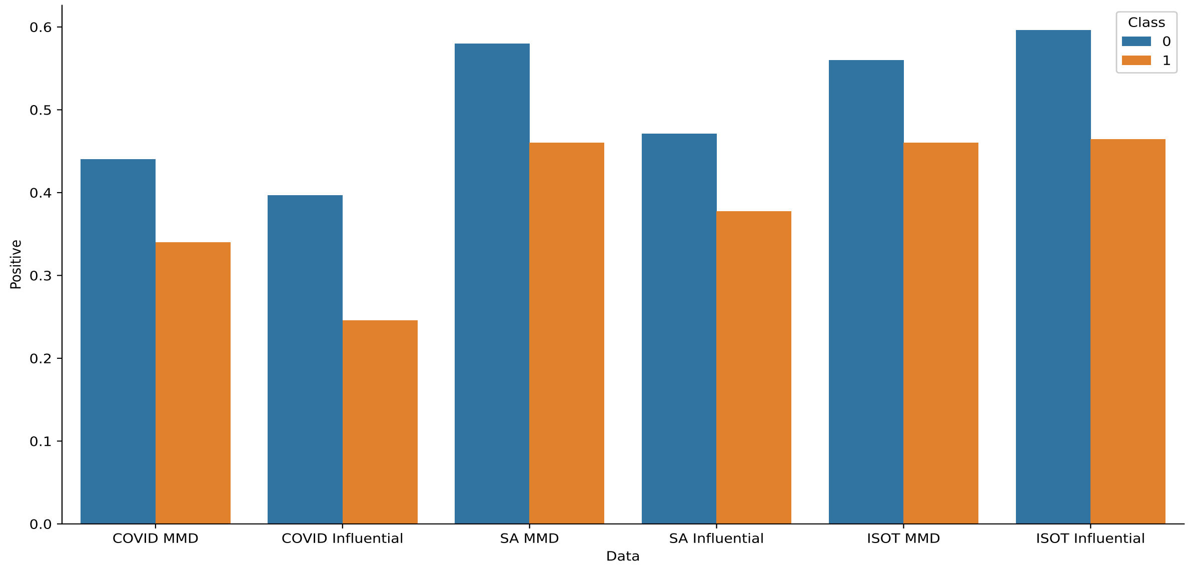


Figure 6.35: Figure showing sentiment results based on polarity measure of textblob from all three datasets. It shows the percentage count of the text values that were positive

6.3.3 Word analysis

Word length

Next, we observed if there were any significant differences in the average words between fake and true news data. We find that, on average, between the three datasets, fake news is about 18% less compared to true news, and we further observe that fake news has more words in capital, showing a more aggressive tone. Finally, we show the average words and capital word counts in table 6.9.

Data	Average Words	Average Upper
Covid True	35	3
Covid Fake	14	7
SA True	336	2
SA Fake	274	3.5
ISOT True	463	5
ISOT Fake	328	7.5

Table 6.9: Table showing the average words and upper case words in the prototypical example dataset

Word comparison

We use the Scattertext library to plot how fake and true news words differ in words and the intersection points between the groups. For example, in figure 6.36-6.38, we note that true words for covid data that are not used in fake news are words such as http, story, covidoutbreak, trending, and social; which are very neutral words that are not persuasive or negative. While fake news has words such as viral, country names, image, and kill, which are commonly associated with controversial words or topics.

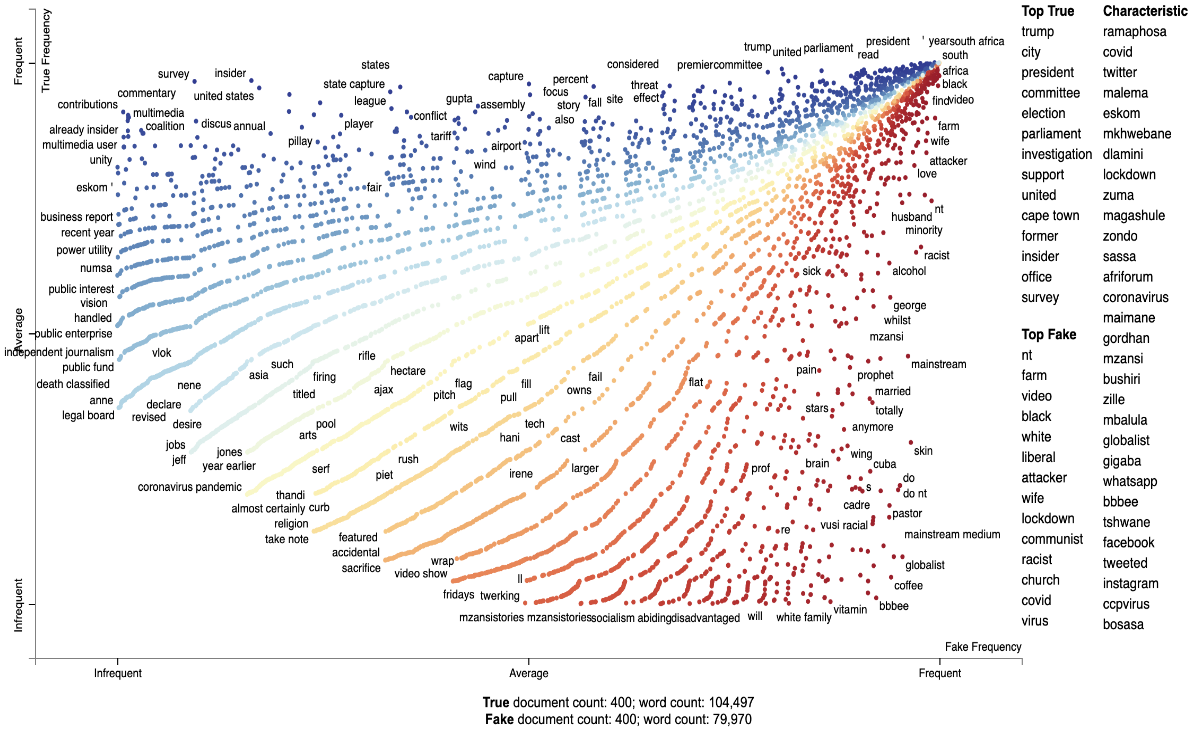
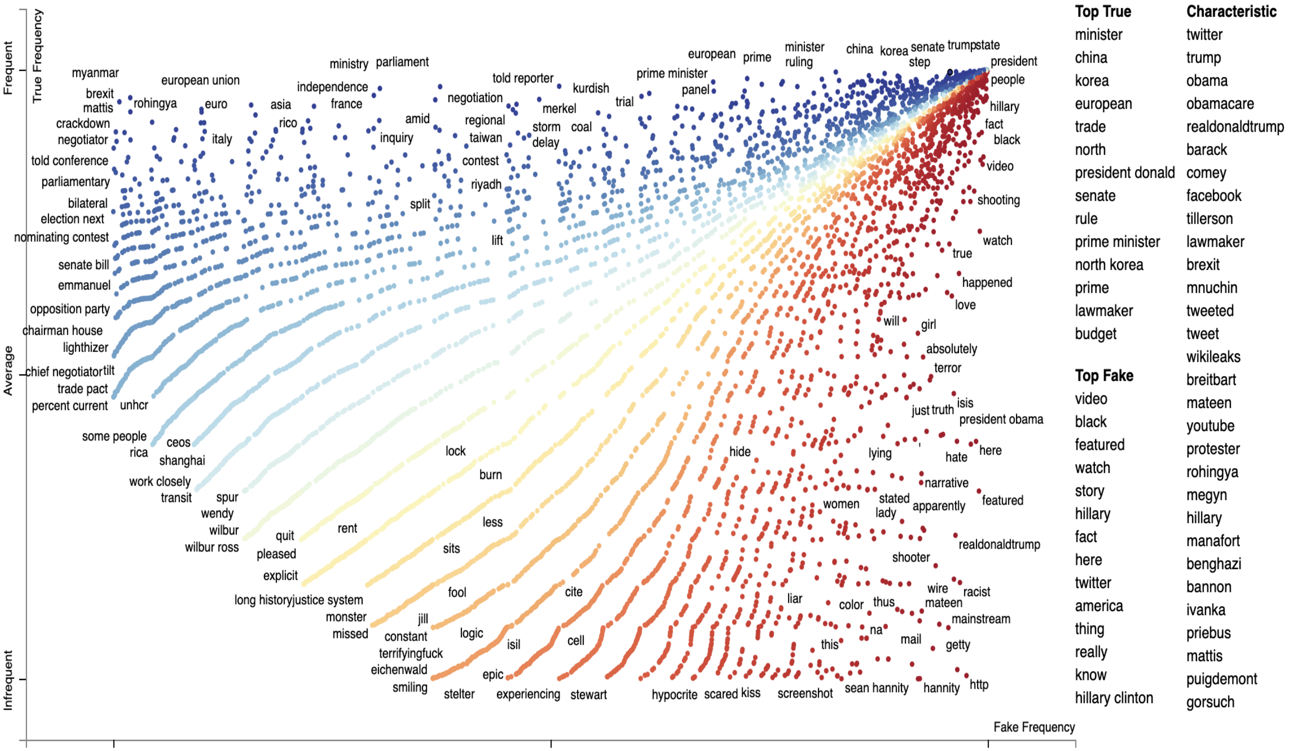


Figure 6.36: Figure showing word differences and similarities between the two class groups in SA prototypical examples



True document count: 400; word count: 113,575
Fake document count: 400; word count: 138,664

Figure 6.37: Figure showing word differences and similarities between the two class groups in ISOT prototypical examples

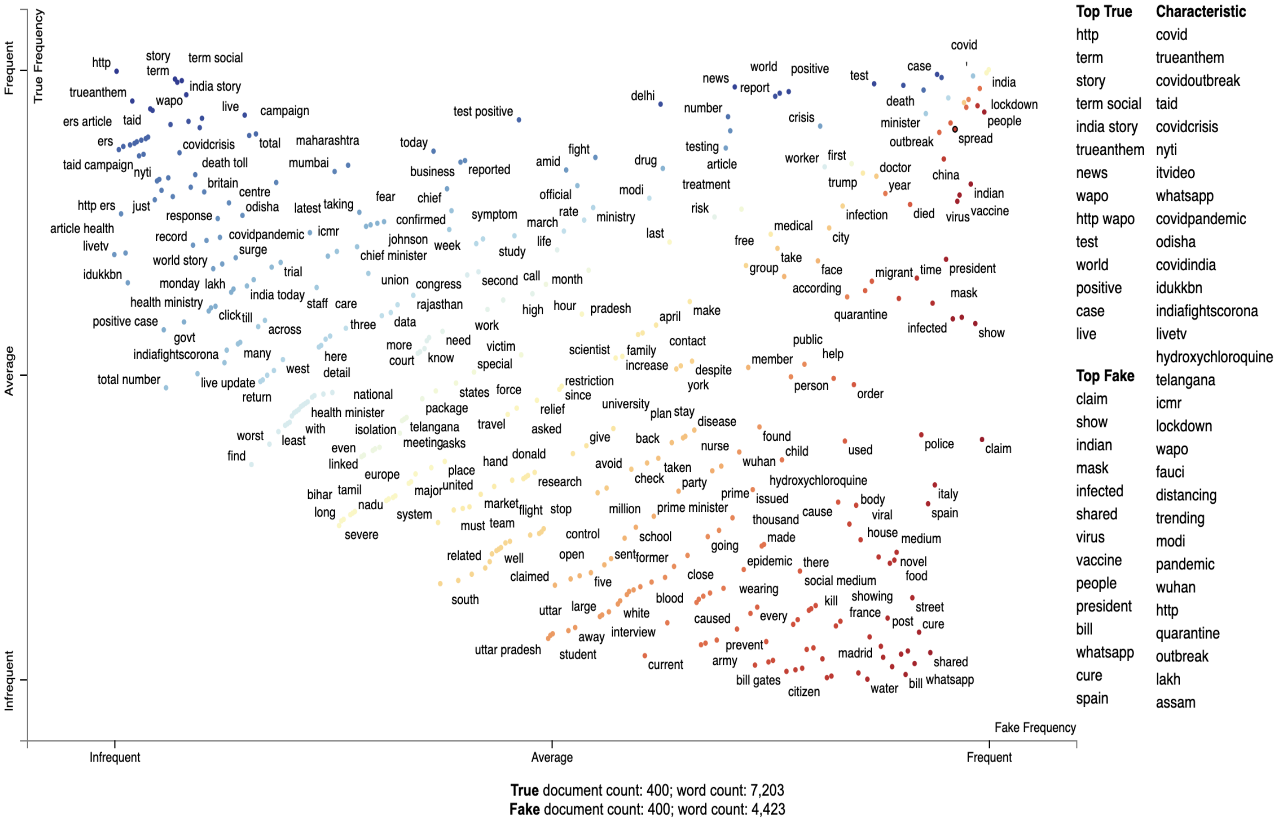


Figure 6.38: Figure showing word differences and similarities between the two class groups in COVID prototypical examples

6.3.4 Linguistic Properties

Given that we are dealing with language data, we also need to observe characteristic differences in linguistic properties. We, thus, observe the part of speech from each class in the prototypes shown in figure 6.39 to 6.41. The most significant differences are noted regarding noun usage in the ISOT prototype examples, where true news uses more nouns than fake news. Overall across the datasets, true news text use past tense verbs(VBD) and prepositions(IN) more compared to fake news, while fake news has more adjective (JJ), verb base form(VB), and verb 3rd person(VBZ) compared to true news. It is further noted that ISOT fake news text has more pronouns than true text compared to the rest of the datasets. South African text data has more foreign words(FW) than the rest of the dataset's text; this may be because South African text is usually used with a combination of colloquial words.

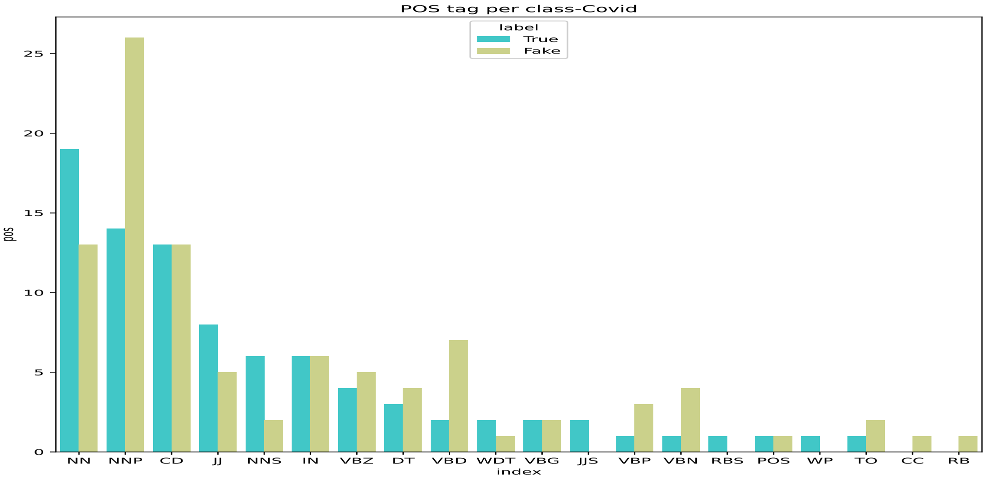


Figure 6.39: Figure showing part of speech tag for Covid prototypes

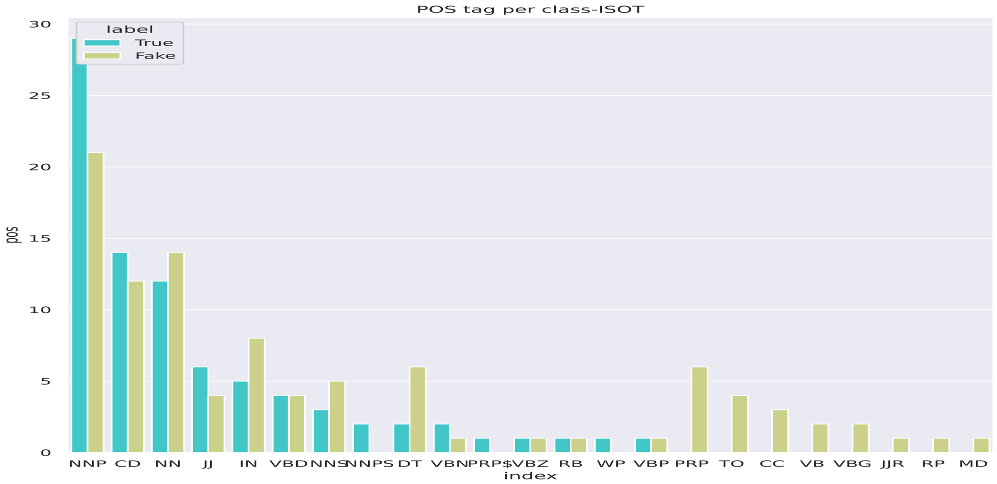


Figure 6.40: Figure showing part of speech tag for ISOT prototypes

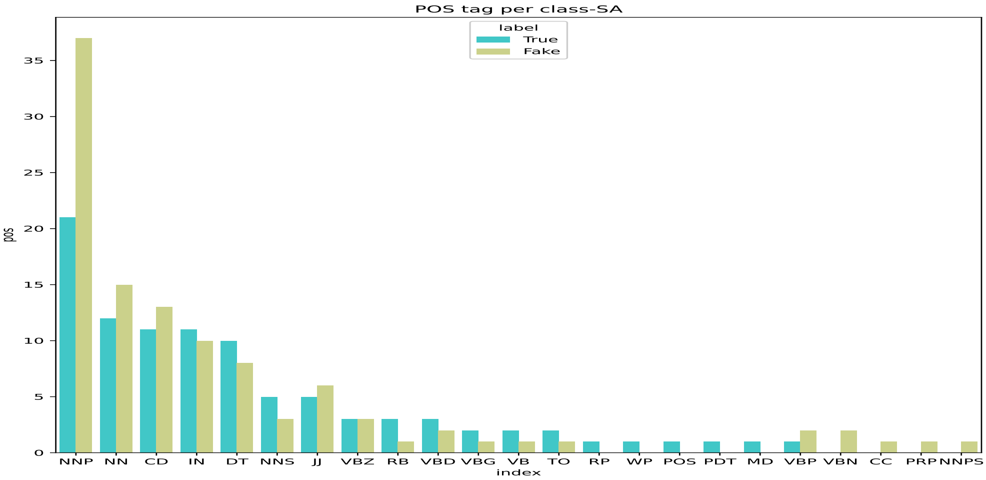
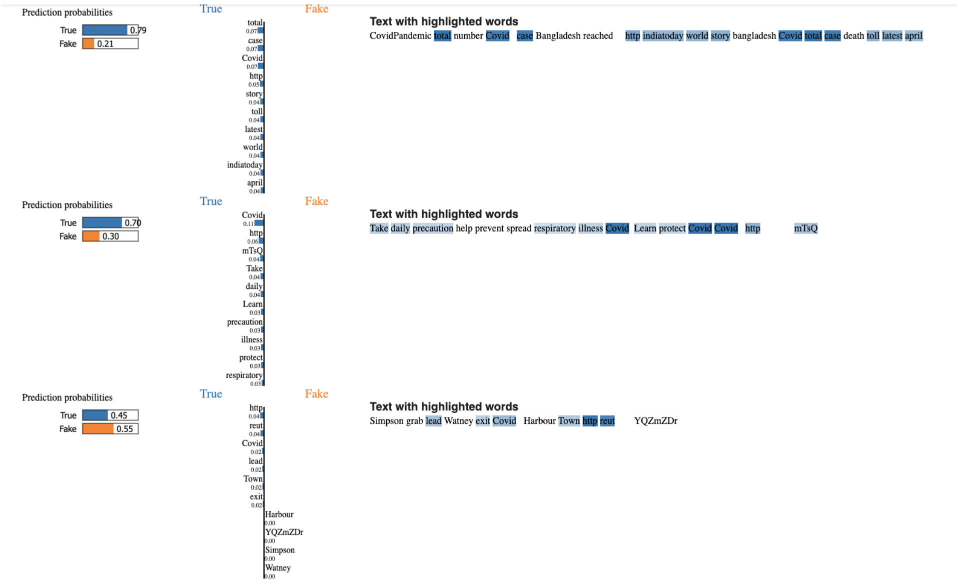


Figure 6.41: Figure showing part of speech tag for SA prototypes

6.3.5 Prototypical examples used with SP-Lime

We use the LTSM trained to perform model agnostic global explainer SP-Lime to determine any additional information that can be extracted using both example-based explainer and model agnostic explainer. The results of the explainer can be seen in figure 6.42-6.43. We find that the most relevant words indicated by the explainer are similar to what we observed in the scattertext output, where text classified as fake news has words that evoke emotion and are centered around a controversial topic whilst true news is more neutral and objective.



Text with highlighted words

CovidPandemic total number Covid case Bangladesh reached http indiatoday world story bangladesh Covid total case death toll latest april

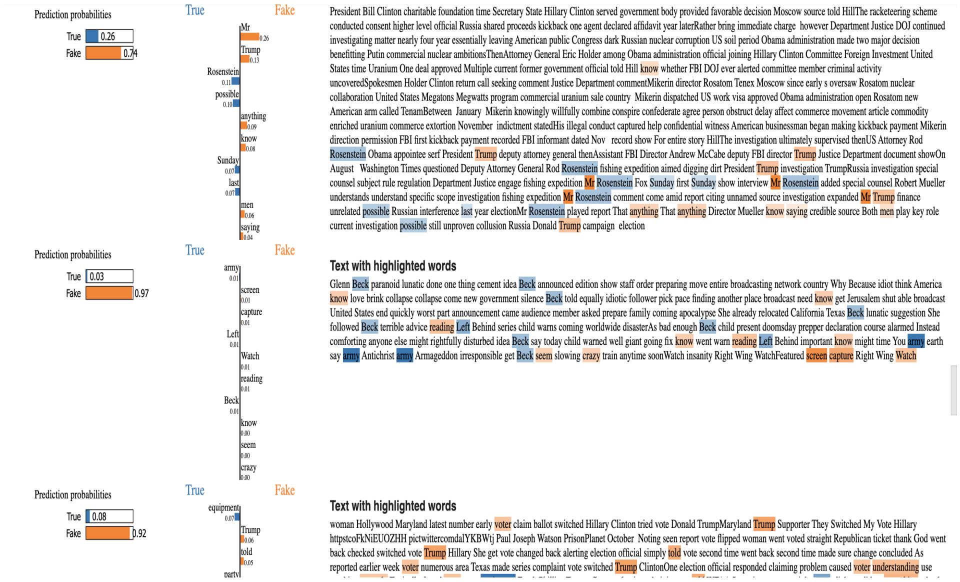
Text with highlighted words

Take daily precaution help prevent spread respiratory illness Covid Learn protect Covid Covid http mTsQ

Text with highlighted words

Simpson grab lead Watney exit Covid Harbour Town http reut YQZmZDr

Figure 6.42: Output of using SP-LIME with Prototypical examples to explain and indicate prevalent words for true news class



President Bill Clinton charitable foundation time Secretary State Hillary Clinton served government body provided favorable decision Moscow source told HillThe racketeering scheme conducted consent higher level official Russia shared proceeds kickback one agent declared affidavit year laterRather bring immediate charge however Department Justice DOJ continued investigating matter nearly four year essentially leaving American public Congress dark Russian nuclear corruption US soil period Obama administration made two major decision benefitting Putin commercial nuclear ambitionsThenAttorney General Eric Holder among Obama administration official joining Hillary Clinton Committee Foreign Investment United States time Uranium One deal approved Multiple current former government official told Hill know whether FBI DOJ ever alerted committee member criminal activity uncoveredSpokesmen Holder Clinton return call seeking comment Justice Department commentMikerin director Rosatom Tenex Moscow since early s oversaw Rosatom nuclear collaboration United States Megatons Megwatts program commercial uranium sale country Mikerin dispatched US work visa approved Obama administration open Rosatom new American arm called TenamBetween January Mikerin knowingly willfully combine conspire confederate agree person obstruct delay affect commerce movement article commodity enriched uranium commerce extortion November indictment statedHis illegal conduct captured help confidential witness American businessman began making kickback payment Mikerin direction permission FBI first kickback payment recorded FBI informant dated Nov record show For entire story HillThe investigation ultimately supervised thenUS Attorney Rod Rosenstein Obama appointee serf President Trump deputy attorney general thenAssistant FBI Director Andrew McCabe deputy FBI director Trump Justice Department document showOn August Washington Times questioned Deputy Attorney General Rod Rosenstein fishing expedition aimed digging dirt President Trump investigation TrumpRussia investigation special counsel subject rule regulation Department Justice engage fishing expedition Mr Rosenstein Fox Sunday first Sunday show interview Mr Rosenstein added special counsel Robert Mueller understands understand specific scope investigation fishing expedition Mr Rosenstein comment come amid report citing unnamed source investigation expanded Mr Trump finance unrelated possible Russian interference last year electionMr Rosenstein played report That anything That anything Director Mueller know saying credible source Both men play key role current investigation possible still unproven collusion Russia Donald Trump campaign election

Text with highlighted words

Glenn Beck paranoid lunatic done one thing cement idea Beck announced edition show staff order preparing move entire broadcasting network country Why Because idiot think America know love brink collapse collapse come new government silence Beck told equally idiotic follower pick pace finding another place broadcast need know get Jerusalem shut able broadcast United States end quickly worst part announcement came audience member asked prepare family coming apocalypse She already relocated California Texas Beck lunatic suggestion She followed Beck terrible advice reading Left Behind series child warns coming worldwide disasterAs bad enough Beck child present doomsday prepper declaration course alarmed Instead comforting anyone else might rightfully disturbed idea Beck say today child warned well giant going fix know went warn reading Left Behind important know might time You army earth say army Antichrist army Armageddon irresponsible get Beck seem slowing crazy train anytime soonWatch insanity Right Wing WatchFeatured screen capture Right Wing Watch

Text with highlighted words

woman Hollywood Maryland latest number early voter claim ballot switched Hillary Clinton tried vote Donald TrumpMaryland Trump Supporter They Switched My Vote Hillary httpstcoFkNiEUOZHH pictwittercomdalYKBWtj Paul Joseph Watson PrisonPlanet October Noting seen report vote flipped woman went voted straight Republican ticket thank God went back checked switched vote Trump Hillary She get vote changed back alerting election official simply told vote second time went back second time made sure change concluded As reported earlier week voter numerous area Texas made series complaint vote switched Trump ClintonOne election official responded claiming problem caused voter understanding use

Figure 6.43: Output of using SP-LIME with Prototypical examples to explain and indicate prevalent words for Fake news class

6.3.6 Critics and harmful instances

The advantage of both MMD-Critic and Influential instances is that they indicate which instances are harmful or are outliers to a class. MMD automatically selects the critics for each class, and 10% of the selected critic set belonged to true news, whilst 90% of the instances belonged to fake news. This indicates that models are more likely to misclassify fake news instances as true news.

We analyse phrases and the output of an attribution map for the top harmful influential instance and MMD-critic instance for fake news class (1).In figure 6.44, we can see that the phrases selected for the harmful instances are very similar to that of true news, making it difficult to distinguish between the classes.

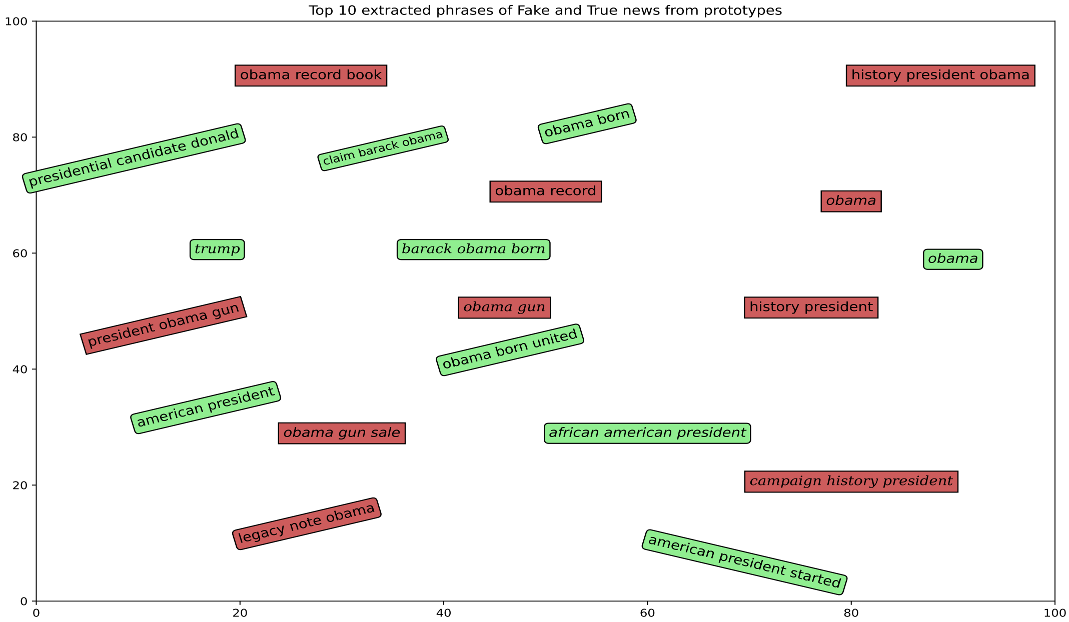
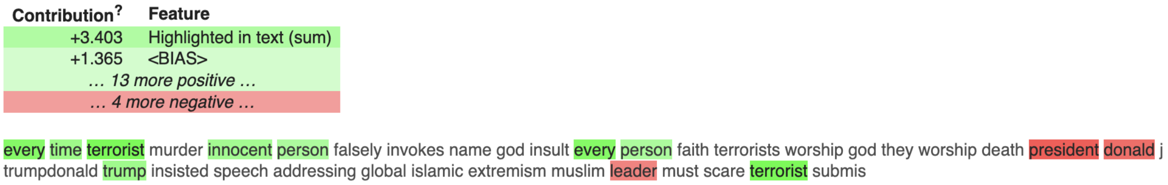


Figure 6.44: Output showing comparison of phrases selected for harmful instance of fake news data to that of a true news instance



Contribution?	Feature
+3.403	Highlighted in text (sum)
+1.365	<BIAS>
… 13 more positive …	
… 4 more negative …	

every time terrorist murder innocent person falsely invokes name god insult every person faith terrorists worship god they worship death president donald j trumpdonald trump insisted speech addressing global islamic extremism muslim leader must scare terrorist submis

Figure 6.45: Output of using attribution map to determine key features for harmful instance indicated by influential and MMD-critic

6.4 Summary

We used four approaches to select prototypes on three of our datasets; once we had our chosen sets based on the algorithms, we measured for each set the quality of the prototypes. Influential instances and MMD-Critic produced prototype sets with the highest quality; for these two sets, we explore if there are any character differences between the two classes. This is applied across the datasets to determine any differences between the datasets. Below we list some of the significant differences which we observe:

- In section 6.3.3, we performed word analysis, and the result across the datasets led to the conclusion that Fake news is shorter in length and contains more capital words.
- In section 6.3.2, we measured the subjectivity of the prototypes and observed that Fake news is more subjective and often opinionated.
- Using POS tagging in 6.3.4, we were able to conclude that Fake news uses more adjectives, verb in 3rd person, and pronouns in linguistic properties.
- By observing the output of the word cloud images, word comparison's from scattertext, and phrase extraction, we note that Fake news phrases are more aggressive and are always around controversial topics.
- In figure 6.40 POS tag results for South African data and 6.36 the scattertext output; We determine that South African fake data has more foreign words as more colloquial words are used.
- Further, from the POS tag in 6.3.4, we can also say South African text is either positive or negative, with hardly any neutral text.
- Lastly, South African fake news is less subjective and more objective compared to the rest of the datasets as per the output of subjectivity analysis performed in 6.3.2.

We also observed from harmful instances that fake news text could be very difficult to detect by a layperson. However, by using linguistic properties, a linguistic expert can detect the slight differences that exist. Consumers of news can use the above characteristics to determine the probability of fakeness from text data.

Chapter 7

Conclusions

This chapter summarizes the research work presented in this study. We revisit our research questions and discuss the main findings. We also discuss proposed future work.

7.1 Summary of Conclusions

In this research, we discuss the problem of interpretability in natural language processing, specifically with respect to misinformation classification tasks. We want to determine what behavioural factors can be used to distinguish a misinformation class and which interpretability model is best suited for interpreting misinformation data.

In chapter 4 we used the commonly used interpretation methods for natural language processing, methods which output attribution maps by highlighting the most important words and phrases in a prediction to explain the prediction outcome. We found that the attribution methods were faithful in their interpretation. However, given that they explained a single instance, it was challenging to determine common characteristics of a class and clearly determine the characteristics, if any, of misinformation data.

In chapter 5 we used example-based techniques for model interpretation to determine if any key characteristics could be used to explain the prediction between the two classes. We found that with natural language; prototype selection methods, unlike images, one cannot easily determine the differences and explanations of the predictions of the two classes. We, therefore, further probed the prototypes of each class extracted using various techniques and tools to determine the character and linguistic differences. We found some significant differences in word length, tone, subjectivity, part of speech, and sentiment as described in chapter 6. We further found that more fake news instances are critics compared to true news, showing that fake news curated tries to mimic the properties and authenticity of the news piece.

From the results obtained in chapter 5 and 4 we were able to compare prototype/example-based methodology with the most commonly used algorithms. We note that with example-based methods, one can extract a sample of the data which best represents a particular class. Then from the sample, easily extract characteristics that make up the class. The key difference's between example-based techniques and common techniques is that with example-based techniques, one could make characteristic differences before having to train a model. In contrast, most techniques require a model to be trained to extract patterns. Further, examples extracted may help build mental models using the prototypes and critics to achieve better human interpretability. However, when it comes to natural language processing, using both techniques to achieve interpretability in misinformation is best suited to truly understand the data and extract key information between the classes. Lastly, given that most of the characteristic differences between the two classes observed require a language expert, the method may be extended to include human-in-the-loop to determine the prototypes' robustness, faithfulness, and quality.

7.2 Future Work

There are more insights that we can achieve from the prototype methodology. We list below some potential avenues for future research, specifically in disinformation:

- The quality of prototypes and critic sets selected can change based on the kernel used in MMD-critic and how we represent the data. Using MMD-critic as the algorithm, we can explore how the prototype and critic sets differ if a string kernel or sequence embeddings is used to capture the sequence of the documents.

- Determining the quality of the prototypes is crucial and needs further exploring. Current literature proposes, together with training, a model to have humans in the loops to assess the prototypes to determine their quality and characteristics. In the future, we can extend our research by adding a human-in-the-loop measure, more specifically, linguistic experts, to examine the quality of prototypes.

- Lastly, Given the distinct key patterns observed, we can explore building an intrinsic interpretable pipeline. Which will have two phases, with the first phase extracting key feature attributes from the text and then using the feature attributes as feature inputs in a logistic regression model or any white-box model. Such that we can distinguish if, over time, we can have key features that can be used in white-box models to classifier misinformation.

Bibliography

[1] A Convolutional Approach for Misinformation Identification. *IJCAI International Joint Conference on Artificial Intelligence*, 0:3901–3907, 2017.

[2] SmoothGrad: removing noise by adding noise. 2017.

[3] Interpretable Machine Learning. A Guide for Making Black Box Models Explainable. *Book*, page 247, 2019.

[4] Raed Alharbi, Minh N. Vu, and My T. Thai. Evaluating Fake News Detection Models from Explainable Machine Learning Perspectives. *IEEE International Conference on Communications*, 2021.

[5] Jackie Ayoub, X. Jessie Yang, and Feng Zhou. Combat COVID-19 infodemic using explainable natural language processing models. *Information Processing and Management*, 58(4), 2021.

[6] Jasmijn Bastings and Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? pages 149–155, 2020.

[7] Kim Been, Khanna Rajiv, and Oluwasanmi Koyejo. Examples are not Enough, Learn to Criticize! Criticism for Interpretability Been. *NIPS*, 2016.

[8] Jacob Bien and Robert Tibshirani. PROTOTYPE SELECTION FOR INTERPRETABLE CLASSIFICATION. *Institute of Mathematical Statistics*, 5(4):2403–2424, 2011.

[9] Jacob Bien and Robert Tibshirani. Prototype selection for interpretable classification. *Annals of Applied Statistics - ANN APPL STAT*, 5, 02 2012.

[10] Peng Ce and Bao Tie. An analysis method for interpretability of CNN text classification model. *Future Internet*, 12(12):1–14, 2020.

[11] Akif Cinar. Overview of existing approaches for the interpretation of machine learning models. pages 1–11.

[12] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What Does BERT Look at? An Analysis of BERT's Attention. pages 276–286, 2019.

[13] Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A Survey of the State of Explainable AI for Natural Language Processing. (Section 5), 2020.

[14] Arun Das and Paul Rad. Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey. pages 1–24, 2020.

[15] Hwang Doosung and Kimt Daewon. Nearest neighbor based prototype classification preserving class regions. *Journal of Information Processing Systems*, 5:1345–1357, 02 2017.

[16] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. feb 2017.

[17] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2020.

[18] Keyur Faldu and Amit Sheth. Discovering the Encoded Lingusitic Knowledge in NLP models. *Towards Data Science*, pages 7–8, 2020.

[19] Guillaume Gadek and Paul Guélorget. An interpretable model to measure fakeness and emotion in news. *Procedia Computer Science*, 176:78–87, 2020.

[20] Moyank Giri, Tarun Aditya, Prasad B Honnavalli, and Sivaraman Eswaran. Automated and Interpretable Fake News Detection with Explainable Artificial Intelligence. *SSRN Electronic Journal*, 2022.

[21] Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269, 2019.

[22] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks, 2019.

[23] Linwei Hu, Jie Chen, Vijayan N Nair, and Agus Sudjianto. Surrogate Locally-Interpretable Models with Supervised Machine Learning Algorithms. *arXiv:2007.14528 [cs, stat]*, pages 1–24, 2020.

[24] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. On the Validity of Self-Attention as Explanation in Transformer Models. *arXiv*, 2017.

[25] Been Kim, Cynthia Rudin, and Julie Shah. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in Neural Information Processing Systems*, 3(January):1952–1960, 2014.

[26] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions, 2017.

[27] Hima Lakkaraju, Julius Adebayo, and Sameer Singh. Explaining Machine Learning Predictions: State-of-the-art, Challenges, Opportunities Slides and Video: explainml-tutorial.github.io Motivation. 2020.

[28] Qing Liao, Heyan Chai, Hao Han, Xiang Zhang, Xuan Wang, Wen Xia, and Ye Ding. An integrated multi-task model for fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5154–5165, 2022.

[29] Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc Interpretability for Neural NLP: A Survey. 2021.

[30] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[31] Ken Mishima and Hayato Yamana. A Survey on Explainable Fake News Detection. *IEICE Transactions on Information and Systems*, E105D(7):1249–1257, 2022.

[32] Sina Mohseni, Fan Yang, Shiva Pentyala, Mengnan Du, Yi Liu, Nic Lupfer, Xia Hu, Shuiwang Ji, and Eric Ragan. Machine Learning Explanations to Prevent Overtrust in Fake News Detection. *Proceedings of the International AAAI Conference on Web and Social Media*, 15:421–431, 2021.

[33] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. *Communications in Computer and Information Science*, 1323(01):417–431, 2020.

[34] Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15, 2018.

[35] Ian E. Nielsen, Ghulam Rasool, Dimah Dera, Nidhal Bouaynaya, and Ravi P. Ramachandran. Robust Explainability: A Tutorial on Gradient-Based Attribution Methods for Deep Neural Networks. pages 1–21, 2021.

[36] Yu Qiao, Daniel Wiechmann, and Elma Kerz. A Language-Based Approach to Fake News Detection Through Interpretable Features and BRNN. *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 14–31, 2020.

[37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Model-Agnostic Interpretability of Machine Learning. (Whi), 2016.

[38] Alexis Ross, Ana Marasović, and Matthew Peters. Explaining NLP Models via Minimal Contrastive Editing (MiCE). pages 3840–3852, 2021.

[39] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences, 2019.

[40] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. Defend: Explainable fake news detection. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 395–405, 2019.

[41] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. Hierarchical propagation networks for fake news detection: Investigation and exploitation. *Proceedings of the 14th International AAAI Conference on Web and Social Media, ICWSM 2020*, pages 626–637, 2020.

[42] Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. Interpreting Deep Learning Models in Natural Language Processing: A Review. pages 1–36, 2021.

[43] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Integrated Gradient Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118, 2017.

[44] Martin Tutek and Jan Šnajder. Iterative Recursive Attention Model for Interpretable Sequence Classification. pages 249–257, 2019.

[45] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models. pages 6707–6723, 2021.

[46] Pengtao Xie, Yuntian Deng, and Eric Xing. Latent Variable Modeling with Diversity-Inducing Mutual Angular Regularization. 2015.

[47] Fan Yang, Mengnan Du, Eric D. Ragan, Shiva K. Pentyala, Hao Yuan, Shuiwang Ji, Sina Mohseni, Rhema Linder, and Xia Hu. XFake: Explainable fake news detector with visualizations. *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*, 2:3600–3604, 2019.

[48] Die Zhang, Huilin Zhou, Hao Zhang, Xiaoyi Bao, Da Huo, Ruizhao Chen, Xu Cheng, Mengyue Wu, and Quanshi Zhang. Building Interpretable Interaction Trees for Deep NLP Models. 2020.

[49] Wei Zhao, Rahul Singh, Joshi Tarun, Sudjianto Agus, and Nair Vijayan N. Self-interpretable Convolutional Neural Networks for Text Classification. *Lecture Notes in Electrical Engineering*, 715:309–316, 2021.

Appendix A

Appendix 1

The first Appendix shows the word analysis images discussed in the exploratory data analysis section in chapter 3. We show visually word cloud and tri-gram output of our three datasets.

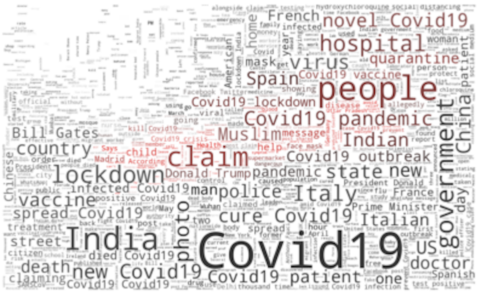


Figure A.1: Word Cloud class 1(Fake) word clod image.

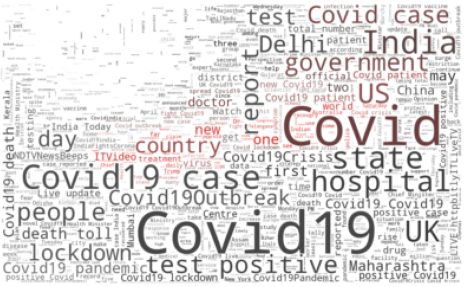


Figure A.2: Word Cloud class 0(True) word clod image.

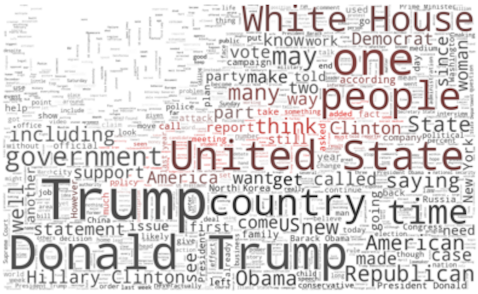


Figure A.3: Word Cloud class 1(Fake) word clod image.

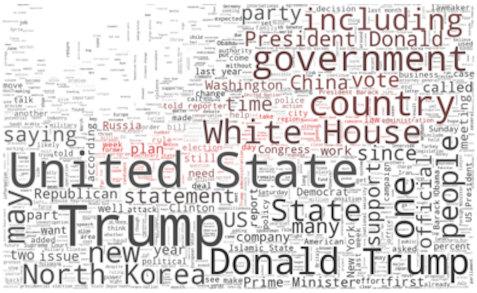


Figure A.4: Word Cloud class 0(True) word clod image.

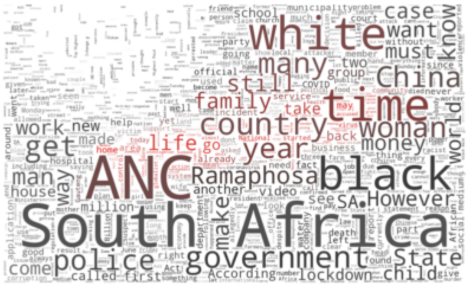


Figure A.5: Word Cloud class 1(Fake) word clod image.

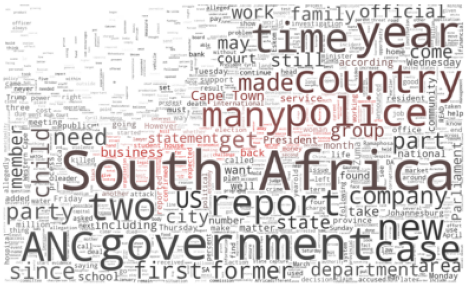


Figure A.6: Word Cloud class 0(True) word clod image.

A.1 Word Cloud analysis

We observe that the prominent words in fake news are similar to that of true news. Showing it is difficult to separate fake and true news by just using prominent words.

Appendix B

Appendix 2

In this Appendix we show the results of unsupervised topic modelling with the gensim model as described in Chapter 3.

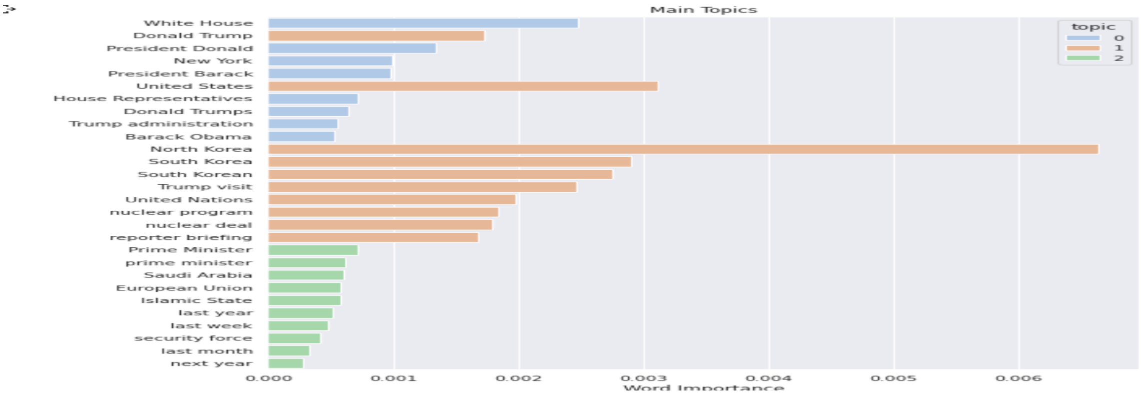


Figure B.1: Bi-gram topic analysis for ISOT Data

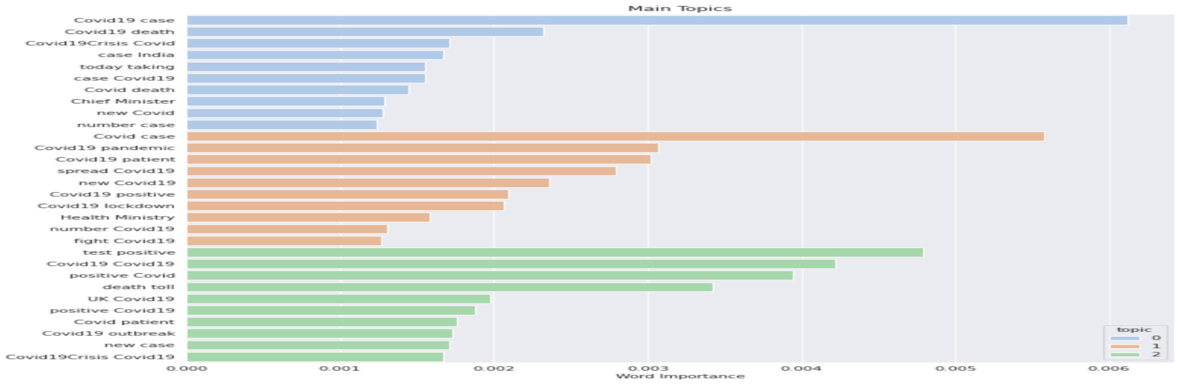


Figure B.2: Bi-gram topic analysis for Covid Data

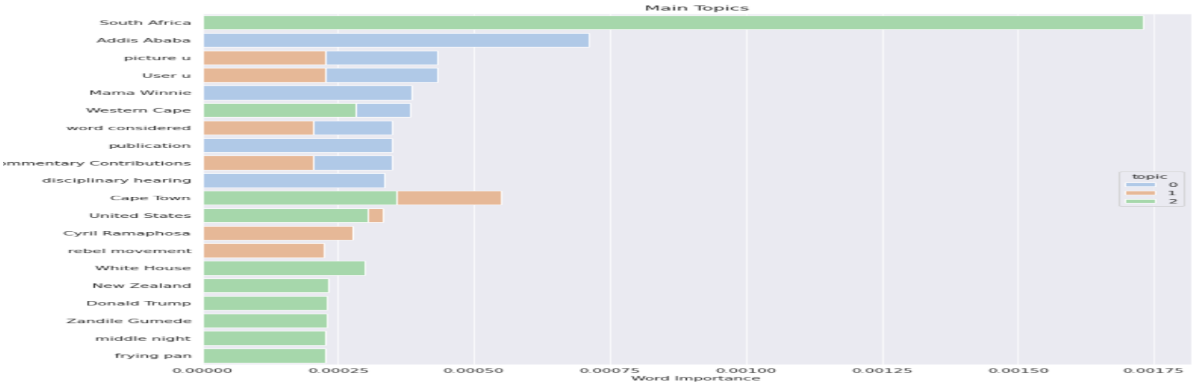


Figure B.3: Bi-gram topic analysis for SA Data

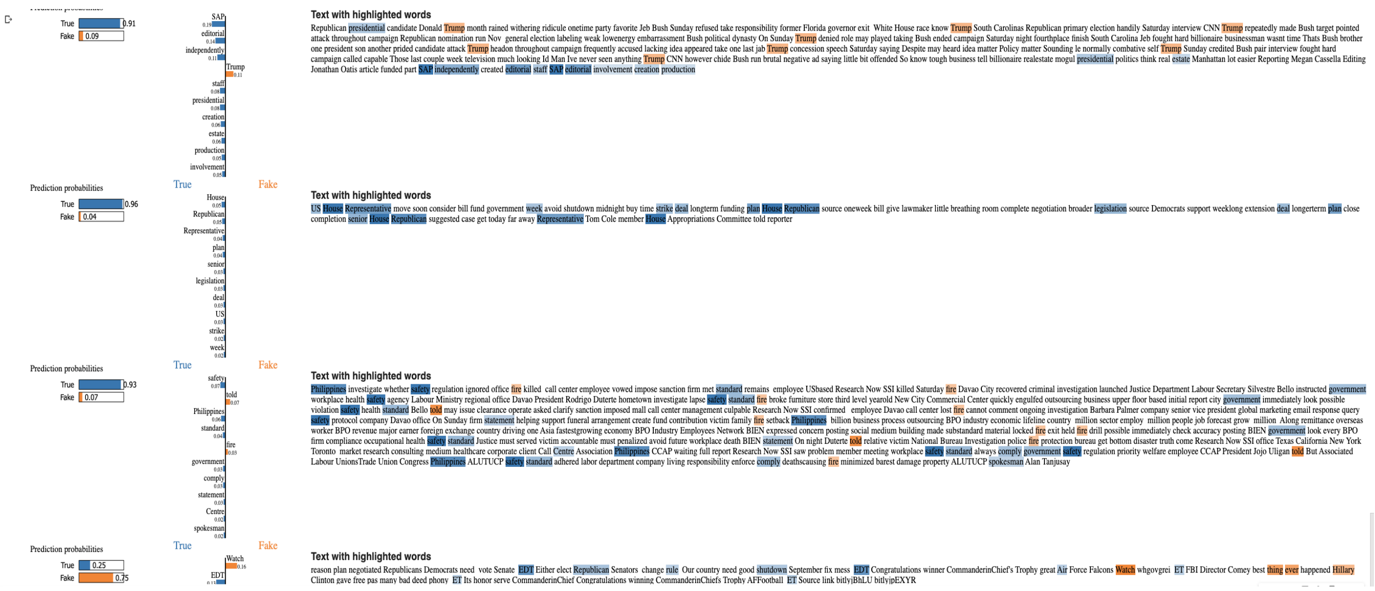
B.1 Topic Modeling

We show the three main topic themes across each dataset.

Appendix C

Appendix 3

In this Appendix we show the results of unsupervised topic modelling with the gensim model as described in Chapter 4.



Text with highlighted words

Republican presidential candidate Donald Trump month rained withering ridicule onetime party favorite Jeb Bush Sunday refused take responsibility former Florida governor exit White House race know Trump South Carolinas Republican primary election handily Saturday interview CNN Trump repeatedly made Bush target pointed attack throughout campaign Republican nomination run Nov general election labeling weak lowenergy embarrassment Bush political dynasty On Sunday Trump denied role may played taking Bush ended campaign Saturday night fourthplace finish South Carolina Jeb fought hard billionaire businessman wasnt time Thats Bush brother one president son another prided candidate attack Trump headon throughout campaign frequently accused lacking idea appeared take one last jab Trump concession speech Saturday saying Despite may heard idea matter Policy matter Sounding le normally combative self Trump Sunday credited Bush pair interview fought hard campaign called capable Those last couple week television much looking Id Man Ive never seen anything Trump CNN however chide Bush run brutal negative ad saying little bit offended So know tough business tell billionaire realestate mogul presidential politics think real estate Manhattan lot easier Reporting Megan Cassella Editing Jonathan Oatis article funded part SAP independently created editorial staff SAP editorial involvement creation production

Text with highlighted words

US House Representative move soon consider bill fund government week avoid shutdown midnight buy time strike deal longterm funding plan House Republican source oneweek bill give lawmaker little breathing room complete negotiation broader legislation source Democrats support weeklong extension deal longerterm plan close completion senior House Republican suggested case get today far away Representative Tom Cole member House Appropriations Committee told reporter

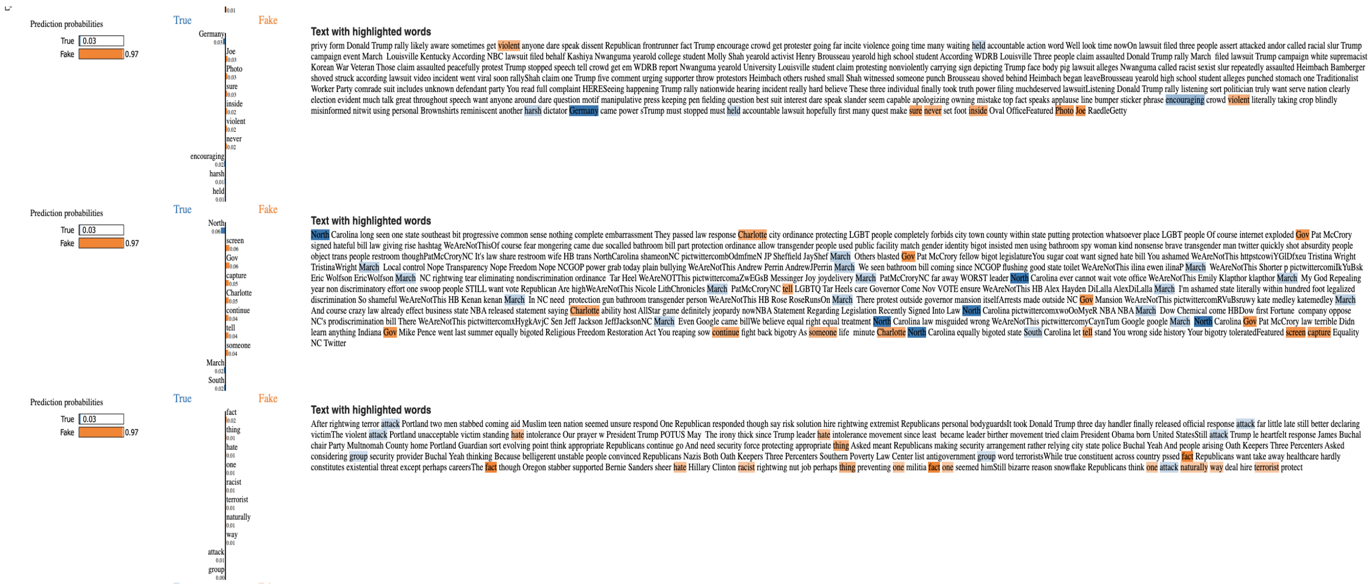
Text with highlighted words

Philippines investigate whether safety regulation ignored office fire killed call center employee vowed impose sanction firm met standard remains employee USbased Research Now SSI killed Saturday fire Davao City recovered criminal investigation launched Justice Department Labour Secretary Silvestre Bello instructed government workplace health safety agency Labour Ministry regional office Davao President Rodrigo Duterte hometown investigate lapse safety standard fire broke furniture store third level yearold New City Commercial Center quickly engulfed outsourcing business upper floor based initial report city government immediately look possible violation safety health standard Bello told may issue clearance operate asked clarify sanction imposed mall call center management culpable Research Now SSI confirmed employee Davao call center lost fire cannot comment ongoing investigation Barbara Palmer company senior vice president global marketing email response query safety protocol company Davao office On Sunday firm statement helping support funeral arrangement create fund contribution victim family fire setback Philippines billion business process outsourcing BPO industry economic lifeline country million sector employ million people job forecast grow million Along remittance overseas worker BPO revenue major earner foreign exchange country driving one Asia fastestgrowing economy BPO Industry Employees Network BIEN expressed concern posting social medium building made substandard material locked fire exit held fire drill possible immediately check accuracy posting BIEN government look every BPO firm compliance occupational health safety standard Justice must served victim accountable must penalized avoid future workplace death BIEN statement On night Duterte told relative victim National Bureau Investigation police fire protection bureau get bottom disaster truth come Research Now SSI office Texas California New York Toronto market research consulting medium healthcare corporate client Call Centre Association Philippines CCAP waiting full report Research Now SSI saw problem member meeting workplace safety standard always comply government safety regulation priority welfare employee CCAP President Jojo Uligan told But Associated Labour UnionsTrade Union Congress Philippines ALUTUCP safety standard adhered labor department company living responsibility enforce comply deathscausing fire minimized barest damage property ALUTUCP spokesman Alan Tanjusay

Text with highlighted words

reason plan negotiated Republicans Democrats need vote Senate EDT Either elect Republican Senators change rule Our country need good shutdown September fix mess EDT Congratulations winner CommanderinChief's Trophy great Air Force Falcons Watch whgovgrei ET FBI Director Comey best thing ever happened Hillary Clinton gave free pas many bad deed phony ET Its honor serve CommanderinChief Congratulations winning CommanderinChiefs Trophy AFFootball ET Source link bitlyjBhLU bitlyjpEXYR

Figure C.1: SP-LIME results from a representative sample of fake news dataset



Text with highlighted words

privy form Donald Trump rally likely aware sometimes get violent anyone dare speak dissent Republican frontrunner fact Trump encourage crowd get protester going far incite violence going time many waiting held accountable action word Well look time nowOn lawsuit filed three people assert attacked andor called racial slur Trump campaign event March Louisville Kentucky According NBC lawsuit filed behalf Kashiya Nwanguma yearold college student Molly Shah yearold activist Henry Brousseau yearold high school student According WDRB Louisville Three people claim assaulted Donald Trump rally March filed lawsuit Trump campaign white supremacist Korean War Veteran Those claim assaulted peacefully protest Trump stopped speech tell crowd get em WDRB report Nwanguma yearold University Louisville student claim protesting nonviolently carrying sign depicting Trump face body pig lawsuit alleges Nwanguma called racist sexist slur repeatedly assaulted Heimbach Bamberger shoved struck according lawsuit video incident went viral soon rallyShah claim one Trump five comment urging supporter throw protestors Heimbach others rushed small Shah witnessed someone punch Brousseau shoved behind Heimbach began leaveBrousseau yearold high school student alleges punched stomach one Traditionalist Worker Party comrade suit includes unknown defendant party You read full complaint HERESeeing happening Trump rally nationwide hearing incident really hard believe These three individual finally took truth power filing muchdeserved lawsuitListening Donald Trump rally listening sort politician truly want serve nation clearly election evident much talk great throughout speech want anyone around dare question motif manipulative press keeping pen fielding question best suit interest dare speak slander seem capable apologizing owning mistake top fact speaks applause line bumper sticker phrase encouraging crowd violent literally taking crop blindly misinformed nitwit using personal Brownshirts reminiscent another harsh dictator Germany came power sTrump must stopped must held accountable lawsuit hopefully first many quest make sure never set foot inside Oval OfficeFeatured Photo Joe RaedleGetty

Text with highlighted words

North Carolina long seen one state southeast bit progressive common sense nothing complete embarrassment They passed law response Charlotte city ordinance protecting LGBT people completely forbids city town county within state putting protection whatsoever place LGBT people Of course internet exploded Gov Pat McCrory signed hateful bill law giving rise hashtag WeAreNotThisOf course fear mongering came due socalled bathroom bill part protection ordinance allow transgender people used public facility match gender identity bigot insisted men using bathroom spy woman kind nonsense brave transgender man twitter quickly shot absurdity people object trans people restroom thoughPatMcCroryNC It's law share restroom wife HB trans NorthCarolina shameonNC pictwittercombOdmfmeN JP Sheffield JayShef March Others blasted Gov Pat McCrory fellow bigot legislatureYou sugar coat want signed hate bill You ashamed WeAreNotThis httpstcowiYGlDfxeu Tristina Wright TristinaWright March Local control Nope Transparency Nope Freedom Nope NCGOP power grab today plain bullying WeAreNotThis Andrew Perrin AndrewJPerrin March We seen bathroom bill coming since NCGOP flushing good state toilet WeAreNotThis ilina ewen ilinaP March WeAreNotThis Shorter p pictwittercomiIkYuBsk Eric Wolfson EricWolfson March NC rightwing tear eliminating nondiscrimination ordinance Tar Heel WeAreNOTThis pictwittercomaZwEGsB Messinger Joy joydelivery March PatMcCroryNC far away WORST leader North Carolina ever cannot wait vote office WeAreNotThis Emily Klapthor klapthor March My God Repealing year non discriminatory effort one swoop people STILL want vote Republican Are highWeAreNotThis Nicole LithChronicles March PatMcCroryNC tell LGBTQ Tar Heels care Governor Come Nov VOTE ensure WeAreNotThis HB Alex Hayden DiLalla AlexDiLalla March I'm ashamed state literally within hundred foot legalized discrimination So shameful WeAreNotThis HB Kenan kenan March In NC need protection gun bathroom transgender person WeAreNotThis HB Rose RoseRunsOn March There protest outside governor mansion itselfArrests made outside NC Gov Mansion WeAreNotThis pictwittercomRVuBsruwy kate medley katemedley March And course crazy law already effect business state NBA released statement saying Charlotte ability host AllStar game definitely jeopardy nowNBA Statement Regarding Legislation Recently Signed Into Law North Carolina pictwittercomxwoOoMyeR NBA NBA March Dow Chemical come HBDow first Fortune company oppose NC's prodiscrimination bill There WeAreNotThis pictwittercomxHygkAvjC Sen Jeff Jackson JeffJacksonNC March Even Google came billWe believe equal right equal treatment North Carolina law misguided wrong WeAreNotThis pictwittercomyCaynTum Google google March North Carolina Gov Pat McCrory law terrible Didn learn anything Indiana Gov Mike Pence went last summer equally bigoted Religious Freedom Restoration Act You reaping sow continue fight back bigotry As someone life minute Charlotte North Carolina equally bigoted state South Carolina let tell stand You wrong side history Your bigotry toleratedFeatured screen capture Equality NC Twitter

Text with highlighted words

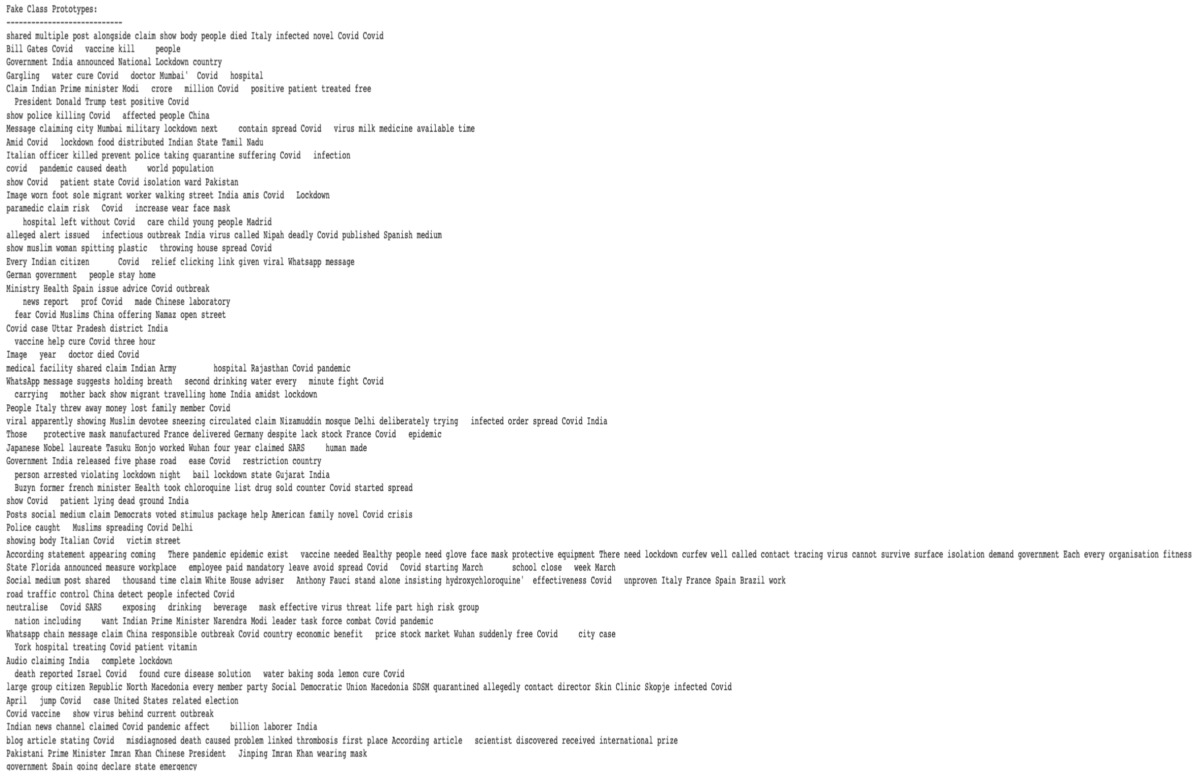
After rightwing terror attack Portland two men stabbed coming aid Muslim teen nation seemed unsure respond One Republican responded though say risk solution hire rightwing extremist Republicans personal bodyguardsIt took Donald Trump three day handler finally released official response attack far little late still better declaring victimThe violent attack Portland unacceptable victim standing hate intolerance Our prayer w President Trump POTUS May The irony thick since Trump leader hate intolerance movement since least became leader birther movement tried claim President Obama born United StatesStill attack Trump le heartfelt response James Buchal chair Party Multnomah County home Portland Guardian sort evolving point think appropriate Republicans continue go And need security force protecting appropriate thing Asked meant Republicans making security arrangement rather relying city state police Buchal Yeah And people arising Oath Keepers Three Percenters Asked considering group security provider Buchal Yeah thinking Because belligerent unstable people convinced Republicans Nazis Both Oath Keepers Three Percenters Southern Poverty Law Center list antigovernment group word terroristsWhile true constituent across country pssed fact Republicans want take away healthcare hardly constitutes existential threat except perhaps careersThe fact though Oregon stabber supported Bernie Sanders sheer hate Hillary Clinton racist rightwing nut job perhaps thing preventing one militia fact one seemed himStill bizarre reason snowflake Republicans think one attack naturally way deal hire terrorist protect

Figure C.2: SP-LIME results from a representative sample of True news dataset

Appendix D

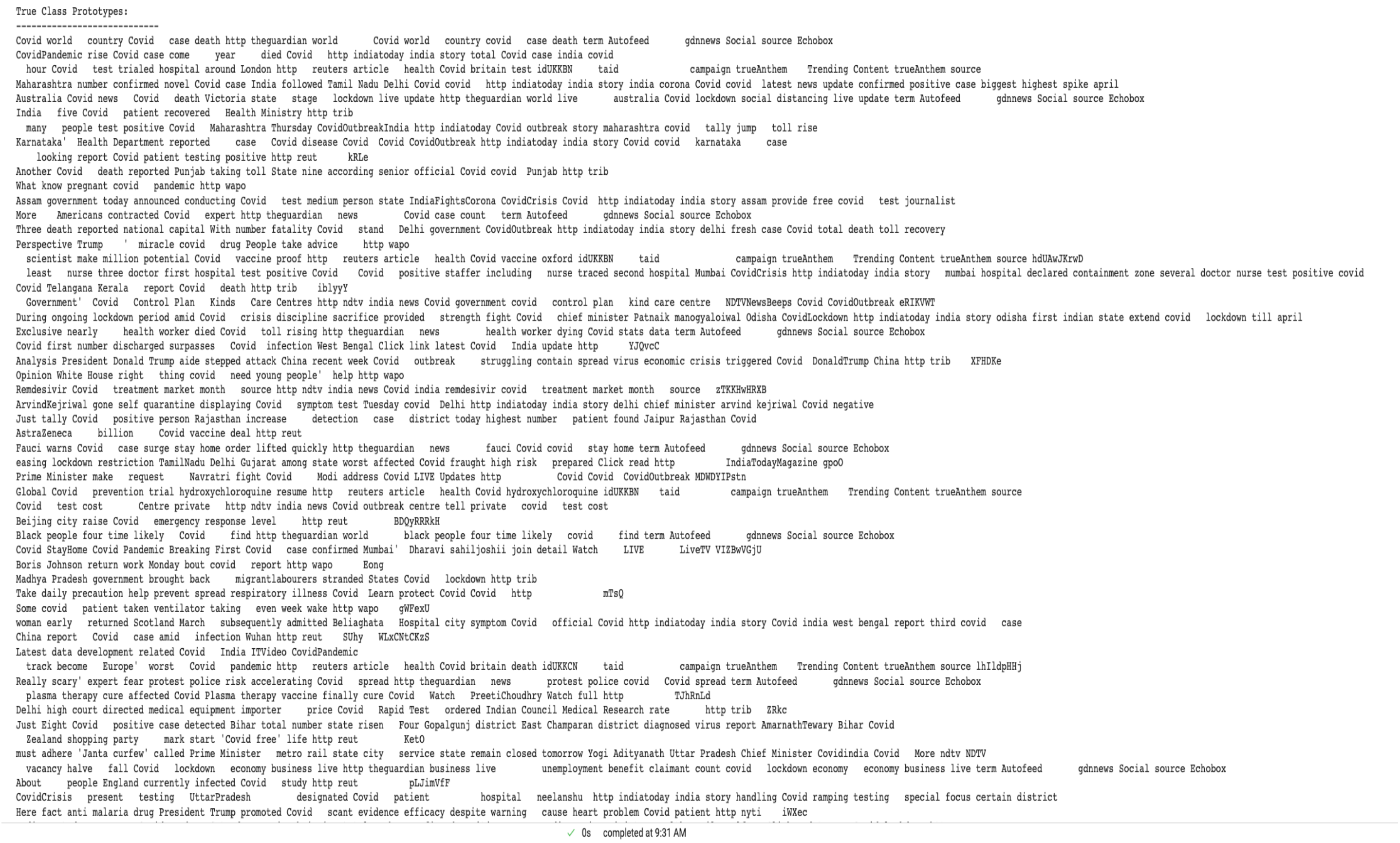
Appendix 4

In this Appendix we show the results of using MMD-critic approach to extract prototypes as discussed in chapter 6. We extracted prototypes for each dataset.



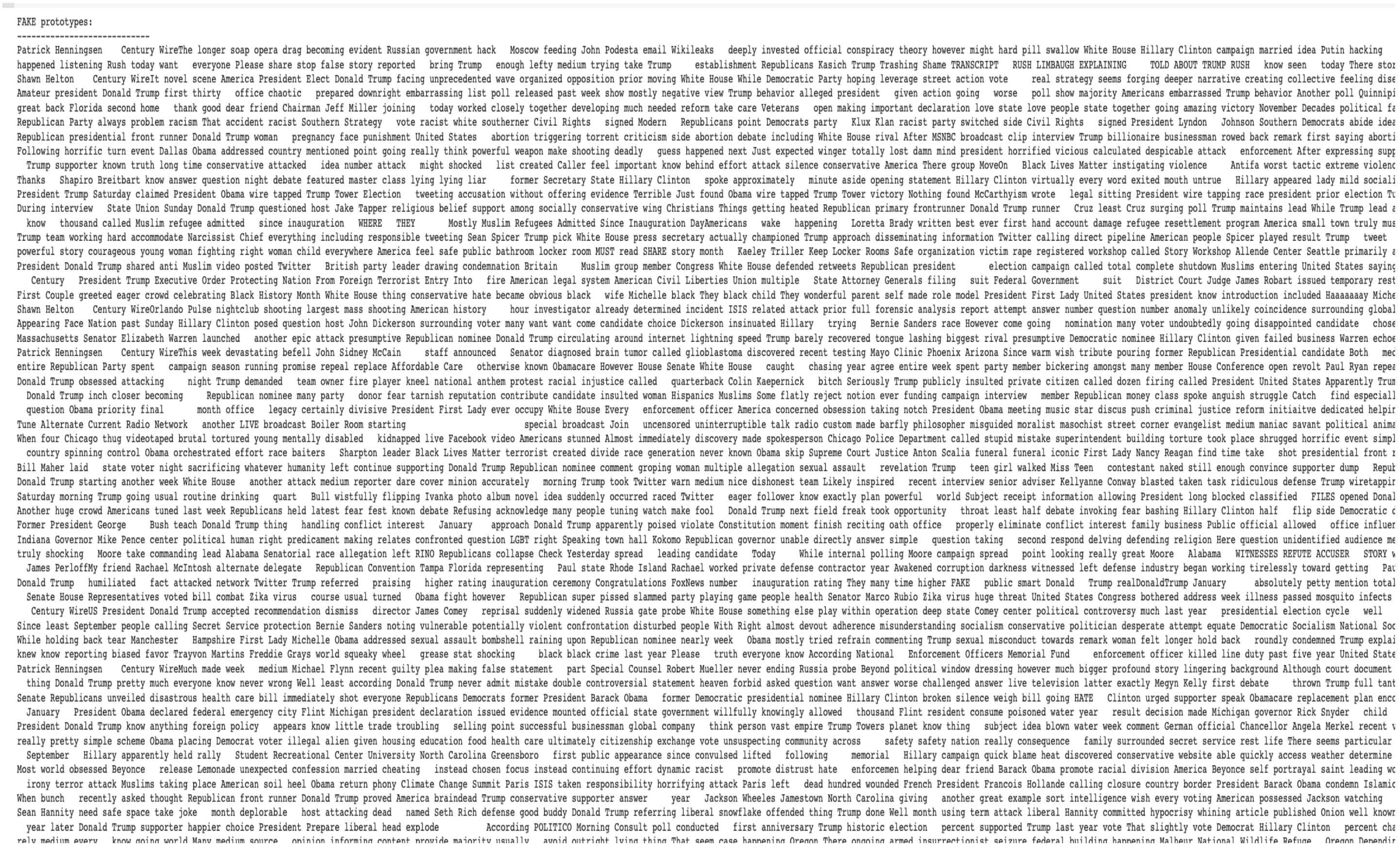
```
Fake Class Prototypes:
----------------------------
shared multiple post alongside claim show body people died Italy infected novel Covid Covid
Bill Gates Covid   vaccine kill     people
Government India announced National Lockdown country
Gargling   water cure Covid   doctor Mumbai'  Covid   hospital
Claim Indian Prime minister Modi   crore   million Covid   positive patient treated free
  President Donald Trump test positive Covid
show police killing Covid   affected people China
Message claiming city Mumbai military lockdown next     contain spread Covid   virus milk medicine available time
Amid Covid   lockdown food distributed Indian State Tamil Nadu
Italian officer killed prevent police taking quarantine suffering Covid   infection
covid   pandemic caused death     world population
show Covid   patient state Covid isolation ward Pakistan
Image worn foot sole migrant worker walking street India amis Covid   Lockdown
paramedic claim risk   Covid   increase wear face mask
    hospital left without Covid   care child young people Madrid
alleged alert issued   infectious outbreak India virus called Nipah deadly Covid published Spanish medium
show muslim woman spitting plastic   throwing house spread Covid
Every Indian citizen       Covid   relief clicking link given viral Whatsapp message
German government   people stay home
Ministry Health Spain issue advice Covid outbreak
    news report   prof Covid   made Chinese laboratory
  fear Covid Muslims China offering Namaz open street
Covid case Uttar Pradesh district India
  vaccine help cure Covid three hour
Image   year   doctor died Covid
medical facility shared claim Indian Army         hospital Rajasthan Covid pandemic
WhatsApp message suggests holding breath   second drinking water every   minute fight Covid
  carrying   mother back show migrant travelling home India amidst lockdown
People Italy threw away money lost family member Covid
viral apparently showing Muslim devotee sneezing circulated claim Nizamuddin mosque Delhi deliberately trying   infected order spread Covid India
Those    protective mask manufactured France delivered Germany despite lack stock France Covid   epidemic
Japanese Nobel laureate Tasuku Honjo worked Wuhan four year claimed SARS     human made
Government India released five phase road   ease Covid   restriction country
  person arrested violating lockdown night   bail lockdown state Gujarat India
  Buzyn former french minister Health took chloroquine list drug sold counter Covid started spread
show Covid   patient lying dead ground India
Posts social medium claim Democrats voted stimulus package help American family novel Covid crisis
Police caught   Muslims spreading Covid Delhi
showing body Italian Covid   victim street
According statement appearing coming   There pandemic epidemic exist   vaccine needed Healthy people need glove face mask protective equipment There need lockdown curfew well called contact tracing virus cannot survive surface isolation demand government Each every organisation fitness
State Florida announced measure workplace   employee paid mandatory leave avoid spread Covid   Covid starting March       school close   week March
Social medium post shared   thousand time claim White House adviser   Anthony Fauci stand alone insisting hydroxychloroquine'  effectiveness Covid   unproven Italy France Spain Brazil work
road traffic control China detect people infected Covid
neutralise   Covid SARS     exposing   drinking   beverage   mask effective virus threat life part high risk group
  nation including     want Indian Prime Minister Narendra Modi leader task force combat Covid pandemic
Whatsapp chain message claim China responsible outbreak Covid country economic benefit   price stock market Wuhan suddenly free Covid     city case
  York hospital treating Covid patient vitamin
Audio claiming India   complete lockdown
  death reported Israel Covid   found cure disease solution   water baking soda lemon cure Covid
large group citizen Republic North Macedonia every member party Social Democratic Union Macedonia SDSM quarantined allegedly contact director Skin Clinic Skopje infected Covid
April   jump Covid   case United States related election
Covid vaccine   show virus behind current outbreak
Indian news channel claimed Covid pandemic affect     billion laborer India
blog article stating Covid   misdiagnosed death caused problem linked thrombosis first place According article   scientist discovered received international prize
Pakistani Prime Minister Imran Khan Chinese President   Jinping Imran Khan wearing mask
government Spain going declare state emergency
```

Figure D.1: Extracted fake-Covid class prototypical examples



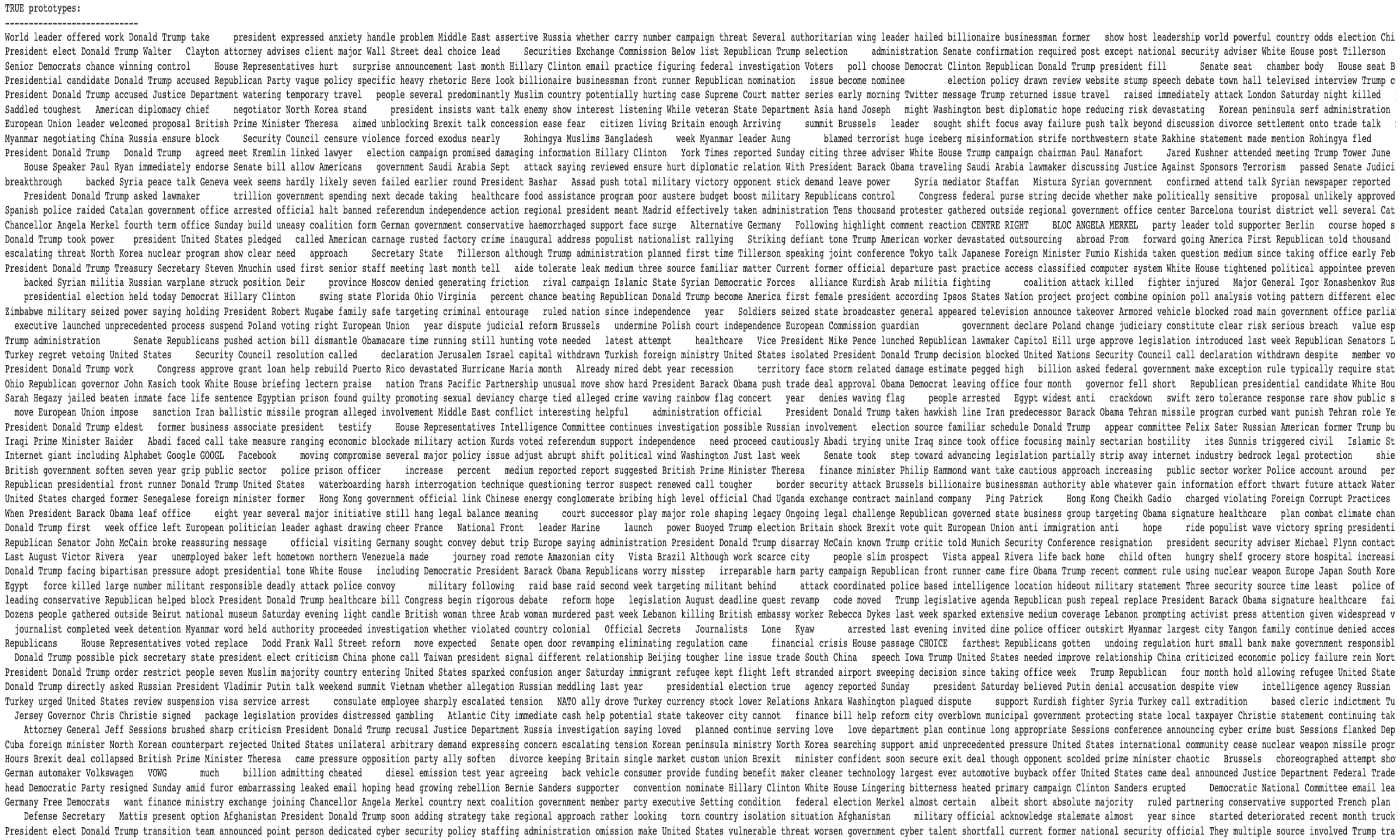
```
True Class Prototypes:
----------------------------
Covid world   country Covid   case death http theguardian world       Covid world   country covid   case death term Autofeed       gdnnews Social source Echobox
CovidPandemic rise Covid case come     year     died Covid   http indiatoday india story total Covid case india covid
  hour Covid   test trialed hospital around London http   reuters article   health Covid britain test idUKKBN     taid              campaign trueAnthem    Trending Content trueAnthem source
Maharashtra number confirmed novel Covid case India followed Tamil Nadu Delhi Covid covid   http indiatoday india story india corona Covid covid  latest news update confirmed positive case biggest highest spike april
Australia Covid news   Covid   death Victoria state   stage   lockdown live update http theguardian world live       australia Covid lockdown social distancing live update term Autofeed       gdnnews Social source Echobox
India   five Covid   patient recovered   Health Ministry http trib
  many   people test positive Covid   Maharashtra Thursday CovidOutbreakIndia http indiatoday Covid outbreak story maharashtra covid   tally jump   toll rise
Karnataka'  Health Department reported     case   Covid disease Covid  Covid CovidOutbreak http indiatoday india story Covid covid   karnataka     case
    looking report Covid patient testing positive http reut      kRLe
Another Covid   death reported Punjab taking toll State nine according senior official Covid covid  Punjab http trib
What know pregnant covid   pandemic http wapo
Assam government today announced conducting Covid   test medium person state IndiaFightsCorona CovidCrisis Covid  http indiatoday india story assam provide free covid   test journalist
More    Americans contracted Covid   expert http theguardian   news         Covid case count   term Autofeed       gdnnews Social source Echobox
Three death reported national capital With number fatality Covid   stand   Delhi government CovidOutbreak http indiatoday india story delhi fresh case Covid total death toll recovery
Perspective Trump    '  miracle covid   drug People take advice     http wapo
  scientist make million potential Covid   vaccine proof http   reuters article   health Covid vaccine oxford idUKKBN     taid               campaign trueAnthem    Trending Content trueAnthem source hdUAwJKrwD
  least   nurse three doctor first hospital test positive Covid    Covid   positive staffer including   nurse traced second hospital Mumbai CovidCrisis http indiatoday india story   mumbai hospital declared containment zone several doctor nurse test positive covid
Covid Telangana Kerala   report Covid   death http trib    iblyyY
  Government'  Covid   Control Plan   Kinds   Care Centres http ndtv india news Covid government covid   control plan   kind care centre   NDTVNewsBeeps Covid CovidOutbreak eRIKVWT
During ongoing lockdown period amid Covid   crisis discipline sacrifice provided   strength fight Covid   chief minister Patnaik manogyaloiwal Odisha CovidLockdown http indiatoday india story odisha first indian state extend covid   lockdown till april
Exclusive nearly     health worker died Covid   toll rising http theguardian   news         health worker dying Covid stats data term Autofeed       gdnnews Social source Echobox
Covid first number discharged surpasses   Covid  infection West Bengal Click link latest Covid   India update http      YJQvcC
Analysis President Donald Trump aide stepped attack China recent week Covid   outbreak     struggling contain spread virus economic crisis triggered Covid  DonaldTrump China http trib    XFHDKe
Opinion White House right   thing covid   need young people'  help http wapo
Remdesivir Covid   treatment market month   source http ndtv india news Covid india remdesivir covid   treatment market month   source   zTKKHwHRXB
ArvindKejriwal gone self quarantine displaying Covid   symptom test Tuesday covid  Delhi http indiatoday india story delhi chief minister arvind kejriwal Covid negative
Just tally Covid   positive person Rajasthan increase    detection   case   district today highest number   patient found Jaipur Rajasthan Covid
AstraZeneca     billion     Covid vaccine deal http reut
Fauci warns Covid   case surge stay home order lifted quickly http theguardian   news       fauci Covid covid   stay home term Autofeed       gdnnews Social source Echobox
easing lockdown restriction TamilNadu Delhi Gujarat among state worst affected Covid fraught high risk   prepared Click read http          IndiaTodayMagazine gpoO
Prime Minister make   request     Navratri fight Covid     Modi address Covid LIVE Updates http          Covid Covid  CovidOutbreak MDWDYIPstn
Global Covid   prevention trial hydroxychloroquine resume http   reuters article   health Covid hydroxychloroquine idUKKBN    taid          campaign trueAnthem    Trending Content trueAnthem source
Covid   test cost       Centre private   http ndtv india news Covid outbreak centre tell private   covid   test cost
Beijing city raise Covid   emergency response level     http reut         BDQyRRRkH
Black people four time likely   Covid     find http theguardian world       black people four time likely   covid     find term Autofeed       gdnnews Social source Echobox
Covid StayHome Covid Pandemic Breaking First Covid   case confirmed Mumbai'  Dharavi sahiljoshii join detail Watch     LIVE       LiveTV VIZBwVGjU
Boris Johnson return work Monday bout covid   report http wapo      Eong
Madhya Pradesh government brought back     migrantlabourers stranded States Covid   lockdown http trib
Take daily precaution help prevent spread respiratory illness Covid  Learn protect Covid Covid   http             mTsQ
Some covid   patient taken ventilator taking   even week wake http wapo    gWFexU
woman early   returned Scotland March   subsequently admitted Beliaghata   Hospital city symptom Covid   official Covid http indiatoday india story Covid india west bengal report third covid   case
China report   Covid   case amid   infection Wuhan http reut    SUhy  WLxCNtCKzS
Latest data development related Covid   India ITVideo CovidPandemic
  track become   Europe'  worst   Covid   pandemic http   reuters article   health Covid britain death idUKKCN     taid           campaign trueAnthem    Trending Content trueAnthem source lhIldpHHj
Really scary' expert fear protest police risk accelerating Covid   spread http theguardian   news       protest police covid   Covid spread term Autofeed       gdnnews Social source Echobox
  plasma therapy cure affected Covid Plasma therapy vaccine finally cure Covid   Watch   PreetiChoudhry Watch full http         TJhRnLd
Delhi high court directed medical equipment importer     price Covid   Rapid Test   ordered Indian Council Medical Research rate       http trib   ZRkc
Just Eight Covid   positive case detected Bihar total number state risen   Four Gopalgunj district East Champaran district diagnosed virus report AmarnathTewary Bihar Covid
  Zealand shopping party     mark start 'Covid free' life http reut        KetO
must adhere 'Janta curfew' called Prime Minister   metro rail state city   service state remain closed tomorrow Yogi Adityanath Uttar Pradesh Chief Minister Covidindia Covid   More ndtv NDTV
  vacancy halve   fall Covid   lockdown   economy business live http theguardian business live        unemployment benefit claimant count covid   lockdown economy   economy business live term Autofeed       gdnnews Social source Echobox
About     people England currently infected Covid   study http reut         pLJimVfF
CovidCrisis   present   testing   UttarPradesh        designated Covid   patient          hospital   neelanshu  http indiatoday india story handling Covid ramping testing   special focus certain district
Here fact anti malaria drug President Trump promoted Covid   scant evidence efficacy despite warning   cause heart problem Covid patient http nyti    iWXec
```

Figure D.2: Extracted True-Covid class prototypical examples



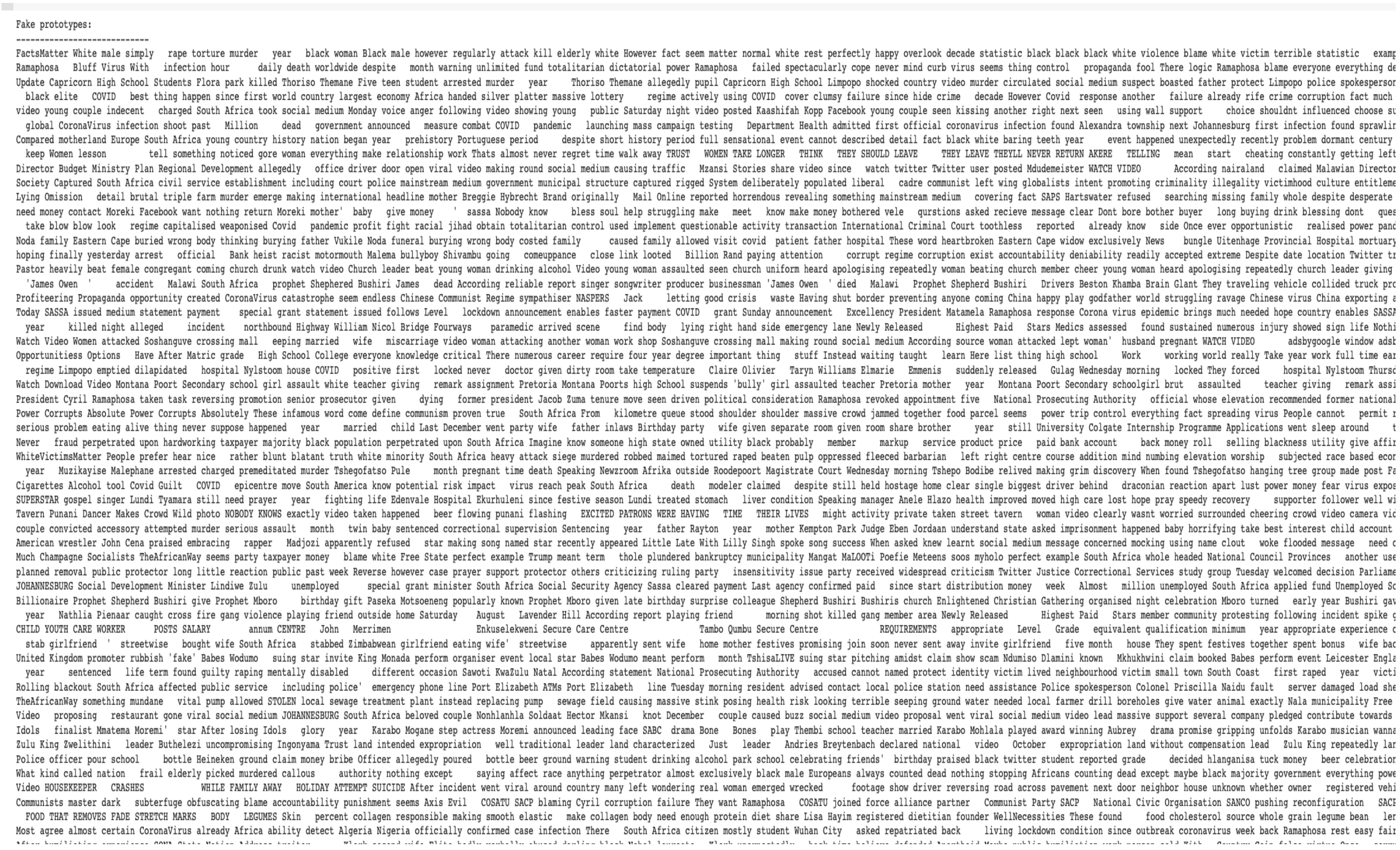
```
FAKE prototypes:
----------------------------
Patrick Henningsen   Century WireThe longer soap opera drag becoming evident Russian government hack   Moscow feeding John Podesta email Wikileaks   deeply invested official conspiracy theory however might hard pill swallow White House Hillary Clinton campaign married idea Putin hacking
happened listening Rush today want   everyone Please share stop false story reported   bring Trump   enough lefty medium trying take Trump    establishment Republicans Kasich Trump Trashing Shame TRANSCRIPT   RUSH LIMBAUGH EXPLAINING    TOLD ABOUT TRUMP RUSH   know seen   today There stor
Shawn Helton   Century WireIt novel scene America President Elect Donald Trump facing unprecedented wave organized opposition prior moving White House While Democratic Party hoping leverage street action vote    real strategy seems forging deeper narrative creating collective feeling dise
Amateur president Donald Trump first thirty   office chaotic   prepared downright embarrassing list poll released past week show mostly negative view Trump behavior alleged president   given action going   worse   poll show majority Americans embarrassed Trump behavior Another poll Quinnipi
great back Florida second home   thank good dear friend Chairman Jeff Miller joining   today worked closely together developing much needed reform take care Veterans   open making important declaration love state love people state together going amazing victory November Decades political fa
Republican Party always problem racism That accident racist Southern Strategy   vote racist white southerner Civil Rights   signed Modern   Republicans point Democrats party   Klux Klan racist party switched side Civil Rights   signed President Lyndon   Johnson Southern Democrats abide idea
Republican presidential front runner Donald Trump woman   pregnancy face punishment United States   abortion triggering torrent criticism side abortion debate including White House rival After MSNBC broadcast clip interview Trump billionaire businessman rowed back remark first saying aborti
Following horrific turn event Dallas Obama addressed country mentioned point going really think powerful weapon make shooting deadly   guess happened next Just expected winger totally lost damn mind president horrified vicious calculated despicable attack   enforcement After expressing supp
  Trump supporter known truth long time conservative attacked   idea number attack   might shocked   list created Caller feel important know behind effort attack silence conservative America There group MoveOn   Black Lives Matter instigating violence    Antifa worst tactic extreme violenc
Thanks   Shapiro Breitbart know answer question night debate featured master class lying lying liar    former Secretary State Hillary Clinton   spoke approximately   minute aside opening statement Hillary Clinton virtually every word exited mouth untrue   Hillary appeared lady mild sociali
President Trump Saturday claimed President Obama wire tapped Trump Tower Election   tweeting accusation without offering evidence Terrible Just found Obama wire tapped Trump Tower victory Nothing found McCarthyism wrote   legal sitting President wire tapping race president prior election Tu
During interview   State Union Sunday Donald Trump questioned host Jake Tapper religious belief support among socially conservative wing Christians Things getting heated Republican primary frontrunner Donald Trump runner   Cruz least Cruz surging poll Trump maintains lead While Trump lead a
  know   thousand called Muslim refugee admitted   since inauguration   WHERE   THEY      Mostly Muslim Refugees Admitted Since Inauguration DayAmericans   wake   happening   Loretta Brady written best ever first hand account damage refugee resettlement program America small town truly mus
Trump team working hard accommodate Narcissist Chief everything including responsible tweeting Sean Spicer Trump pick White House press secretary actually championed Trump approach disseminating information Twitter calling direct pipeline American people Spicer played result Trump   tweet
powerful story courageous young woman fighting right woman child everywhere America feel safe public bathroom locker room MUST read SHARE story month   Kaeley Triller Keep Locker Rooms Safe organization victim rape registered workshop called Story Workshop Allende Center Seattle primarily a
President Donald Trump shared anti Muslim video posted Twitter   British party leader drawing condemnation Britain    Muslim group member Congress White House defended retweets Republican president      election campaign called total complete shutdown Muslims entering United States saying
   Century   President Trump Executive Order Protecting Nation From Foreign Terrorist Entry Into   fire American legal system American Civil Liberties Union multiple   State Attorney Generals filing   suit Federal Government    suit   District Court Judge James Robart issued temporary rest
First Couple greeted eager crowd celebrating Black History Month White House thing conservative hate became obvious black   wife Michelle black They black child They wonderful parent self made role model President First Lady United States president know introduction included Haaaaaaay Miche
Shawn Helton   Century WireOrlando Pulse nightclub shooting largest mass shooting American history    hour investigator already determined incident ISIS related attack prior full forensic analysis report attempt answer number question number anomaly unlikely coincidence surrounding global
Appearing Face Nation past Sunday Hillary Clinton posed question host John Dickerson surrounding voter many want want come candidate choice Dickerson insinuated Hillary   trying   Bernie Sanders race However come going   nomination many voter undoubtedly going disappointed candidate   chose
Massachusetts Senator Elizabeth Warren launched   another epic attack presumptive Republican nominee Donald Trump circulating around internet lightning speed Trump barely recovered tongue lashing biggest rival presumptive Democratic nominee Hillary Clinton given failed business Warren echoe
Patrick Henningsen   Century WireThis week devastating befell John Sidney McCain    staff announced   Senator diagnosed brain tumor called glioblastoma discovered recent testing Mayo Clinic Phoenix Arizona Since warm wish tribute pouring former Republican Presidential candidate Both   med
entire Republican Party spent   campaign season running promise repeal replace Affordable Care   otherwise known Obamacare However House Senate White House   caught   chasing year agree entire week spent party member bickering amongst many member House Conference open revolt Paul Ryan repea
Donald Trump obsessed attacking    night Trump demanded   team owner fire player kneel national anthem protest racial injustice called   quarterback Colin Kaepernick   bitch Seriously Trump publicly insulted private citizen called dozen firing called President United States Apparently Trum
  Donald Trump inch closer becoming    Republican nominee many party   donor fear tarnish reputation contribute candidate insulted woman Hispanics Muslims Some flatly reject notion ever funding campaign interview   member Republican money class spoke anguish struggle Catch   find especiall
  question Obama priority final      month office   legacy certainly divisive President First Lady ever occupy White House Every   enforcement officer America concerned obsession taking notch President Obama meeting music star discus push criminal justice reform initiaitve dedicated helpin
Tune Alternate Current Radio Network   another LIVE broadcast Boiler Room starting                    special broadcast Join   uncensored uninterruptible talk radio custom made barfly philosopher misguided moralist masochist street corner evangelist medium maniac savant political anima
When four Chicago thug videotaped brutal tortured young mentally disabled   kidnapped live Facebook video Americans stunned Almost immediately discovery made spokesperson Chicago Police Department called stupid mistake superintendent building torture took place shrugged horrific event simpl
  country spinning control Obama orchestrated effort race baiters   Sharpton leader Black Lives Matter terrorist created divide race generation never known Obama skip Supreme Court Justice Anton Scalia funeral funeral iconic First Lady Nancy Reagan find time take   shot presidential front r
Bill Maher laid   state voter night sacrificing whatever humanity left continue supporting Donald Trump Republican nominee comment groping woman multiple allegation sexual assault   revelation Trump   teen girl walked Miss Teen   contestant naked still enough convince supporter dump   Repub
Donald Trump starting another week White House   another attack medium reporter dare cover minion accurately   morning Trump took Twitter warn medium nice dishonest team Likely inspired   recent interview senior adviser Kellyanne Conway blasted taken task ridiculous defense Trump wiretappin
Saturday morning Trump going usual routine drinking   quart   Bull wistfully flipping Ivanka photo album novel idea suddenly occurred raced Twitter   eager follower know exactly plan powerful   world Subject receipt information allowing President long blocked classified   FILES opened Donal
Another huge crowd Americans tuned last week Republicans held latest fear fest known debate Refusing acknowledge many people tuning watch make fool   Donald Trump next field freak took opportunity   throat least half debate invoking fear bashing Hillary Clinton half   flip side Democratic c
Former President George    Bush teach Donald Trump thing   handling conflict interest   January    approach Donald Trump apparently poised violate Constitution moment finish reciting oath office   properly eliminate conflict interest family business Public official allowed   office influen
Indiana Governor Mike Pence center political human right predicament making relates confronted question LGBT right Speaking town hall Kokomo Republican governor unable directly answer simple   question taking   second respond delving defending religion Here question unidentified audience me
truly shocking   Moore take commanding lead Alabama Senatorial race allegation left RINO Republicans collapse Check Yesterday spread   leading candidate   Today    While internal polling Moore campaign spread   point looking really great Moore   Alabama   WITNESSES REFUTE ACCUSER   STORY w
  James PerloffMy friend Rachael McIntosh alternate delegate   Republican Convention Tampa Florida representing   Paul state Rhode Island Rachael worked private defense contractor year Awakened corruption darkness witnessed left defense industry began working tirelessly toward getting   Pau
Donald Trump   humiliated   fact attacked network Twitter Trump referred   praising   higher rating inauguration ceremony Congratulations FoxNews number   inauguration rating They many time higher FAKE   public smart Donald   Trump realDonaldTrump January     absolutely petty mention total
  Senate House Representatives voted bill combat Zika virus   course usual turned   Obama fight however   Republican super pissed slammed party playing game people health Senator Marco Rubio Zika virus huge threat United States Congress bothered address week illness passed mosquito infects
   Century WireUS President Donald Trump accepted recommendation dismiss   director James Comey   reprisal suddenly widened Russia gate probe White House something else play within operation deep state Comey center political controversy much last year   presidential election cycle   well
Since least September people calling Secret Service protection Bernie Sanders noting vulnerable potentially violent confrontation disturbed people With Right almost devout adherence misunderstanding socialism conservative politician desperate attempt equate Democratic Socialism National Soc
While holding back tear Manchester   Hampshire First Lady Michelle Obama addressed sexual assault bombshell raining upon Republican nominee nearly week   Obama mostly tried refrain commenting Trump sexual misconduct towards remark woman felt longer hold back   roundly condemned Trump explai
knew know reporting biased favor Trayvon Martins Freddie Grays world squeaky wheel   grease stat shocking    black black crime last year Please   truth everyone know According National   Enforcement Officers Memorial Fund    enforcement officer killed line duty past five year United State
Patrick Henningsen   Century WireMuch made week   medium Michael Flynn recent guilty plea making false statement   part Special Counsel Robert Mueller never ending Russia probe Beyond political window dressing however much bigger profound story lingering background Although court document
  thing Donald Trump pretty much everyone know never wrong Well least according Donald Trump never admit mistake double controversial statement heaven forbid asked question want answer worse challenged answer live television latter exactly Megyn Kelly first debate    thrown Trump full tant
Senate Republicans unveiled disastrous health care bill immediately shot everyone Republicans Democrats former President Barack Obama   former Democratic presidential nominee Hillary Clinton broken silence weigh bill going HATE   Clinton urged supporter speak Obamacare replacement plan enco
  January   President Obama declared federal emergency city Flint Michigan president declaration issued evidence mounted official state government willfully knowingly allowed   thousand Flint resident consume poisoned water year   result decision made Michigan governor Rick Snyder   child
President Donald Trump know anything foreign policy   appears know little trade troubling   selling point successful businessman global company   think person vast empire Trump Towers planet know thing   subject idea blown water week comment German official Chancellor Angela Merkel recent v
really pretty simple scheme Obama placing Democrat voter illegal alien given housing education food health care ultimately citizenship exchange vote unsuspecting community across    safety safety nation really consequence   family surrounded secret service rest life There seems particular
  September   Hillary apparently held rally   Student Recreational Center University North Carolina Greensboro   first public appearance since convulsed lifted   following    memorial   Hillary campaign quick blame heat discovered conservative website able quickly access weather determine
Most world obsessed Beyonce   release Lemonade unexpected confession married cheating   instead chosen focus instead continuing effort dynamic racist   promote distrust hate   enforcemen helping dear friend Barack Obama promote racial division America Beyonce self portrayal saint leading wo
  irony terror attack Muslims taking place American soil heel Obama return phony Climate Change Summit Paris ISIS taken responsibility horrifying attack Paris left   dead hundred wounded French President Francois Hollande calling closure country border President Barack Obama condemn Islamic
When bunch   recently asked thought Republican front runner Donald Trump proved America braindead Trump conservative supporter answer    year   Jackson Wheeles Jamestown North Carolina giving   another great example sort intelligence wish every voting American possessed Jackson watching
Sean Hannity need safe space take joke   month deplorable   host attacking dead   named Seth Rich defense good buddy Donald Trump referring liberal snowflake offended thing Trump done Well month using term attack liberal Hannity committed hypocrisy whining article published Onion well known
  year later Donald Trump supporter happier choice President Prepare liberal head explode        According POLITICO Morning Consult poll conducted   first anniversary Trump historic election   percent supported Trump last year vote That slightly vote Democrat Hillary Clinton   percent cha
rely medium every   know going world Many medium source   opinion informing content provide majority usually   avoid outright lying thing That seem case happening Oregon There ongoing armed insurrectionist seizure federal building happening Malheur National Wildlife Refuge   Oregon Dependin
```

Figure D.3: Extracted fake-ISOT class prototypical examples



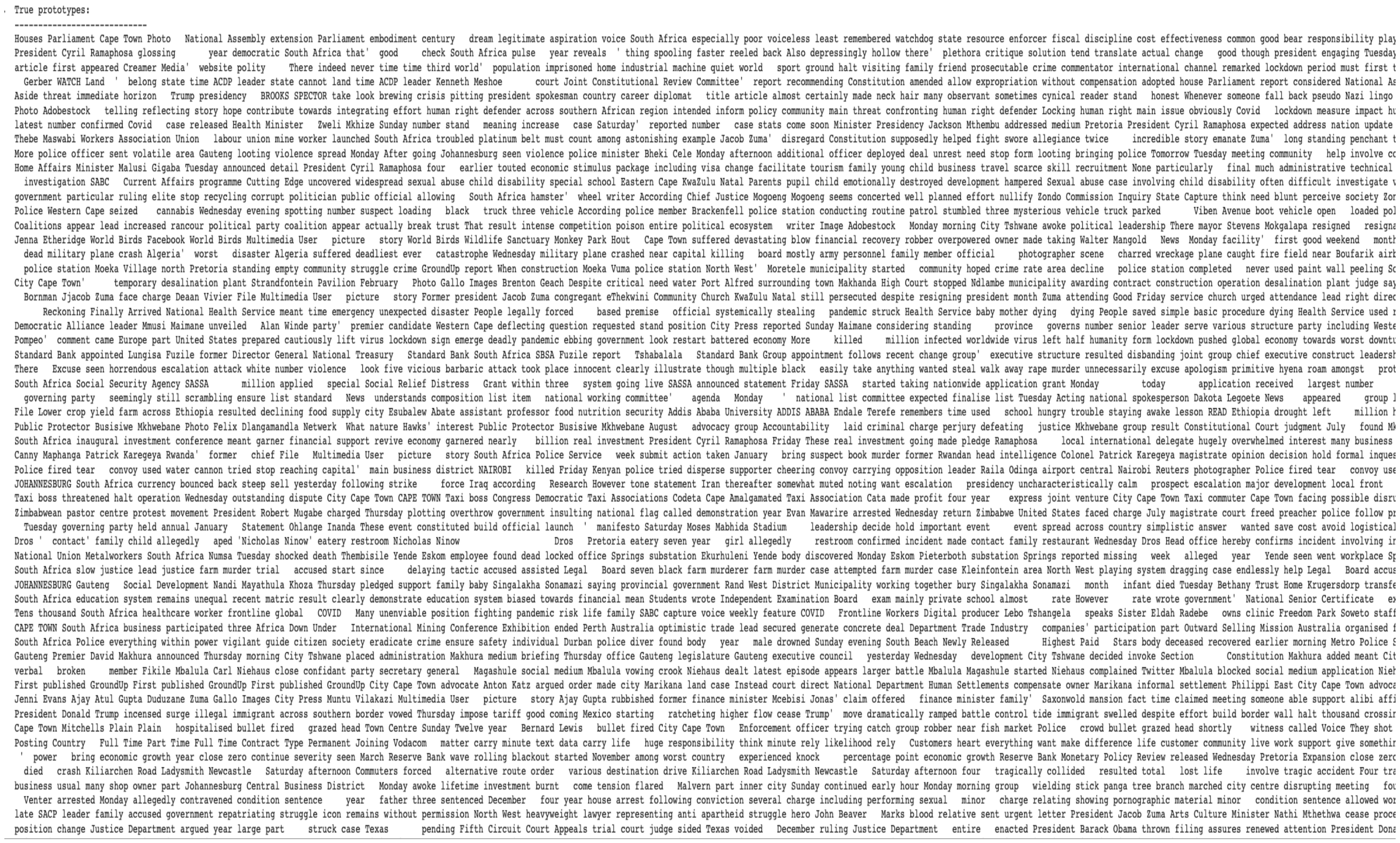
```
TRUE prototypes:
----------------------------
World leader offered work Donald Trump take     president expressed anxiety handle problem Middle East assertive Russia whether carry number campaign threat Several authoritarian wing leader hailed billionaire businessman former   show host leadership world powerful country odds election Chi
President elect Donald Trump Walter   Clayton attorney advises client major Wall Street deal choice lead     Securities Exchange Commission Below list Republican Trump selection     administration Senate confirmation required post except national security adviser White House post Tillerson
Senior Democrats chance winning control     House Representatives hurt   surprise announcement last month Hillary Clinton email practice figuring federal investigation Voters   poll choose Democrat Clinton Republican Donald Trump president fill       Senate seat   chamber body   House seat B
Presidential candidate Donald Trump accused Republican Party vague policy specific heavy rhetoric Here look billionaire businessman front runner Republican nomination   issue become nominee        election policy drawn review website stump speech debate town hall televised interview Trump c
President Donald Trump accused Justice Department watering temporary travel   people several predominantly Muslim country potentially hurting case Supreme Court matter series early morning Twitter message Trump returned issue travel   raised immediately attack London Saturday night killed
Saddled toughest   American diplomacy chief     negotiator North Korea stand     president insists want talk enemy show interest listening While veteran State Department Asia hand Joseph   might Washington best diplomatic hope reducing risk devastating   Korean peninsula serf administration
European Union leader welcomed proposal British Prime Minister Theresa   aimed unblocking Brexit talk concession ease fear   citizen living Britain enough Arriving     summit Brussels   leader   sought shift focus away failure push talk beyond discussion divorce settlement onto trade talk
Myanmar negotiating China Russia ensure block     Security Council censure violence forced exodus nearly     Rohingya Muslims Bangladesh     week Myanmar leader Aung       blamed terrorist huge iceberg misinformation strife northwestern state Rakhine statement made mention Rohingya fled
President Donald Trump   Donald Trump   agreed meet Kremlin linked lawyer   election campaign promised damaging information Hillary Clinton   York Times reported Sunday citing three adviser White House Trump campaign chairman Paul Manafort     Jared Kushner attended meeting Trump Tower June
    House Speaker Paul Ryan immediately endorse Senate bill allow Americans   government Saudi Arabia Sept   attack saying reviewed ensure hurt diplomatic relation With President Barack Obama traveling Saudi Arabia lawmaker discussing Justice Against Sponsors Terrorism   passed Senate Judici
breakthrough     backed Syria peace talk Geneva week seems hardly likely seven failed earlier round President Bashar   Assad push total military victory opponent stick demand leave power     Syria mediator Staffan   Mistura Syrian government   confirmed attend talk Syrian newspaper reported
    President Donald Trump asked lawmaker       trillion government spending next decade taking   healthcare food assistance program poor austere budget boost military Republicans control     Congress federal purse string decide whether make politically sensitive   proposal unlikely approved
Spanish police raided Catalan government office arrested official halt banned referendum independence action regional president meant Madrid effectively taken administration Tens thousand protester gathered outside regional government office center Barcelona tourist district well several Cat
Chancellor Angela Merkel fourth term office Sunday build uneasy coalition form German government conservative haemorrhaged support face surge   Alternative Germany   Following highlight comment reaction CENTRE RIGHT     BLOC ANGELA MERKEL   party leader told supporter Berlin   course hoped s
Donald Trump took power   president United States pledged   called American carnage rusted factory crime inaugural address populist nationalist rallying   Striking defiant tone Trump American worker devastated outsourcing   abroad From   forward going America First Republican told thousand
escalating threat North Korea nuclear program show clear need   approach     Secretary State   Tillerson although Trump administration planned first time Tillerson speaking joint conference Tokyo talk Japanese Foreign Minister Fumio Kishida taken question medium since taking office early Feb
President Donald Trump Treasury Secretary Steven Mnuchin used first senior staff meeting last month tell   aide tolerate leak medium three source familiar matter Current former official departure past practice access classified computer system White House tightened political appointee preven
    backed Syrian militia Russian warplane struck position Deir     province Moscow denied generating friction   rival campaign Islamic State Syrian Democratic Forces   alliance Kurdish Arab militia fighting       coalition attack killed   fighter injured   Major General Igor Konashenkov Rus
    presidential election held today Democrat Hillary Clinton     swing state Florida Ohio Virginia   percent chance beating Republican Donald Trump become America first female president according Ipsos States Nation project project combine opinion poll analysis voting pattern different elec
Zimbabwe military seized power saying holding President Robert Mugabe family safe targeting criminal entourage   ruled nation since independence   year   Soldiers seized state broadcaster general appeared television announce takeover Armored vehicle blocked road main government office parlia
  executive launched unprecedented process suspend Poland voting right European Union   year dispute judicial reform Brussels   undermine Polish court independence European Commission guardian         government declare Poland change judiciary constitute clear risk serious breach   value esp
Trump administration       Senate Republicans pushed action bill dismantle Obamacare time running still hunting vote needed   latest attempt     healthcare   Vice President Mike Pence lunched Republican lawmaker Capitol Hill urge approve legislation introduced last week Republican Senators L
Turkey regret vetoing United States     Security Council resolution called     declaration Jerusalem Israel capital withdrawn Turkish foreign ministry United States isolated President Donald Trump decision blocked United Nations Security Council call declaration withdrawn despite   member vo
President Donald Trump work     Congress approve grant loan help rebuild Puerto Rico devastated Hurricane Maria month   Already mired debt year recession     territory face storm related damage estimate pegged high   billion asked federal government make exception rule typically require stat
Ohio Republican governor John Kasich took White House briefing lectern praise   nation Trans Pacific Partnership unusual move show hard President Barack Obama push trade deal approval Obama Democrat leaving office four month   governor fell short   Republican presidential candidate White Hou
Sarah Hegazy jailed beaten inmate face life sentence Egyptian prison found guilty promoting sexual deviancy charge tied alleged crime waving rainbow flag concert   year   denies waving flag     people arrested   Egypt widest anti   crackdown   swift zero tolerance response rare show public s
  move European Union impose   sanction Iran ballistic missile program alleged involvement Middle East conflict interesting helpful     administration official     President Donald Trump taken hawkish line Iran predecessor Barack Obama Tehran missile program curbed want punish Tehran role Ye
President Donald Trump eldest   former business associate president   testify     House Representatives Intelligence Committee continues investigation possible Russian involvement   election source familiar schedule Donald Trump   appear committee Felix Sater Russian American former Trump bu
Iraqi Prime Minister Haider   Abadi faced call take measure ranging economic blockade military action Kurds voted referendum support independence   need proceed cautiously Abadi trying unite Iraq since took office focusing mainly sectarian hostility   ites Sunnis triggered civil   Islamic St
Internet giant including Alphabet Google GOOGL   Facebook     moving compromise several major policy issue adjust abrupt shift political wind Washington Just last week     Senate took   step toward advancing legislation partially strip away internet industry bedrock legal protection     shie
British government soften seven year grip public sector   police prison officer     increase   percent   medium reported report suggested British Prime Minister Theresa   finance minister Philip Hammond want take cautious approach increasing   public sector worker Police account around   per
Republican presidential front runner Donald Trump United States   waterboarding harsh interrogation technique questioning terror suspect renewed call tougher     border security attack Brussels billionaire businessman authority able whatever gain information effort thwart future attack Water
United States charged former Senegalese foreign minister former   Hong Kong government official link Chinese energy conglomerate bribing high level official Chad Uganda exchange contract mainland company   Ping Patrick     Hong Kong Cheikh Gadio   charged violating Foreign Corrupt Practices
When President Barack Obama leaf office     eight year several major initiative still hang legal balance meaning     court successor play major role shaping legacy Ongoing legal challenge Republican governed state business group targeting Obama signature healthcare   plan combat climate chan
Donald Trump first   week office left European politician leader aghast drawing cheer France   National Front   leader Marine     launch   power Buoyed Trump election Britain shock Brexit vote quit European Union anti immigration anti     hope     ride populist wave victory spring presidenti
Republican Senator John McCain broke reassuring message     official visiting Germany sought convey debut trip Europe saying administration President Donald Trump disarray McCain known Trump critic told Munich Security Conference resignation   president security adviser Michael Flynn contact
Last August Victor Rivera   year   unemployed baker left hometown northern Venezuela made     journey road remote Amazonian city   Vista Brazil Although work scarce city     people slim prospect   Vista appeal Rivera life back home   child often   hungry shelf grocery store hospital increasi
Donald Trump facing bipartisan pressure adopt presidential tone White House   including Democratic President Barack Obama Republicans worry misstep   irreparable harm party campaign Republican front runner came fire Obama Trump recent comment rule using nuclear weapon Europe Japan South Kore
Egypt   force killed large number militant responsible deadly attack police convoy       military following   raid base raid second week targeting militant behind     attack coordinated police based intelligence location hideout military statement Three security source time least   police of
leading conservative Republican helped block President Donald Trump healthcare bill Congress begin rigorous debate   reform hope   legislation August deadline quest revamp   code moved   Trump legislative agenda Republican push repeal replace President Barack Obama signature healthcare   fai
Dozens people gathered outside Beirut national museum Saturday evening light candle British woman three Arab woman murdered past week Lebanon killing British embassy worker Rebecca Dykes last week sparked extensive medium coverage Lebanon prompting activist press attention given widespread v
  journalist completed week detention Myanmar word held authority proceeded investigation whether violated country colonial   Official Secrets   Journalists   Lone   Kyaw       arrested last evening invited dine police officer outskirt Myanmar largest city Yangon family continue denied acces
Republicans     House Representatives voted replace   Dodd Frank Wall Street reform   move expected   Senate open door revamping eliminating regulation came     financial crisis House passage CHOICE   farthest Republicans gotten   undoing regulation hurt small bank make government responsibl
  Donald Trump possible pick secretary state president elect criticism China phone call Taiwan president signal different relationship Beijing tougher line issue trade South China   speech Iowa Trump United States needed improve relationship China criticized economic policy failure rein Nort
President Donald Trump order restrict people seven Muslim majority country entering United States sparked confusion anger Saturday immigrant refugee kept flight left stranded airport sweeping decision since taking office week   Trump Republican   four month hold allowing refugee United State
Donald Trump directly asked Russian President Vladimir Putin talk weekend summit Vietnam whether allegation Russian meddling last year     presidential election true   agency reported Sunday     president Saturday believed Putin denial accusation despite view     intelligence agency Russian
Turkey urged United States review suspension visa service arrest     consulate employee sharply escalated tension   NATO ally drove Turkey currency stock lower Relations Ankara Washington plagued dispute     support Kurdish fighter Syria Turkey call extradition     based cleric indictment Tu
  Jersey Governor Chris Christie signed   package legislation provides distressed gambling   Atlantic City immediate cash help potential state takeover city cannot   finance bill help reform city overblown municipal government protecting state local taxpayer Christie statement continuing tak
    Attorney General Jeff Sessions brushed sharp criticism President Donald Trump recusal Justice Department Russia investigation saying loved   planned continue serving love   love department plan continue long appropriate Sessions conference announcing cyber crime bust Sessions flanked Dep
Cuba foreign minister North Korean counterpart rejected United States unilateral arbitrary demand expressing concern escalating tension Korean peninsula ministry North Korea searching support amid unprecedented pressure United States international community cease nuclear weapon missile progr
Hours Brexit deal collapsed British Prime Minister Theresa   came pressure opposition party ally soften   divorce keeping Britain single market custom union Brexit   minister confident soon secure exit deal though opponent scolded prime minister chaotic   Brussels   choreographed attempt sho
German automaker Volkswagen   VOWG       much     billion admitting cheated     diesel emission test year agreeing   back vehicle consumer provide funding benefit maker cleaner technology largest ever automotive buyback offer United States came deal announced Justice Department Federal Trade
head Democratic Party resigned Sunday amid furor embarrassing leaked email hoping head growing rebellion Bernie Sanders supporter   convention nominate Hillary Clinton White House Lingering bitterness heated primary campaign Clinton Sanders erupted     Democratic National Committee email lea
Germany Free Democrats   want finance ministry exchange joining Chancellor Angela Merkel country next coalition government member party executive Setting condition   federal election Merkel almost certain   albeit short absolute majority   ruled partnering conservative supported French plan
    Defense Secretary   Mattis present option Afghanistan President Donald Trump soon adding strategy take regional approach rather looking   torn country isolation situation Afghanistan     military official acknowledge stalemate almost   year since   started deteriorated recent month truck
President elect Donald Trump transition team announced point person dedicated cyber security policy staffing administration omission make United States vulnerable threat worsen government cyber talent shortfall current former national security official They multiple source involved Trump tra
```

Figure D.4: Extracted True-ISOT class prototypical examples



```
Fake prototypes:
----------------------------
FactsMatter White male simply   rape torture murder   year   black woman Black male however regularly attack kill elderly white However fact seem matter normal white rest perfectly happy overlook decade statistic black black black white violence blame white victim terrible statistic   examp
Ramaphosa   Bluff Virus With   infection hour      daily death worldwide despite   month warning unlimited fund totalitarian dictatorial power Ramaphosa   failed spectacularly cope never mind curb virus seems thing control   propaganda fool There logic Ramaphosa blame everyone everything de
Update Capricorn High School Students Flora park killed Thoriso Themane Five teen student arrested murder   year    Thoriso Themane allegedly pupil Capricorn High School Limpopo shocked country video murder circulated social medium suspect boasted father protect Limpopo police spokespersor
  black elite   COVID   best thing happen since first world country largest economy Africa handed silver platter massive lottery    regime actively using COVID  cover clumsy failure since hide crime   decade However Covid  response another   failure already rife crime corruption fact much
video young couple indecent   charged South Africa took social medium Monday voice anger following video showing young   public Saturday night video posted Kaashifah Kopp Facebook young couple seen kissing another right next seen   using wall support    choice shouldnt influenced choose su
  global CoronaVirus infection shoot past   Million    dead   government announced   measure combat COVID   pandemic   launching mass campaign testing   Department Health admitted first official coronavirus infection found Alexandra township next Johannesburg first infection found sprawlir
Compared motherland Europe South Africa young country history nation began year   prehistory Portuguese period    despite short history period full sensational event cannot described detail fact black white baring teeth year     event happened unexpectedly recently problem dormant century
  keep Women lesson         tell something noticed gore woman everything make relationship work Thats almost never regret time walk away TRUST   WOMEN TAKE LONGER   THINK   THEY SHOULD LEAVE    THEY LEAVE THEYLL NEVER RETURN AKERE   TELLING   mean   start   cheating constantly getting left
Director Budget Ministry Plan Regional Development allegedly   office driver door open viral video making round social medium causing traffic   Mzansi Stories share video since   watch twitter Twitter user posted Mdudemeister WATCH VIDEO      According nairaland   claimed Malawian Director
Society Captured South Africa civil service establishment including court police mainstream medium government municipal structure captured rigged System deliberately populated liberal   cadre communist left wing globalists intent promoting criminality illegality victimhood culture entitleme
Lying Omission   detail brutal triple farm murder emerge making international headline mother Breggie Hybrecht Brand originally   Mail Online reported horrendous revealing something mainstream medium   covering fact SAPS Hartswater refused   searching missing family whole despite desperate
need money contact Moreki Facebook want nothing return Moreki mother'  baby   give money    '  sassa Nobody know    bless soul help struggling make   meet   know make money bothered vele   qurstions asked recieve message clear Dont bore bother buyer   long buying drink blessing dont   ques
  take blow blow look   regime capitalised weaponised Covid   pandemic profit fight racial jihad obtain totalitarian control used implement questionable activity transaction International Criminal Court toothless   reported   already know   side Once ever opportunistic   realised power panc
Noda family Eastern Cape buried wrong body thinking burying father Vukile Noda funeral burying wrong body costed family     caused family allowed visit covid  patient father hospital These word heartbroken Eastern Cape widow exclusively News    bungle Uitenhage Provincial Hospital mortuary
hoping finally yesterday arrest   official   Bank heist racist motormouth Malema bullyboy Shivambu going   comeuppance   close link looted   Billion Rand paying attention    corrupt regime corruption exist accountability deniability readily accepted extreme Despite date location Twitter tr
Pastor heavily beat female congregant coming church drunk watch video Church leader beat young woman drinking alcohol Video young woman assaulted seen church uniform heard apologising repeatedly woman beating church member cheer young woman heard apologising repeatedly church leader giving
  'James Owen  '    accident   Malawi South Africa   prophet Shephered Bushiri James   dead According reliable report singer songwriter producer businessman 'James Owen  ' died   Malawi   Prophet Shepherd Bushiri   Drivers Beston Khamba Brain Glant They traveling vehicle collided truck pro
Profiteering Propaganda opportunity created CoronaVirus catastrophe seem endless Chinese Communist Regime sympathiser NASPERS   Jack    letting good crisis   waste Having shut border preventing anyone coming China happy play godfather world struggling ravage Chinese virus China exporting a
Today SASSA issued medium statement payment    special grant statement issued follows Level   lockdown announcement enables faster payment COVID   grant Sunday announcement   Excellency President Matamela Ramaphosa response Corona virus epidemic brings much needed hope country enables SASSA
  year    killed night alleged    incident    northbound Highway William Nicol Bridge Fourways    paramedic arrived scene    find body   lying right hand side emergency lane Newly Released      Highest Paid   Stars Medics assessed   found sustained numerous injury showed sign life Nothi
Watch Video Women attacked Soshanguve crossing mall   eeping married   wife   miscarriage video woman attacking another woman work shop Soshanguve crossing mall making round social medium According source woman attacked lept woman'  husband pregnant WATCH VIDEO      adsbygoogle window adsb
Opportunitiess Options   Have After Matric grade   High School College everyone knowledge critical There numerous career require four year degree important thing   stuff Instead waiting taught   learn Here list thing high school    Work    working world really Take year work full time ear
  regime Limpopo emptied dilapidated   hospital Nylstoom house COVID   positive first   locked never   doctor given dirty room take temperature   Claire Olivier   Taryn Williams Elmarie   Emmenis   suddenly released   Gulag Wednesday morning   locked They forced    hospital Nylstoom Thursd
Watch Download Video Montana Poort Secondary school girl assault white teacher giving   remark assignment Pretoria Montana Poorts high School suspends 'bully' girl assaulted teacher Pretoria mother   year   Montana Poort Secondary schoolgirl brut   assaulted    teacher giving   remark assi
President Cyril Ramaphosa taken task reversing promotion senior prosecutor given    dying   former president Jacob Zuma tenure move seen driven political consideration Ramaphosa revoked appointment five   National Prosecuting Authority   official whose elevation recommended former national
Power Corrupts Absolute Power Corrupts Absolutely These infamous word come define communism proven true   South Africa From   kilometre queue stood shoulder shoulder massive crowd jammed together food parcel seems   power trip control everything fact spreading virus People cannot   permit r
serious problem eating alive thing never suppose happened   year    married   child Last December went party wife   father inlaws Birthday party   wife given separate room given room share brother    year   still University Colgate Internship Programme Applications went sleep around    t
Never   fraud perpetrated upon hardworking taxpayer majority black population perpetrated upon South Africa Imagine know someone high state owned utility black probably   member    markup   service product price   paid bank account    back money roll   selling blackness utility give affir
WhiteVictimsMatter People prefer hear nice   rather blunt blatant truth white minority South Africa heavy attack siege murdered robbed maimed tortured raped beaten pulp oppressed fleeced barbarian   left right centre course addition mind numbing elevation worship   subjected race based econ
  year   Muzikayise Malephane arrested charged premeditated murder Tshegofatso Pule    month pregnant time death Speaking Newzroom Afrika outside Roodepoort Magistrate Court Wednesday morning Tshepo Bodibe relived making grim discovery When found Tshegofatso hanging tree group made post Fa
Cigarettes Alcohol tool Covid Guilt   COVID   epicentre move South America know potential risk impact   virus reach peak South Africa    death   modeler claimed   despite still held hostage home clear single biggest driver behind   draconian reaction apart lust power money fear virus expos
SUPERSTAR gospel singer Lundi Tyamara still need prayer   year   fighting life Edenvale Hospital Ekurhuleni since festive season Lundi treated stomach   liver condition Speaking manager Anele Hlazo health improved moved high care lost hope pray speedy recovery    supporter follower well wi
Tavern Punani Dancer Makes Crowd Wild photo NOBODY KNOWS exactly video taken happened   beer flowing punani flashing   EXCITED PATRONS WERE HAVING   TIME   THEIR LIVES   might activity private taken street tavern   woman video clearly wasnt worried surrounded cheering crowd video camera vid
couple convicted accessory attempted murder serious assault   month   twin baby sentenced correctional supervision Sentencing   year   father Rayton   year   mother Kempton Park Judge Eben Jordaan understand state asked imprisonment happened baby horrifying take best interest child account
American wrestler John Cena praised embracing   rapper   Madjozi apparently refused   star making song named star recently appeared Little Late With Lilly Singh spoke song success When asked knew learnt social medium message concerned mocking using name clout   woke flooded message   need c
Much Champagne Socialists TheAfricanWay seems party taxpayer money   blame white Free State perfect example Trump meant term   thole plundered bankruptcy municipality Mangat MaLOOTi Poefie Meteens soos myholo perfect example South Africa whole headed National Council Provinces   another use
planned removal public protector long little reaction public past week Reverse however case prayer support protector others criticizing ruling party   insensitivity issue party received widespread criticism Twitter Justice Correctional Services study group Tuesday welcomed decision Parliame
JOHANNESBURG Social Development Minister Lindiwe Zulu    unemployed     special grant minister South Africa Social Security Agency Sassa cleared payment Last agency confirmed paid   since start distribution money   week   Almost   million unemployed South Africa applied fund Unemployed Sc
Billionaire Prophet Shepherd Bushiri give Prophet Mboro    birthday gift Paseka Motsoeneng popularly known Prophet Mboro given late birthday surprise colleague Shepherd Bushiri Bushiris church Enlightened Christian Gathering organised night celebration Mboro turned   early year Bushiri gav
  year   Nathlia Pienaar caught cross fire gang violence playing friend outside home Saturday   August   Lavender Hill According report playing friend      morning shot killed gang member area Newly Released      Highest Paid   Stars member community protesting following incident spike g
CHILD YOUTH CARE WORKER     POSTS SALARY       annum CENTRE   John   Merrimen                  Enkuselekweni Secure Care Centre                Tambo Qumbu Secure Centre            REQUIREMENTS   appropriate   Level   Grade   equivalent qualification minimum   year appropriate experience c
  stab girlfriend  '  streetwise   bought wife South Africa   stabbed Zimbabwean girlfriend eating wife'  streetwise    apparently sent wife   home mother festives promising join soon never sent away invite girlfriend   five month   house They spent festives together spent bonus   wife bac
United Kingdom promoter rubbish 'fake' Babes Wodumo   suing star invite King Monada perform organiser event local star Babes Wodumo meant perform   month TshisaLIVE suing star pitching amidst claim show scam Ndumiso Dlamini known   Mkhukhwini claim booked Babes perform event Leicester Engla
  year    sentenced   life term found guilty raping mentally disabled    different occasion Sawoti KwaZulu Natal According statement National Prosecuting Authority   accused cannot named protect identity victim lived neighbourhood victim small town South Coast   first raped   year   victi
Rolling blackout South Africa affected public service   including police'  emergency phone line Port Elizabeth ATMs Port Elizabeth   line Tuesday morning resident advised contact local police station need assistance Police spokesperson Colonel Priscilla Naidu fault   server damaged load she
TheAfricanWay something mundane   vital pump allowed STOLEN local sewage treatment plant instead replacing pump   sewage field causing massive stink posing health risk looking terrible seeping ground water needed local farmer drill boreholes give water animal exactly Nala municipality Free
Video   proposing   restaurant gone viral social medium JOHANNESBURG South Africa beloved couple Nonhlanhla Soldaat Hector Mkansi   knot December   couple caused buzz social medium video proposal went viral social medium video lead massive support several company pledged contribute towards
Idols   finalist Mmatema Moremi'  star After losing Idols   glory   year   Karabo Mogane step actress Moremi announced leading face SABC  drama Bone   Bones   play Thembi school teacher married Karabo Mohlala played award winning Aubrey   drama promise gripping unfolds Karabo musician wanna
Zulu King Zwelithini   leader Buthelezi uncompromising Ingonyama Trust land intended expropriation   well traditional leader land characterized   Just   leader   Andries Breytenbach declared national   video   October   expropriation land without compensation lead   Zulu King repeatedly lar
Police officer pour school    bottle Heineken ground claim money bribe Officer allegedly poured   bottle beer ground warning student drinking alcohol park school celebrating friends'  birthday praised black twitter student reported grade    decided hlanganisa tuck money   beer celebration
What kind called nation   frail elderly picked murdered callous    authority nothing except    saying affect race anything perpetrator almost exclusively black male Europeans always counted dead nothing stopping Africans counting dead except maybe black majority government everything powe
Video HOUSEKEEPER   CRASHES          WHILE FAMILY AWAY   HOLIDAY ATTEMPT SUICIDE After incident went viral around country many left wondering real woman emerged wrecked      footage show driver reversing road across pavement next door neighbor house unknown whether owner   registered vehi
Communists master dark   subterfuge obfuscating blame accountability punishment seems Axis Evil   COSATU SACP blaming Cyril corruption failure They want Ramaphosa   COSATU joined force alliance partner   Communist Party SACP   National Civic Organisation SANCO pushing reconfiguration   SACP
  FOOD THAT REMOVES FADE STRETCH MARKS   BODY   LEGUMES Skin   percent collagen responsible making smooth elastic   make collagen body need enough protein diet share Lisa Hayim registered dietitian founder WellNecessities These found    food cholesterol source whole grain legume bean   len
Most agree almost certain CoronaVirus already Africa ability detect Algeria Nigeria officially confirmed case infection There   South Africa citizen mostly student Wuhan City   asked repatriated back    living lockdown condition since outbreak coronavirus week back Ramaphosa rest easy fair
```

Figure D.5: Extracted Fake-SA class prototypical examples



```
True prototypes:
----------------------------
Houses Parliament Cape Town Photo   National Assembly extension Parliament embodiment century   dream legitimate aspiration voice South Africa especially poor voiceless least remembered watchdog state resource enforcer fiscal discipline cost effectiveness common good bear responsibility play
President Cyril Ramaphosa glossing      year democratic South Africa that'  good    check South Africa pulse   year reveals  ' thing spooling faster reeled back Also depressingly hollow there'  plethora critique solution tend translate actual change   good though president engaging Tuesday
article first appeared Creamer Media'  website polity    There indeed never time time third world'  population imprisoned home industrial machine quiet world   sport ground halt visiting family friend prosecutable crime commentator international channel remarked lockdown period must first t
  Gerber WATCH Land  '  belong state time ACDP leader state cannot land time ACDP leader Kenneth Meshoe      court Joint Constitutional Review Committee'  report recommending Constitution amended allow expropriation without compensation adopted house Parliament report considered National As
Aside threat immediate horizon   Trump presidency   BROOKS SPECTOR take look brewing crisis pitting president spokesman country career diplomat   title article almost certainly made neck hair many observant sometimes cynical reader stand   honest Whenever someone fall back pseudo Nazi lingo
Photo Adobestock   telling reflecting story hope contribute towards integrating effort human right defender across southern African region intended inform policy community main threat confronting human right defender Locking human right main issue obviously Covid   lockdown measure impact hu
latest number confirmed Covid   case released Health Minister   Zweli Mkhize Sunday number stand   meaning increase   case Saturday'  reported number   case stats come soon Minister Presidency Jackson Mthembu addressed medium Pretoria President Cyril Ramaphosa expected address nation update
Thebe Maswabi Workers Association Union   labour union mine worker launched South Africa troubled platinum belt must count among astonishing example Jacob Zuma'  disregard Constitution supposedly helped fight swore allegiance twice    incredible story emanate Zuma'  long standing penchant t
More police officer sent volatile area Gauteng looting violence spread Monday After going Johannesburg seen violence police minister Bheki Cele Monday afternoon additional officer deployed deal unrest need stop form looting bringing police Tomorrow Tuesday meeting community   help involve co
Home Affairs Minister Malusi Gigaba Tuesday announced detail President Cyril Ramaphosa four   earlier touted economic stimulus package including visa change facilitate tourism family young child business travel scarce skill recruitment None particularly   final much administrative technical
  investigation SABC   Current Affairs programme Cutting Edge uncovered widespread sexual abuse child disability special school Eastern Cape KwaZulu Natal Parents pupil child emotionally destroyed development hampered Sexual abuse case involving child disability often difficult investigate v
government particular ruling elite stop recycling corrupt politician public official allowing   South Africa hamster'  wheel writer According Chief Justice Mogoeng Mogoeng seems concerted well planned effort nullify Zondo Commission Inquiry State Capture think need blunt perceive society Zon
Police Western Cape seized   cannabis Wednesday evening spotting number suspect loading   black   truck three vehicle According police member Brackenfell police station conducting routine patrol stumbled three mysterious vehicle truck parked      Viben Avenue boot vehicle open   loaded pol
Coalitions appear lead increased rancour political party coalition appear actually break trust That result intense competition poison entire political ecosystem   writer Image Adobestock   Monday morning City Tshwane awoke political leadership There mayor Stevens Mokgalapa resigned   resigna
Jenna Etheridge World Birds Facebook World Birds Multimedia User   picture   story World Birds Wildlife Sanctuary Monkey Park Hout   Cape Town suffered devastating blow financial recovery robber overpowered owner made taking Walter Mangold   News  Monday facility'  first good weekend   month
  dead military plane crash Algeria'  worst   disaster Algeria suffered deadliest ever   catastrophe Wednesday military plane crashed near capital killing   board mostly army personnel family member official    photographer scene   charred wreckage plane caught fire field near Boufarik airb
  police station Moeka Village north Pretoria standing empty community struggle crime GroundUp report When construction Moeka Vuma police station North West'  Moretele municipality started   community hoped crime rate area decline   police station completed   never used paint wall peeling Sc
City Cape Town'     temporary desalination plant Strandfontein Pavilion February   Photo Gallo Images Brenton Geach Despite critical need water Port Alfred surrounding town Makhanda High Court stopped Ndlambe municipality awarding contract construction operation desalination plant judge say
  Bornman Jjacob Zuma face charge Deaan Vivier File Multimedia User   picture   story Former president Jacob Zuma congregant eThekwini Community Church KwaZulu Natal still persecuted despite resigning president month Zuma attending Good Friday service church urged attendance lead right direc
      Reckoning Finally Arrived National Health Service meant time emergency unexpected disaster People legally forced    based premise   official systemically stealing   pandemic struck Health Service baby mother dying   dying People saved simple basic procedure dying Health Service used r
Democratic Alliance leader Mmusi Maimane unveiled   Alan Winde party'  premier candidate Western Cape deflecting question requested stand position City Press reported Sunday Maimane considering standing    province   governs number senior leader serve various structure party including Weste
Pompeo'  comment came Europe part United States prepared cautiously lift virus lockdown sign emerge deadly pandemic ebbing government look restart battered economy More    killed    million infected worldwide virus left half humanity form lockdown pushed global economy towards worst downtu
Standard Bank appointed Lungisa Fuzile former Director General National Treasury   Standard Bank South Africa SBSA Fuzile report   Tshabalala   Standard Bank Group appointment follows recent change group'  executive structure resulted disbanding joint group chief executive construct leadersh
There   Excuse seen horrendous escalation attack white number violence   look five vicious barbaric attack took place innocent clearly illustrate though multiple black   easily take anything wanted steal walk away rape murder unnecessarily excuse apologism primitive hyena roam amongst   prot
South Africa Social Security Agency SASSA      million applied   special Social Relief Distress   Grant within three   system going live SASSA announced statement Friday SASSA   started taking nationwide application grant Monday       today      application received   largest number
  governing party   seemingly still scrambling ensure list standard   News  understands composition list item   national working committee'    agenda   Monday    '  national list committee expected finalise list Tuesday Acting national spokesperson Dakota Legoete News    appeared    group l
File Lower crop yield farm across Ethiopia resulted declining food supply city Esubalew Abate assistant professor food nutrition security Addis Ababa University ADDIS ABABA Endale Terefe remembers time used   school hungry trouble staying awake lesson READ Ethiopia drought left     million h
Public Protector Busisiwe Mkhwebane Photo Felix Dlangamandla Netwerk  What nature Hawks' interest Public Protector Busisiwe Mkhwebane August   advocacy group Accountability   laid criminal charge perjury defeating   justice Mkhwebane group result Constitutional Court judgment July   found Mk
South Africa inaugural investment conference meant garner financial support revive economy garnered nearly    billion real investment President Cyril Ramaphosa Friday These real investment going made pledge Ramaphosa    local international delegate hugely overwhelmed interest many business
Canny Maphanga Patrick Karegeya Rwanda'  former   chief File   Multimedia User   picture   story South Africa Police Service   week submit action taken January   bring suspect book murder former Rwandan head intelligence Colonel Patrick Karegeya magistrate opinion decision hold formal inques
Police fired tear   convoy used water cannon tried stop reaching capital'  main business district NAIROBI   killed Friday Kenyan police tried disperse supporter cheering convoy carrying opposition leader Raila Odinga airport central Nairobi Reuters photographer Police fired tear   convoy use
JOHANNESBURG South Africa currency bounced back steep sell yesterday following strike    force Iraq according   Research However tone statement Iran thereafter somewhat muted noting want escalation   presidency uncharacteristically calm   prospect escalation major development local front
Taxi boss threatened halt operation Wednesday outstanding dispute City Cape Town CAPE TOWN Taxi boss Congress Democratic Taxi Associations Codeta Cape Amalgamated Taxi Association Cata made profit four year    express joint venture City Cape Town Taxi commuter Cape Town facing possible disru
Zimbabwean pastor centre protest movement President Robert Mugabe charged Thursday plotting overthrow government insulting national flag called demonstration year Evan Mawarire arrested Wednesday return Zimbabwe United States faced charge July magistrate court freed preacher police follow pr
  Tuesday governing party held annual January   Statement Ohlange Inanda These event constituted build official launch  '  manifesto Saturday Moses Mabhida Stadium    leadership decide hold important event    event spread across country simplistic answer   wanted save cost avoid logistical
Dros '  contact' family child allegedly   aped 'Nicholas Ninow' eatery restroom Nicholas Ninow           Dros   Pretoria eatery seven year   girl allegedly    restroom confirmed incident made contact family restaurant Wednesday Dros Head office hereby confirms incident involving ir
National Union Metalworkers South Africa Numsa Tuesday shocked death Thembisile Yende Eskom employee found dead locked office Springs substation Ekurhuleni Yende body discovered Monday Eskom Pieterboth substation Springs reported missing   week   alleged   year   Yende seen went workplace Sp
South Africa slow justice lead justice farm murder trial   accused start since    delaying tactic accused assisted Legal   Board seven black farm murderer farm murder case attempted farm murder case Kleinfontein area North West playing system dragging case endlessly help Legal   Board accus
JOHANNESBURG Gauteng   Social Development Nandi Mayathula Khoza Thursday pledged support family baby Singalakha Sonamazi saying provincial government Rand West District Municipality working together bury Singalakha Sonamazi   month   infant died Tuesday Bethany Trust Home Krugersdorp transfe
South Africa education system remains unequal recent matric result clearly demonstrate education system biased towards financial mean Students wrote Independent Examination Board   exam mainly private school almost    rate However    rate wrote government'  National Senior Certificate   ex
Tens thousand South Africa healthcare worker frontline global   COVID   Many unenviable position fighting pandemic risk life family SABC capture voice weekly feature COVID   Frontline Workers Digital producer Lebo Tshangela   speaks Sister Eldah Radebe   owns clinic Freedom Park Soweto staff
CAPE TOWN South Africa business participated three Africa Down Under   International Mining Conference Exhibition ended Perth Australia optimistic trade lead secured generate concrete deal Department Trade Industry   companies' participation part Outward Selling Mission Australia organised f
South Africa Police everything within power vigilant guide citizen society eradicate crime ensure safety individual Durban police diver found body   year   male drowned Sunday evening South Beach Newly Released      Highest Paid   Stars body deceased recovered earlier morning Metro Police S
Gauteng Premier David Makhura announced Thursday morning City Tshwane placed administration Makhura medium briefing Thursday office Gauteng legislature Gauteng executive council   yesterday Wednesday   development City Tshwane decided invoke Section      Constitution Makhura added meant Cit
verbal   broken    member Fikile Mbalula Carl Niehaus close confidant party secretary general   Magashule social medium Mbalula vowing crook Niehaus dealt latest episode appears larger battle Mbalula Magashule started Niehaus complained Twitter Mbalula blocked social medium application Nieh
First published GroundUp First published GroundUp First published GroundUp City Cape Town advocate Anton Katz argued order made city Marikana land case Instead court direct National Department Human Settlements compensate owner Marikana informal settlement Philippi East City Cape Town advoca
Jenni Evans Ajay Atul Gupta Duduzane Zuma Gallo Images City Press Muntu Vilakazi Multimedia User   picture   story Ajay Gupta rubbished former finance minister Mcebisi Jonas' claim offered   finance minister family'  Saxonwold mansion fact time claimed meeting someone able support alibi affi
President Donald Trump incensed surge illegal immigrant across southern border vowed Thursday impose tariff good coming Mexico starting   ratcheting higher flow cease Trump'  move dramatically ramped battle control tide immigrant swelled despite effort build border wall halt thousand crossin
Cape Town Mitchells Plain Plain   hospitalised bullet fired   grazed head Town Centre Sunday Twelve year   Bernard Lewis   bullet fired City Cape Town   Enforcement officer trying catch group robber near fish market Police   crowd bullet grazed head shortly    witness called Voice They shot
Posting Country   Full Time Part Time Full Time Contract Type Permanent Joining Vodacom   matter carry minute text data carry life   huge responsibility think minute rely likelihood rely   Customers heart everything want make difference life customer community live work support give somethin
 '  power   bring economic growth year close zero continue severity seen March Reserve Bank wave rolling blackout started November among worst country   experienced knock    percentage point economic growth Reserve Bank Monetary Policy Review released Wednesday Pretoria Expansion close zero
  died   crash Kiliarchen Road Ladysmith Newcastle   Saturday afternoon Commuters forced   alternative route order   various destination drive Kiliarchen Road Ladysmith Newcastle   Saturday afternoon four   tragically collided   resulted total   lost life    involve tragic accident Four tra
business usual many shop owner part Johannesburg Central Business District   Monday awoke lifetime investment burnt   come tension flared   Malvern part inner city Sunday continued early hour Monday morning group   wielding stick panga tree branch marched city centre disrupting meeting   fou
  Venter arrested Monday allegedly contravened condition sentence    year   father three sentenced December   four year house arrest following conviction several charge including performing sexual   minor   charge relating showing pornographic material minor   condition sentence allowed wor
late SACP leader family accused government repatriating struggle icon remains without permission North West heavyweight lawyer representing anti apartheid struggle hero John Beaver   Marks blood relative sent urgent letter President Jacob Zuma Arts Culture Minister Nathi Mthethwa cease proce
position change Justice Department argued year large part    struck case Texas     pending Fifth Circuit Court Appeals trial court judge sided Texas voided   December ruling Justice Department   entire   enacted President Barack Obama thrown filing assures renewed attention President Dona
```

Figure D.6: Extracted True-SA class prototypical examples