PL-NCC: A Novel Approach for Fake News Detection through Data Augmentation

by

Keshopan Arunthavachelvan

Bachelor of Science, Ontario Tech University, 2021

A thesis

presented to Toronto Metropolitan University

in partial fulfillment of the

requirements for the degree of

Master of Science (M.Sc.)

in the program of

Computer Science

Toronto, Ontario, Canada, 2023

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Toronto Metropolitan University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Toronto Metropolitan University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

PL-NCC: A Novel Approach for Fake News Detection through Data Augmentation

Master of Science (M.Sc.), 2023

Keshopan Arunthavachelvan

Computer Science

Toronto Metropolitan University

Abstract

Fake news has become rampant with the prominence of social media, jeopardizing information integrity and the reliability of news. It becomes imperative for media platforms to develop robust and efficient strategies to mitigate the spread of fake news. We make two contributions to fake news research in this work. Firstly, we propose the Psycho-Linguistic News Content and Comments (PL-NCC) dataset, a consolidated dataset of two leading fake news datasets, NELA-GT and Fakeddit, which leverages linguistic and psychological characteristics from the news article and user comments to improve the classification accuracy of benchmark models. Secondly, we propose the News Content and Comments (NCC) classification model to leverage the psychological features extracted from our PL-NCC dataset, which introduces a feed-forward layer to a deep learning model, enhancing the efficacy of the extracted psychological features to better identify fake news. Our approach achieves a classification accuracy of over ninety percent, exceeding baseline results.

Acknowledgements

Firstly, I would like to thank you as a reader for taking an interest in my research and taking the time to read through my work. I would like to express my greatest gratitude to my supervisor Dr. Cherie Ding for her support and guidance on my research throughout my time at Toronto Metropolitan University. I would also like to thank Dr. Shaina Raza for taking the time out of her busy schedule to collaborate and share her expertise in research academia throughout my research and previous papers. I would like to thank the Department of Computer Science and Dr. Cherie Ding for providing me with funding.

Additionally, I extend my thanks to the members of the examination committee, Dr. Alex Ferworn, Dr. Vivian Hu, and Dr. Alireza Sadeghian for taking the time to review my work and for their valuable feedback and revisions to better articulate the writing of my research.

Finally, I would like to give special thanks to my mother, my father, and my loving family for encouraging me throughout all my years in academia, allowing me to reach the point where I am today. I thank you all.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1

Introduction

1.1 Background and Problem Statement

Social media has gained immense prominence in modern-day society, changing the way news and information are shared and accessed. The convenience of disseminating information through social media has made the process of spreading news much simpler; however, the resulting consequence is the increased spread of fake information, which poses a significant threat to maintaining the integrity and trust of news [62, 69].

Fake news is the deliberate fabrication or sharing of misleading information which is presented as factual news, including fake stories and manipulated facts. The propagation of fake news on social media poses significant consequences in modern-day society, such as the manipulation of the 2016 election [3], which caused many voters to question the reliability of their presidential candidates.

As news credibility worsens, it becomes increasingly difficult for the public to distinguish between fake and truthful news, causing the general public to question the credibility and reliability of trustworthy news and information sources. To combat the growing problem of fake news, it becomes crucial for social media platforms to take preventative measures to prioritize the development of identifying and mitigating the spread

of fake news. Such measures include implementing machine learning systems to detect and flag fake or misleading news content.

Earlier fake news detection primarily relied on the news article itself for news classification by analyzing the news content, writing style, and source credibility. However, researchers have started including user engagements and propagation data (user profiles for news distribution and tracking) as part of the classification task. Previous studies show that user engagement and news propagation provide valuable information for the classification of fake news. However, other work have identified through analysis that identifying the propagation graph or news cascade can be time-consuming, and this type of information may not be available if we want to detect fake news in the early stages. In this study, we developed a simplified and efficient approach for early detection of fake news by incorporating the news content itself and accompanying users' comments data.

Additionally, recent work in fake news detection explores the use of Cognitive Natural Language Processing (NLP) [1, 2, 11], which studies elements of cognitive science, linguistics, psychology and artificial intelligence to better understand human language. Cognitive NLP provides valuable details about the writing style of an article, including grammar and language analysis, as well as emotions portrayed in the writing. This data greatly enhances the ability for classification models to determine the truthfulness of news.

Studies have observed the value of psychological characteristics when performing fake news classification, with results indicating performance improvements when including various sentiment-based characteristics extracted from the article's news content and user comments. These studies analyze emotions, swear words, and social behaviour to determine the veracity of the news. Our detailed analysis of these feature groups indicates that social behavioural traits such as politeness, interpersonal conflict, moralization, prosocial behaviour, and communication influence the accuracy of fake news classification. Fake news often exhibits emotional bias and profanity, whereas truthful articles tend

to remain more neutral. Positive and negative emotions, along with tones of anxiety, sadness, and anger, are frequently seen in fake news and can be helpful in improving fake news classification methods [17].

Existing fake news datasets such as NELA-GT [18] or Fakeddit [37] only focus on one content type or are much too small such as FakeNewsNet [51]. In addition, early detection models consider only the news article's text and fake news detection models, which consider propagation patterns, are much too complex for the classification task [33, 54]. Many classification models utilizing user propagation patterns assume that all propagation data is available at the time of publication; however, for early fake news detection, this information is often not available [33], and extracting propagation graphs/paths and analyzing the up-to-date user profiles will take time to complete, making it not ideal for early detection tasks. Developing a classification model to rely mostly on news content would be a better option for early fake news detection. In this work, we would like to address the above mentioned issues for early fake news detection by coming up with a dataset that includes both news articles and user comments on these articles, as well as a detection model that makes full use of the linguistic and psychological features extracted from the news content and user comments. The model should function well when only relying on the news content, and it should also have the flexibility to take into consideration the user comments as they become available.

1.2 Objectives

The main research objective for our work is to first develop a consolidated Psycho-Linguistic News Content and Comments (PL-NCC) dataset built on state-of-the-art NELA-GT and Fakeddit fake news datasets and expand on them by including various linguistic and psychological features unique to fake and real news. In addition, when the propagation data is available, we develop a simple way to include the propagation data

in our dataset by including the related user comments for each news article, making it a desirable approach for early detection models. Secondly, we would like to come up with an early fake news detection model using these linguistic and psychological features of both news content and user comments. We aim to improve on standard classification models by introducing a feed-forward linear layer to a standard multilayer perceptron model to better enhance the efficacy of the proposed psychological features in this work.

In this research, we try to answer the following research questions:

RQ1 How do linguistic and psychological features affect the performance of fake news classification?

RQ2 Does the consolidated dataset improve the performance of existing fake news classification?

RQ3 Can the proposed model leverage psycho-linguistic features and user comments to improve existing content-based classification?

RQ4 Which linguistic and psychological feature groups exhibit the best performance?

1.3 Proposed Approach

To build the PL-NCC dataset, we take the news content of NELA-GT combined with the user engagement from Fakeddit to create a more diverse and larger dataset for the classification task. Additionally, we expand on the consolidated dataset by introducing additional pre-processed linguistic and psychological features from the article's text and user comments to be ready to use for classification models. Psychological features are not widely studied in the field of fake news research, with only a few studies experimenting with sentimental analysis and emotions [19, 68]. We provide the embedding values of these extracted features in the PL-NCC dataset proposed in this thesis.

After preparing the dataset, we analyse the patterns of these linguistic and psychological features existing in fake and real news, and find out that leveraging these characteristics can enhance the fake news classification. We further analyze the patterns in the user comments and observe that these features sometimes exhibit different patterns in fake and real news. By including the user comments, we could further enhance the detection accuracy. Based on these findings, we propose a content-based early fake news detection model - News Content and Comments (NCC) based model, which combines deep neural networks with the lexical, semantic, syntactical, and psychological features extracted in the proposed dataset. To enhance the efficacy of the proposed psychological features in our PL-NCC dataset, we build upon a standard MLP classification model by introducing a feed-forward linear layer, allowing the classifier to better train on unique characteristics of fake news.

Finally, we conduct an extensive set of experiments to compare the performance of the proposed NCC classification model against the latest state-of-the-art fake news detection models. Using fake news datasets from NELA-GT and Fakeddit as benchmarks, we evaluate the performance of each model with various feature sets and embedding systems. Our model outperformed existing classifiers with fake news detection, achieving an overall accuracy and F1 score of ninety seven percent.

1.4 Structure

The following chapters of this thesis are organized as follows:

2. **Literature Review** provides an overview of related work in the field of fake news research and outlines the different types of fake news implementations and dataset types.

3. **Methodology** presents a detailed overview of the design of our classification model and dataset.

4. In the **Experiments and Evaluation** chapter, we review the performance of our model and dataset, as well as perform a detailed analysis of the effectiveness of the proposed features and model. We answer the proposed research questions in this chapter.

5. We conclude the thesis with the **Conclusion and Future Work** chapter where we provide our closing remarks, outline the limitations of our work and discuss future development of our model and dataset.

Chapter 2

Literature Review

The study of fake news research can be categorized into two forms of research categories: improvement of classification methods, and the development of robust fake news datasets aimed to aid in improving fake news classification. We discuss both types of research in this chapter.

2.1 Fake News Classification Methods

Fake news classification can be categorized into two general types: manual and automatic detection methods. While both methods are effective for the classification task, each has drawbacks which the other method attempts to improve. Manual detection includes the use of fact-checking websites, such as PolitiFact and Reporterslab, which contain a curated list of news articles that have been manually reviewed and labelled by humans. These websites provide the ground truth labels (real or fake labels) to identify the veracity of each news article. This detection method is very accurate; however, the act of performing the classification is a very lengthy and tedious process. As a result, researchers and social media platforms have begun leveraging automated detection methods to perform the classification task. These detection methods are performed using

machine learning systems and perform the classification task much more efficiently. A survey by Zhou et al. [73] categorizes automated detection methods into the following two types of models:

- knowledge-based: analyzing the written content of a provided text, including the writing style of the text
- propagation-based: studying how a provided text spreads to different users, and analyzes source credibility, which studies the credibility of the writer of a text

The survey suggests that developing an efficient fake news detection model should utilize a combination of these two detection methods to leverage the benefits of each approach while overcoming any limitations that each model has.

2.1.1 Knowledge Based Detection

Knowledge and Information Detection

Knowledge-based detection models capture fake news by validating the news content in a text with existing truthful information. There are different methods to determine the accuracy of facts in a news article, but they can be generalized into automatic and manual testing. Fact-checking can be conducted by either utilizing experts in a field or communities and the general public to analyze the validity of the text. Although the utilization of experts produced more accurate results, it is not as efficient as it seems to be. It can be very costly to do so and is difficult to do with the increasing demand for news that is being spread. On the other hand, crowd sourcing the detection to communities or the general public is more efficient, as more resources are available; however, the detection can be less credible as personal bias can be introduced into the detection results.

Graph learning frameworks [38, 67], such as the CompareNet model [23] focus on generating two graphical representations of the textual content in an article to perform

source-based classification. The first graph extracts entities, key topics and sentences from an article, where the entities and topics are connected to each sentence in a one-directional link. A one-directional link is used to perform semantic analysis of the entities while preventing the influence of other true entities from affecting the news representation. This graph is then utilized to create an attention network graph to learn the most linked topics within the paper. These topics are then compared with the knowledge base using an entity comparison network. This entity network performs structural, textual, and gating embedding on the previously generated graph by comparing the contents with the knowledge bases used. Finally, the entities obtained from the article are run through a fake news classifier to perform its fake news classification. The CompareNet is a valuable tool that can be implemented before running existing classification models to improve the accuracy of the model's detection and add the additional source-based classification. Source-based entity classification has a lot of value in fake news detection, and creating an easy-to-use model that can be incorporated into several different fake news detection models is very valuable.

Content and Writing Style

Many trustworthy news publications follow a standard format when writing their articles, consisting of a title and the article's text content. Many fake news pieces attempt to copy these writing styles to attract more readers into believing the content in their text. The survey by Zhou et al. [70] explains that fake news detection platforms must capture any irregularities between the writing style of a fake news text and the style of a truthful text. Standard features that these detection models analyze are lexicons (frequency of words), syntax (frequency of verbs and nouns), disclosure (frequency of rhetorical words in sentences), and semantics (frequency of multiple factors such as complexity of words, sentiment, etc.).

Work by Kai Shu [49, 53] discusses the psychology behind fake news detection. Par-

ticularly the psychological traits that humans exhibit, which lead them to believe in fake news. These two traits are Naive Realism and Confirmation Bias. Naive Realism is the ideology that humans believe only their perspective on a given topic and strongly disagree with any opposing views. They tend to claim that individuals with opposing views are uninformed. Confirmation Bias is the concept that readers tend to consume content that sides their stance on a given topic. Both traits work together to allow readers to believe fake news more easily. This topic works hand in hand with the echo chamber effect discussed in the papers, which explains that individuals within the same social circle tend to share information, in which all parties share similar views. As a result, individuals are more likely to believe in fake news that their colleagues may spread on social media platforms.

The FNDNet [29] model utilizes convolution neural network models to identify crucial feature sets which differentiate real and fake news. This model utilizes deep learning convolutional neural networks, allowing it to learn and test new features that it deems valuable. The most significant benefit of this model is its improvement to the automation of existing models by constantly learning about new patterns and features of fake news implementations. This is a large benefit compared to existing machine learning fake news detection classifications as they require manual implementation of the features that the model needs to test.

Compared to the other fake news detection models, the FNED model [33] focuses on detecting fake news early on in its propagation (within five minutes since the initial post), allowing the model to focus on preventing fake news from spreading before it can cause harm. Many existing fake news detection models focus on reducing the further spread of fake news once it has already been propagated to multiple users. There is sufficient data that can be used to classify text as fake news at this point; however, the article would have spread and been read by numerous users by then. When using the same models earlier in the propagation, these models can experience overfitting as the amount of data

is not sufficient to correctly classify these texts. The work discussed in this paper states that most text articles reach their peak propagation within the first twenty-four hours of being posted before slowing down their spread. Most existing fake news detection models begin their detection near the end of the first twenty-four hours, and by this point, the fake news has already spread to many readers. The goal behind the FNED model is to prevent the initial propagation of fake news before it can begin spreading to multiple users.

The FNED model attempts to fix a significant problem with existing fake news detection models while incorporating multiple efficient detection methods into its fake news detection. The FNED detection model uses a combination of a variation of sentiment analysis using user feedback (typically within the first fifty responses) and a CNN-based news classifier. The first component of this model uses both text content (the user response) and the user profile to detect the likelihood of an article being fake based on past and current textual responses from a given user. If multiple users frequently share positive sentiments towards fake news, the article is likely to be classified as fake news. The model also introduces the PU-Learning framework, which focuses on using epochs (weak classifiers) of the unlabeled news dataset to help train the model to detect fake news. Each epoch of the unlabeled news dataset is classified as truthful. Once all the epochs have been used for training, the results are averaged to obtain the probability of an article being fake.

The use of sentiment analysis promises very high accuracy rates as user feedback provides a lot of valuable information, which is often overlooked in fake new detection models. Likewise, using the PU-learning framework that this research team implemented makes it possible to train a robust fake news detection model using a very small sample of unlabeled articles. This allows fake news detection models to be implemented early into the news' propagation to help reduce the spread of fake news much sooner, decreasing the reach that the article can achieve.

The theory-driven model by Zhou et al. [70] discusses the theoretical approach for fake news detection, mainly focused on clickbait articles. This model uses a combination of both text content within an article and the propagation patterns of news spread to detect fake news. The implementation intends to analyze the text content of an article to detect fake news using a combination of lexical, syntactical, and semantic features.

- The lexical level focuses on analyzing the word frequency using a standardized bag of words model to study the writing style of the article.
- The syntactical level is categorized into two specific categories of classification: shallow syntactic features and deep syntactic features. The shallow syntactic feature analyzes the frequency of parts of speech, such as nouns, verbs and determiners. The deep syntactic features analyze the frequency of rewrite rules obtained using the Probability Context-Free Grammar parsing tree.
- The semantic level focuses on analyzing sentiments within the article's text content. This level detects clickbait using a combination of general clickbait patterns (such as "this can change your life" phrases), the value of the news discussed in the article, the readability of the content, and sensationalism, which determines the sentiment within the content. This language level is helpful in detecting clickbait within an article.

Like the FNED detection model, the UFD model [65] attempts to achieve an efficient fake news detection model that can run on a small training data sample. It states that many existing fake news classification models require large training data and take a longer amount of time (late in a news article's propagation) to detect fake news effectively. This model attempts to remedy this problem by introducing an unsupervised training model that only utilizes data present in the article being classified. The FNED and UFD classification models analyze user engagement as their primary classification model.

However, the UFD model also incorporates a text content classification model that uses a random sampling model with epochs to perform the fake news classification.

Unlike other models, the UFD model works around problems associated with noise introduced with user engagement in news articles. The work attempts to reduce the noise introduced into the classification model by leveraging Twitter's verification feature on user profiles. The model only performs the classification on user posts (articles) made by a Twitter-verified user and incorporates the user engagement from unverified users within the same post. This method removes multiple posts from unverified users, which may not have a lot of helpful user engagement to aid in the model's training.

The UFD classification model uses an efficient approach to extract recent user feedback on news articles. The model obtains a dataset of news articles containing titles of the news articles and automates a search process using Twitter's search API to collect tweets related to the selected news article. By doing so, the model can obtain all user engagement that occurs on a specific article shared by multiple users on the platform.

The Grover model [66] is a fake news generator which uses artificial intelligence to generate full-text content and metadata of an article using only the title. The researchers who developed this system address the ethical issues of distributing this model due to the possibility of increased spread of fake news. However, the Grover model has the additional benefit of high accuracy in detecting fake news with Grover-generated news articles. The Grover model uses artificially generated text content to aid in its fake news detection. However, the Grover model does not mimic human psychological behaviour when generating its text, allowing it to portray real human-like text accurately.

Work by Horne et al. [20] perform their classification only on the title of an article and analyze fake, real and satirical news articles. Additionally, their work has been tested on three different datasets, all of which have been personally scrubbed by the researchers. The core assumption in this paper indicates that satire and fake news share similar properties in terms of writing style compared to real news articles. Additionally,

the paper states that satire and fake news-based articles attempt to persuade the readers through the article's title. In contrast, truthful news articles use the title to engage the reader. The paper analyzes three critical features within the title of an article, consisting of the stylistic, complexity and psychological traits. The stylistic features of the article are captured using the Bird POS tagger, obtaining data such as the frequency of stop words, punctuation, etc. The complexity is calculated using the article's title at the sentence and word level. The complexity is analyzed by computing the number of each respective term and its categorization (such as verb and phrase syntax depths). The LIWC dictionaries are utilized to obtain psychological features within the title content, including factors such as personal concerns.

Horne et al.'s later work [22] introduces the concept of content drift and its effects on fake news detection systems. Their work first detects fake news using existing classification models using the NELA-GT 2018 dataset [39]. This classification analyzes how content drift currently affects fake news detection systems. It concludes that these models have a negative performance impact; however, the impact occurs very slowly. The researchers proceed to analyze how different labelling and features affect these models. The model utilizes several feature groups to aid in classification, specifically in writing style, the writing's complexity (such as lexical diversity, level of reading difficulty, and the length of the article), the overall bias that the articles have and the sentiments used in the paper, the morality of the article, the time and location of the topic of the article, as well as the analysis of the source through wiki pages. The two unique features listed are the morality and wiki feature sets. The premise behind the wiki feature group indicates that non-credible sources will typically not have a Wikipedia page associated with them.

Their work also discusses fake news implementations that focus on hindering existing news classification models. In their work, they discuss that fake news implementations are constantly evolving, and thus, models need to adapt to these implementations. The paper discusses different forms of fake news implementations that actively work to hinder

fake news detection systems. In particular, the paper discusses the topics of evasion attacks, poison attacks, and blocking attacks. Evasion attacks focus on utilizing existing content from truthful news articles and incorporating fake news within these articles or web pages. This can be done by stating their stances on topics within the article to avoid being flagged by fake news detection systems. The second tactic is poison attacks, which can be detrimental to classification models that are constantly learning. Poison attacks work by deteriorating the performance of existing fake news classifiers by injecting segments of fake news using features commonly used to train these classifiers. Similar to evasion attacks, poison attacks use articles that appear as real news articles but have embedded fake news to hinder the classifiers' performance. The final attack discussed is blocking attacks, which attempt to prevent the classifiers from obtaining fake news from the articles and allowing them to extract only positive samples. This attack is the more potent attack of the three, as the training classifiers will not be able to train their models on the fake news presented in these articles. The effects of these attacks appear in the classifiers several weeks after the initial attack. The fake news articles will be presented to the classifier that the classifier is unfamiliar with, and as a result, the classifier will see a significant reduction in performance.

2.1.2 Propagation Based Detection

Source Based Detection

Source-based detection models compare articles published by individuals or organizations to analyze the writer's credibility. Credibility is used to determine whether a text is fake or real. The general idea behind this model is that publishers/users with low credibility have a higher chance of spreading fake news than individuals with a much higher credibility level. Users who frequently post fake news will have lower credibility than individuals who frequently post truthful news.

The model introduced by Kai Shu, known as multiple-sources of weak social supervision (MWSS) [55], specializes in using weak sources of features from multiple sources such as user comments and the clean features of the main article to classify an article as fake or real. This model aims to implement a fake news classification model that can work effectively with a limited amount of clean annotated features within a dataset. Implementing this model makes it possible to create an early form of fake news detection.

This article discusses problems associated with fake news classification models that utilize user engagement. One of the most considerable problems with using user engagement for fake news detection is that user engagement can also be fabricated, just like fake news articles. It is difficult to determine if user engagement is authentic, and incorrectly classifying these engagements can disrupt the accuracy of the overall classification model.

The MWSS model performs its classification of fake news by adding weights to the weak social features of a news article. These weights are determined based on sentiment, stance, and the user's credibility in posting their engagement on the article. Each of these three attributes has a weight associated with them. These three features are used to determine if the news article leans towards being more truthful or fake. User engagement is conducted over multiple sources of users.

Propagation Detection Models

Research [48] has identified that propagation data proves a valuable role in fake news classification, demonstrating that individuals with a higher risk of spreading or engaging with fake news are more likely to disseminate fake news more frequently. Propagation detection models analyze the spread of news across a platform to determine if the news is fake or truthful. These analysis methods are typically represented in the form of a tree. Each node of the tree represents a set amount of information regarding a user. The edges represent the relationship between two users (i.e., sharing an article between the users). There are two fundamental forms of analysis in regard to propagation: hop-based cascade

and time-based cascade. Hop-based cascade detection analyzes only the relationship between each individual who spreads a specific piece of text. This cascade method analyzes the number of individuals who view a text article shared by a particular user. Time-based cascade analyzes the time intervals between the shares between each user. Propagation-based detection is the most reliable method; however, it is not beneficial for early detection as this model can only detect fake news once it has begun spreading.

Kai Shu et al. discuss their work on the difference between macro and micro-level propagation patterns of fake news [52]. Macro-level propagation patterns involve spreading news from its source by sharing the article with multiple users, including sharing an article through social media posts or retweeting. Micro-level propagation patterns involve user engagement and discussion on a source or shared news article. Micro-level propagation primarily focuses on applying stance and sentiment analysis to the fake news classification models. In their work, they analyze these two propagation patterns by using an article's structure and temporal characteristics to determine the benefits and drawbacks of both levels in fake news detection.

The TriFN model [54] focuses on fake news detection that primarily compares the relationships between user-news engagement and publisher-news engagement. These two engagements are compared to determine if a news article can be deemed fake. One key factor used to determine the authenticity of news is analyzing partisan bias. This bias is used to compare the likelihood of a user/publisher spreading fake news. Individuals with high biases on either left or right sides of the spectrum are more likely to skew the truthfulness of an article since they are more likely to add preference to their own viewpoints. In contrast, individuals with a smaller partisan bias are more likely to avoid spreading fake news because they are more likely to avoid allowing their personal views to impact their statements/articles. User relationships are also a key focus of this model, which analyzes the relationships between users to determine the likelihood of the users spreading fake news. The article explains that friends on social media platforms are more

likely to share views. By analyzing the relationships between users, it is possible to group individuals who are more likely to spread fake news and users that spread truthful news.

Many existing fake news detection models classify an article as fake or real; however, a user typically does not have an extensive explanation of why the article is considered fake. The dEFEND model [50] addresses this issue by leveraging user comments to explain why specific sentences classify an article as fake. The system works by encoding news articles and user comments into sentence representations. Then, these two components are linked to one another to explain why specific sentences cause an article to be deemed fake.

The FANG model [38] implementation analyzes the relationships between different users on a social media platform to study how fake news is distributed by analyzing user and company/publisher profiles, as well as propagation paths. The FANG model analyzes the user distribution patterns and interactions within a given news article, but it also incorporates data about the publisher and users, as well as the stance that the author takes. For instance, this model can look into the "About Us" section of a web page to determine if the publisher mentions anything suspicious or biased, such as the publisher actively stating that their posts are their own thoughts and opinions about events. By analyzing this data, fake news detection models can categorize certain publishers as more likely to spread fake news than publishers who do not mention any bias and only state facts.

Zhou and Zafarani's work [72] analyzes the propagation networks of fake and real news articles to explain the dissemination of fake news, mainly through the node, ego, triad and community networks. These networks are then used as features within the model to aid in its fake news detection.

The work discusses in great detail how propagation networks function within a news article. The paper analyzes fake and real news networks separately to identify key statistics. It concludes that fake news networks are more likely to exhibit much larger network trees and will experience a slowdown in their propagation much later than true news

networks. A denser network pattern is introduced in this work, which is a new form of fake news implementation involving the formation of internal networks by fake news distributors. Fake news spreaders specifically create these networks to improve the spread of their fake news. Similar to other fake news detection models focusing on news propagation, this model assigns a label to user profiles to identify their likelihood of believing or spreading fake news. Users who share similar weights for spreading fake news are likely to belong to the same dense network. By using these dense networks, it will be possible to flag groups of users for the spread of fake news.

The user-characteristic enhanced model (UCEM) model [26] focuses on implementing fake news detection utilizing user networks and propagation patterns. Their work groups fake news spreaders together based on their historical data with news dissemination. However, the UCEM model counters this stance by stating that fake news spreaders typically do not have any connection with one another and will have very few followers during initial propagation. As a result, the UCEM model attempts to remedy this problem by artificially creating networks between fake news spreaders using their profile data. They define this implementation as social proximity using network embedding. This connection is made based on the premise that two nodes are similar if there is a connection between them. Using this concept, the UCEM model generates a relationship between two similar users who are likely to spread fake news to make them closer within an article's news network, calculated using cosine similarity.

Many existing classifiers that utilize propagation trees strictly focus their fake news detection after an article has completed its propagation. This form of classification serves more as an analysis of the data rather than a plausible classifier for fake news detection. The STS-NN model [24] aims to treat temporal and spatial information of message propagation as one system, which is done by capturing both spatial and temporal information in the classifier separately and then integrating them together to obtain a final classification. Each component of this model iterates through every message in the

article's propagation. The first component of the model is the spatial capturer, which collects the spatial information of the parent message, which is then used to capture all spatial information up to the current message. Likewise, the temporal capturer captures the temporal information of the parent message using a gated recurrent unit. This unit obtains the temporal information of the total propagation. The final stage of the STS-NN model is the integrator which uses both hidden representations of the spatial and temporal capturers to form a whole representation using a two-layer perceptron. This final representation is the spatial-temporal component of the message. The unique aspect of this model is its ability to perform early fake news detection because the classifications happen at each layer in the propagation (at each message). Therefore, this model can be used for both early fake news detection and fake news analysis.

Cheng et al. [9] address their analysis of unbiased recommendation patterns and introduce the term confounders, which are variables that correlate user profile features with the user's susceptibility to fake news. The paper states that users are likely to click on an article they are interested in and the amount of exposure they have received to the article's topic. However, this results in particular articles seeing fewer engagements than the more appealing alternatives. Many classification models only analyze positive interactions with articles but miss negative (non-clicked articles) interactions with articles. The work attempts to address this problem by creating an unbiased classification system. The paper uses Inverse Propensity Scoring (IPS) to train the model to include negative and positively engaged articles. Doing so allows the model to perform unbiased fake news detection regarding user profiles. Using the unbiased dataset, the model analyzes attributes of user profiles that condition positive fake news engagement and the sharing behaviours these attributes have with fake and real news. We see similar work with the use of recommendation systems for fake news detection, such as work by Balcar et al. [4].

Kai Shu et al. have explored the use of emotions in fake news detection. Their

work [19, 68] addresses an emotion-based fake news detection model known as EFN, which discusses the relationship of emotions between publisher and user emotions. The relationship between these two emotions is classified as dual emotion. Emotions can have a lot of value in stance and sentiment analysis in fake news classification; however, this feature is severely underutilized in fake news analysis. The primary analysis between these emotions comes in the forms of emotional resonance and emotional dissonance. Emotion resonance is when the publisher and audience share a common emotion in the article, while different emotions between the two parties are classified as emotion dissonance. The core focus of the EFN model is to identify how to capture these forms of emotions and how these emotions can be exploited in fake news detection. The EFN model performs its classification by training its model on five specific emotion categories (anger, sadness, doubt, happiness and none). The model performs sentiment analysis on the textual and user responses by associating particular terms with a specific emotion category. The model can perform fake news classification by comparing the results by analyzing the resonance and dissonance of emotions between the two parties. Kai Shu et al. state in their work that many fake news share an emotional resonance of anger or other negative emotions.

Among the different fields of fake news research discussed, there have been promising studies conducted that combine different detection methods. Work by Raza et al. [46] has identified that a combination of content-based and social contexts has proven useful in the field, allowing detection models to predict fake news more accurately and early on in a news article's life cycle.

2.2 Fake News Datasets

Another field of fake news research involves the development of fake news-based datasets. We categorize fake news datasets into two types: content-based, and propagation-

based datasets.

2.2.1 Content-Based Datasets

The most common type of fake news dataset are content-based datasets [46, 21], including state-of-the-art datasets such as NELA-GT, Credbank [36], FakeNewsNet [51], and Liar [63]. Content-based datasets such as these are highly diverse in their news content due to their large sample of news articles from multiple different news sources. Datasets such as FakeNewsNet [5] obtain their labels from sources such as PolitiFact and GossipCop, which perform manual labelling of their dataset, unlike the previously mentioned datasets which are automatically labelled using machine learning. However, the process of manually labelling these datasets is a tedious task and results in a much smaller dataset size. Recent work [20, 47, 70] have explored including pre-processed feature sets as part of their dataset to enhance the classification accuracy of the detection models, such as the Welfake dataset [59], which incorporates linguistic patterns such as readability, subjectivity and sentence length as well as sentiment analysis. Recent work [64, 71] in fake news classification has explored incorporating multi-modal aspects of the news articles, such as images, to perform the classification task; however, this form of classification is out of scope for our research work.

2.2.2 Propagation-Based Datasets

Recent studies [24, 44, 37] explore the use of propagational data to aid in the classification task. Datasets such as Truth Seeker [10], PHEME [31], and RumourEval [12] are specifically created for rumour detection, which involves analyzing how news travels by using user engagements and propagational paths from Twitter's tweets.

2.2.3 Multi-Class Labeling

Recent fake news research [20] has explored beyond the conventional binary classification for fake news detection by identifying the type of fake news presented in each article. Datasets such as Liar, BuzzFeed, FEVER [58], Fakeddit, and Credbank developed their datasets to include three-, four-, or six-way labelling, which accurately categorizes fake news into types like satire, manipulation, and false image connection, which has proven to yield significant benefits in psychological-based fake news classification.

2.2.4 Multi-Category Datasets

Lastly, datasets such as Fakeddit, Credbank, and FakeNewsNet have become state-of-the-art due to their inclusion of multiple categories of data by offering a wide array of propagation-related features and including both the original headline and text content of articles. Specifically, Fakeddit provides user engagement details through user comments and IDs, which can be expanded to include user profiles. Additionally, Fakeddit incorporates images from all collected Reddit posts to enhance multi-modal classification. Thirdly, the user comment data from Reddit posts are extracted to aid text-based classification.

For our work, we improve on existing state-of-the-art fake news classification by developing a consolidated fake news dataset, which includes user comment data extracted from the Fakeddit dataset, as well as the news content and labelling inherited from NELA-GT-2019 as components for our PL-NCC dataset. The scope of this work focuses solely on the development of a content-based fake news detection with the inclusion of propagation-related elements such as user comments. The PL-NCC dataset contains not only user comments but also the news article itself, which is not commonly seen in existing fake news datasets. To make the classification task simple and efficient for early fake news classification, we exclude the propagation cascade from our dataset and

only explore the text content of user comments. To better enhance the efficacy of our dataset, we further expand on the consolidated dataset by incorporating a diverse set of psychological features obtained using cognitive Natural Language Processing (NLP) techniques, which goes beyond conventional sentiment analysis and emotional assessment used in previous research, as well as commonly used linguistic features which have been identified to enhance the performance accuracy of fake news detection. In addition to the development of the PL-NCC dataset, we introduce a novel approach to better enhance the efficacy of psychological features extracted in our PL-NCC dataset by introducing a feed-forward linear layer to a multilayer perceptron (MLP) deep learning model, which allows the model to emphasize certain psychological characteristics unique to fake news which allows the NCC model to better classify the veracity of fake news.

Chapter 3

Methodology

3.1 Problem Definition

Our text-based Psycho-Linguistic News Content and Comments (PL-NCC) fake news dataset is designed to aid in identifying fake news by analyzing the text within news articles and user comments. We further leverage this dataset to develop a News Content and Comments (NCC) based classification model to detect fake news. The goal of this work is to determine the veracity of news articles in the input news corpus using a binary classification model.

3.2 Data Description

Our first contribution in this work is the development of a consolidated dataset, which merges the news content and user comments from two state-of-the-art fake news datasets: NELA-GT-2019 and Fakeddit. For our work, we merge the text-based news content, headlines and labels from the NELA-GT-2019 dataset [18] with the related user comments from Fakeddit [37]. We use the consolidated dataset to extract additional linguistic and psychological features from the article's text and user comments to include

as part of the PL-NCC dataset proposed in this work. A breakdown of the PL-NCC dataset is presented in table 3.1. Examples of the extracted news stories and related user comments are presented in the Appendix.

Table 3.1: Data Summary

Total Articles	2,929
Total User Comments	41,867
Collection Year	2019
Fake News Percentage	69.9%
Real News Percentage	30.1%
Data Components	Articles' headline, text, comments, classification label, and metadata
Data Sources	NELA-GT (for articles' text and headlines), Fakeddit (for comment IDs)
Extra Data Retrieval	Reddit's API (for full user comments, date-time stamps, and upvotes)
Comment Data Window	First eight hours of user comment data after publication on Reddit
Dataset Combination	All the information is concatenated into the combined dataset

3.2.1 Data Pre-Processing

NELA-GT-2019 consists of new articles collected in 2019 from various news sources, containing a collection of real and fake news articles. The dataset contains the article's text content and headlines, metadata about the source of the news, as well as publication time and labelling. The Fakeddit dataset is a compilation of user comments collected from several pre-chosen subreddits over a ten-year period. The dataset extracts user comments, the headline of the user post (which is commonly the news article title), related metadata for the news publication and labelling. To prepare the PL-NCC dataset, we extract all news articles from 2019 where the news headline from both the NELA-GT and Fakeddit datasets match. After merging the two datasets, we extract only the article's

headline, text content, user comments, news publication date and time, as well as the binary classification label from NELA-GT to comprise the PL-NCC dataset.

Recent work in fake news studies [19, 54, 73] has indicated that the analysis of linguistic and psychological characteristics of the news article has significant benefits in the classification tasks of fake news. Thus, the proposed PL-NCC dataset aims to improve on state-of-the-art classification models by including a predefined set of linguistic and psychological features as part of the dataset.

3.2.2 Labeling

To label each news article in the dataset, we inherit the binary classification labels from NELA-GT. While Fakeddit offers labels for two-, three-, and six-way categorization, these labels are assigned based on the overall credibility score of each subreddit rather than a per-article or per-comment basis. As a result, these labels are broader compared to NELA-GT's approach, which labels its news articles based on the credibility of the news source. Thus, we adopt NELA-GT's labelling method for the PL-NCC dataset. Our final dataset contains zero to many user comments for each news article. Each article is assigned one binary veracity label to evaluate the performance of the classification model. A recent study by Raza and Ding [46] suggests that this approach of merging datasets and labelling is effective for the classification of fake news.

3.3 Model Architecture

We express the classification outcome as a binary value, $Y = (0, 1,)$, where zero indicates a truthful news article, and one signifies fake news. We define each article in N number of news articles as a set $P=(P_h, P_c, P_c)$, where P_h represents the headline of an article, P_c represents the article's textual content, and P_u represents the related user comments.

To improve the classification task of existing fake news detection models, we developed the PL-NCC dataset, which includes the extracted linguistic and psychological features from input P. To obtain the linguistic features, we utilize natural language processing (NLP) to obtain the numerical embedding from the article's text content and user comments. We obtain the psychological features by feeding the article text and comments into the Linguistic Inquiry and Word Count (LIWC) dictionary [42] to obtain the respective feature scores. We test our dataset using our proposed NCC classification model, where the numeric embedding and feature scores are processed through a multilayer perception model (MLP). To improve the classification task of the MLP model, we introduce a feed-forward linear layer prior to model training to improve the efficacy of the extracted psychological features. After feature concatenation, our NCC model is trained using the concatenated features as an input, and after the classification task, the output of the model is the final binary prediction of our model Y, which identifies the veracity of the news articles in N.

3.4 PL-NCC Dataset

Our focus in this work is to enhance the effectiveness of textual features extracted from the NELA-GT and Fakeddit datasets for the purpose of better classifying fake news. To achieve this task, we utilize a range of natural language processing (NLP) techniques, including analyzing parts of speech (POS) and context-free grammar (CFG), utilizing disinformation-related attributes (DIA), clickbait-related attributes (CBA), and leveraging emerging language models such as BERT embedding. Furthermore, we make use of the Linguistic Inquiry and Word Count (LIWC) dictionary to extract the psychological features from the input data in P. This approach allows our dataset to improve the accuracy of existing fake news classification models significantly. A detailed breakdown of the PL-NCC dataset is presented in Figure 3.1.

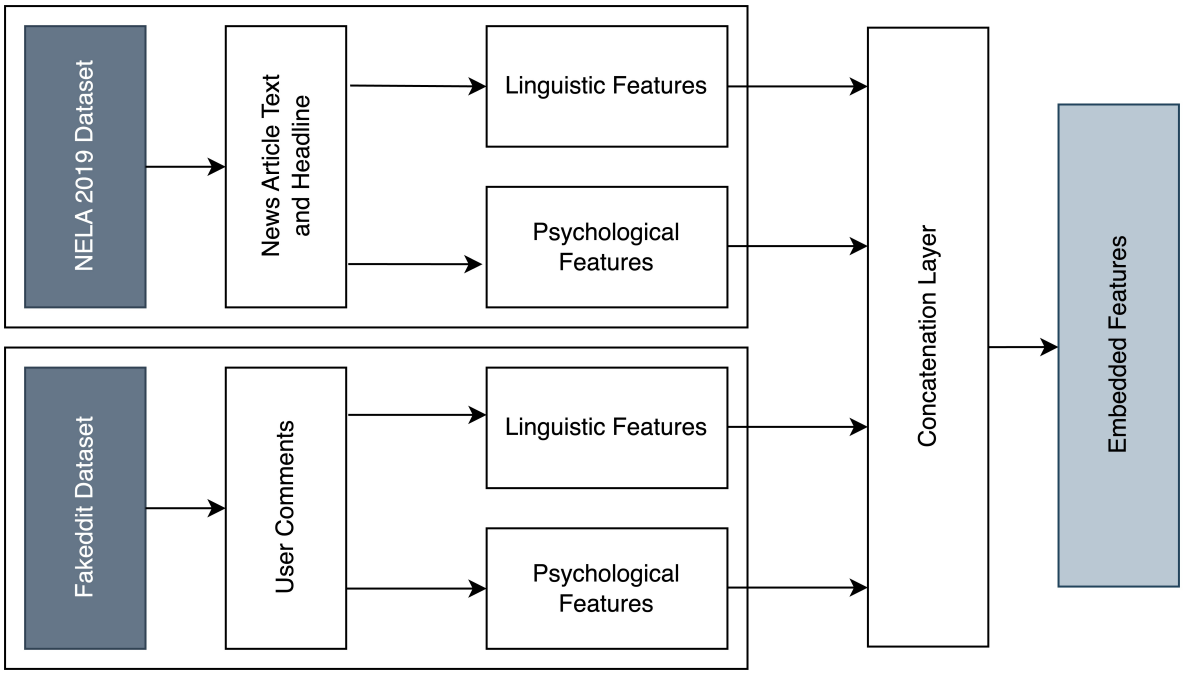


Figure 3.1: Overview of proposed dataset

3.5 Feature extraction

3.5.1 Linguistic Features

Linguistics is a core focus of fake-news research, as the earliest models analyze only the text content of the news article to perform its classification, which remains a core aspect of the modern classification task. The proposed PL-NCC dataset extracts a set of linguistic features from the news article's text content, headline and related user comments, which are either frequently used in modern day classification models or indicate valuable information to aid in the classification task. To improve the simplicity of classification models, we extract and store different sets of linguistic features from the articles' text and user comments in the PL-NCC dataset. The extracted linguistic features are beneficial for the classification task as they provide critical information about the news article's writing style. Existing text-based classification models [66] extract parts of speech

(POS) and context-free grammar (CFG) from the article's text to enhance their model's performance. As a result, these features are included as part of the PL-NCC dataset for easy extraction by classification models. Newer classification models [13, 32, 56, 60] have experimented with the use of pre-trained language models such as BERT to enhance their classification task. While the BERT language model is becoming more prominent in fake news research, we have found better success with the use of the DistilBERT language model in our work to represent the BERT embedding models, as illustrated in Tables 4.4 and 4.5. DistilBERT is a more efficient version of the existing BERT language model, which retains over ninety percent of the BERT language model, while including fourty percent less parameters and runs sixty percent faster[1]. As our goal for this work is to develop an efficient, yet effective dataset for fake news classification, we leverage the benefits of the DistilBERT language model to represent BERT embeddings in this work. We include the extracted BERT embeddings from the article's text and user comments as part of the consolidated dataset. Other classification models, such as the work by Zhou et al. [57, 73] have explored the effects of including clickbait and disinformation-related attributes from the article's text in fake news classification. Our analysis of these patterns presented in Table 4.3 indicates that disinformation traits such as discrepancy (terms which provide reasoning), causation (explanations), tentative (terms defining potential or conditions), insight (individual thoughts and knowledge), certitude (terms defining certainty), and differentiation (terms comparing variance) as well as click-bait attributes such as "this will blow your mind" and "can change your life" provide contrasting differences in fake and real news, which is beneficial in fake news classification. The labels for each clickbait attribute is obtained by comparing the headline of each news article with our dictionary of forty-seven commonly used clickbait headlines used in fake news, where the existence of a clickbait headline receives a score of one, indicating the news article is fake; otherwise, it will receive a score of zero. The higher scores of each feature indicate

[1]https://huggingface.co/docs/transformers/model$_d$*oc*/*distilbert*

that the respective news article is more likely to be fake.

3.5.2 Psychological Features

A key contribution of our work involves the inclusion of psychological features extracted from both the news article's text and user comments. Recent work in fake news classification has explored the user of cognitive NLP to aid in the classification of fake news, and research has shown that traits such as emotions, word toxicity, and social behavioural traits exhibit positive feedback when identifying the veracity of news. For our work, we extract social behavioural traits such as morality, politeness, communication (addressing a topic or subject's stance), interpersonal conflict (conflict between two subjects) [15], and pro-social behaviour (voluntary act to help others) [6], which, as shown in Table 4.3, indicate varying patterns in fake and real news. Additionally, we extract the negative and positive emotions within the article's text and user comments to assist with the classification task. With negative emotional features, tones such as anxiety, anger, and sadness can be extracted from the text and used for classifying fake news. Additionally, the use of swear words is counted to determine the word toxicity of the news article.

Morality is defined as the righteousness of actions taken by a subject and weighs the outcome on a scale [14]. All news articles have a scale of morality in which they are written, and this information is vital when determining the veracity of fake news. Our studies indicate that the writing of real news articles will remain neutrally just, while fake news tends to lean towards either side of the scale. Similarly, social behavioural traits such as politeness, pro-social behaviour, and communication indicate tones of gratitude in the writing, while interpersonal conflict signifies hostility towards the article's subject [7]. These traits are useful in fake news studies, as common words in each feature group can help differentiate real from fake news.

Real news is commonly written in a neutrally biased, professional manner, so swear

words are uncommon in its writing. In contrast, fake news articles are commonly written in unprofessional and vulgar language [43]. The inclusion of word toxicity in our dataset allows classification models to better analyze and predict the veracity of news articles based on its writing tone.

More recent work by Kai Shu [19] has explored the use of sentiment analysis and emotions in the writing of news articles. The analysis can be used to extract tones of emotion such as anger, anxiety and sadness. When analyzing emotions, fake news articles tend to display either positive or negative emotional bias, whereas real news articles maintain a neutral emotional tone in their writing. Emotions play a crucial role in fake news detection, as they can also aid in extracting toxicity and inflammatory language from the input text[2].

3.5.3 LIWC

Each psychological feature group listed above is included in our PL-NCC dataset. To obtain the numeric embedding of each feature, we obtain the occurrences of each term in the input text P and cross-reference them with the Linguistic Inquiry and Word Count (LIWC) dictionary [7] using term frequency–inverse document frequency (TF-IDF) weights. The resulting value is the percentage of each term within the LIWC dictionary in relation to the entire news corpus N.

Classification models are unable to process the raw text inputs of each feature as an input without a pre-processing step to convert the data into embedding values. Thus, to allow the models to easily input our dataset features into their model, we obtain the numerical embedding values of each feature using natural language processing (NLP) and TF-IDF techniques. All linguistic and psychological features in our dataset, excluding BERT embedding, are represented as a percentage score out of one hundred.

[2]https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en$_US$

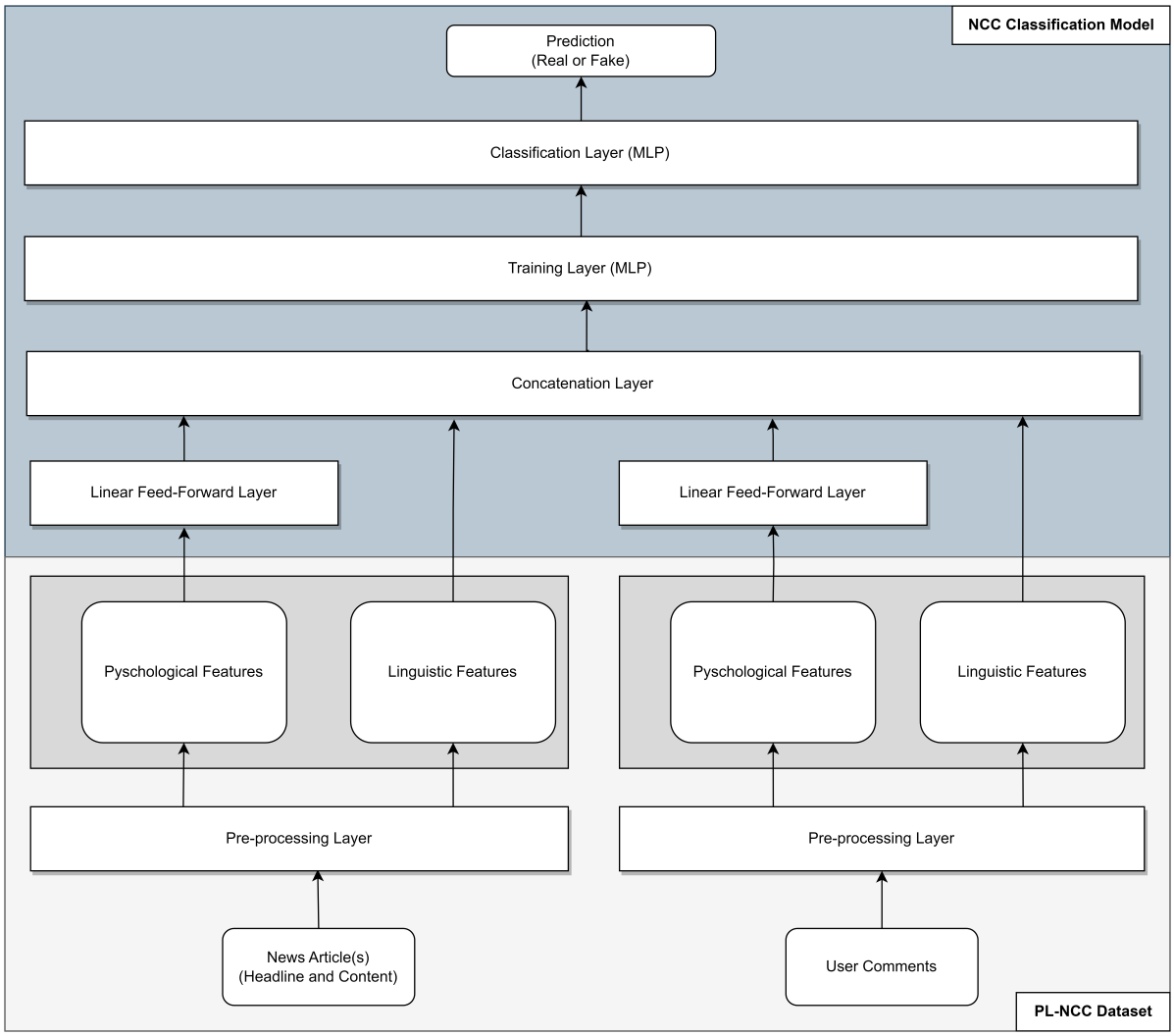


Figure 3.2: Architecture of Proposed NCC Classification Model using PL-NCC Dataset

3.6 NCC Classification Model

Figure 3.2 showcases the architecture of the proposed NCC classification model working in conjunction with the PL-NCC dataset.

3.6.1 Linear feed-forward layer

This work aims to explore the benefits of using psychological features with the classification of fake news. Before performing the classification task, we run the extracted psychological features from the PL-NCC dataset through a feed-forward linear layer. The inclusion of this layer emphasizes characteristics extracted from the psychological features which are unique to fake news articles. By emphasizing these traits, our classification model can better identify the veracity of fake news articles during the classification task. The output of this layer is an updated numeric embedding of the psychological features with emphasized fake news values.

3.6.2 Multilayer perceptron (MLP)

For the classification task, we utilize a standard multilayer perceptron (MLP), acting as a deep neural network to train and classify both the linguistic and psychological features extracted from the news content and user comments, as defined in Figure 3.3. We fit our neural model with one hundred layers, consisting of one input layer, one output layer, and ninety-eight hidden layers. After training and classification, our neural network returns the binary prediction Y of the input dataset N. Our model is optimized using the Adam optimizer [35] and utilizes the rectified linear unit (ReLU) activation function defined below, where x defines the value of the input.

$$f(x) = max(0, x) \tag{3.1}$$

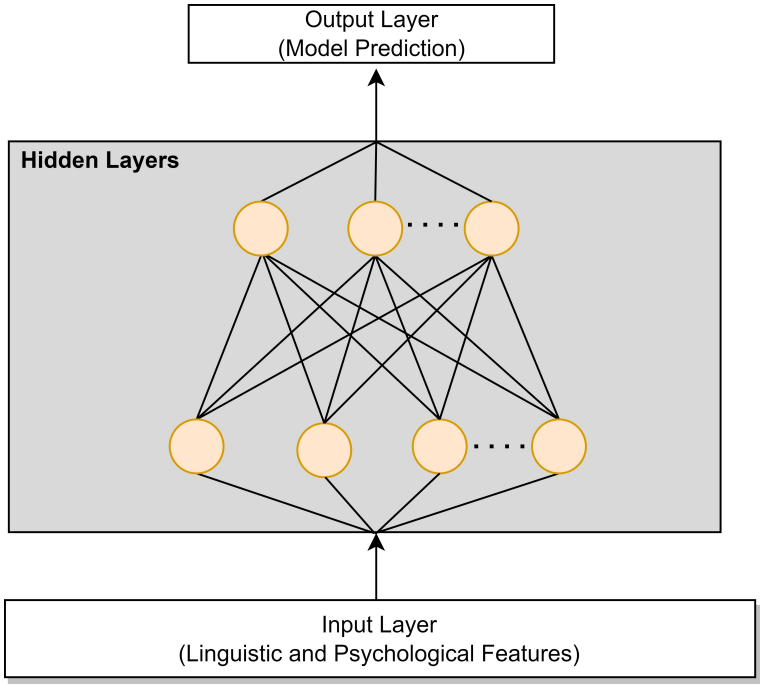


Figure 3.3: Overview of Deep Neural Network (MLP)

When training our model, we define the following sparse categorical cross-entropy loss function to evaluate the performance of our model's training:

$$L = -(ylog(p) + (1 - y)log(1 - p)) \tag{3.2}$$

where p represents the predicted value of our model and y represents the actual value.

3.6.3 Model training

The NCC model is trained on seventy percent of the input dataset N and the remaining thirty percent is used to test the model. Table 3.2 illustrates the data model for the input and output of each layer in the proposed model.

Table 3.2: Datatype for layers in proposed model

		Datatype	
Layer	Feature	input	output
dataset input	headline	string	string
dataset input	text content	string	string
dataset input	user comments	string	string
pre-processing	linguistic features	string	float
pre-processing	psychological features	string	float
concatenation layer	linguistic and psychological features	float	float
linear feed-forward	processed psychological features	float	float
classification (MLP)	concatenated features	float	integer (0, 1)
prediction	classification prediction	integer (0, 1)	-

Chapter 4

Experiments and Evaluation

In this work, we conduct a series of experiments to showcase the effectiveness of linguistic and psychological features in fake news classification, as well as the efficacy of our proposed NCC classification model. In this chapter, we address the following discussion questions:

E1 Does the proposed PL-NCC dataset improve the performance of state-of-the-art classification models?

E2 What patterns can be visualized when analyzing the linguistic and psychological features of fake news?

E3 Which linguistic and psychological feature groups have the largest impact in fake news classification?

E4 What effect does the inclusion of user comments have in fake news classification?

4.1 Experimental setup

We conduct our experiments using Keras [30], NLTK [34], and Scikit-Learn [41] libraries to process and classify the data. All experiments are conducted using a computer

with sixty-four gigabytes of RAM, a sixteen-core thirty-two-thread processor, and an RTX 2070 graphics card. When classifying the input N, thirty percent of the dataset is used for testing purposes, while the remaining seventy percent of the dataset is used to train the NCC classification model.

We test the effectiveness of the proposed linguistic and psychological features and user comments in our PL-NCC by running the dataset against state-of-the-art classification models, including XGBoost [8], convolutional neural networks (CNN) [40], multilayer perceptron models (MLP) [45], and decision tree classifiers (DTC). Recent work in the field of fake news research [28, 56] have adapted BERT embedding models to improve the performance of the fake news classification task. In our analysis, we experiment with different language models, including BERT embedding and Bag of Words models, to showcase the benefits of BERT embedding models in fake news classification.

4.1.1 Evaluation protocol

Our goal in these experiments is to understand the patterns of linguistic and psychological features in fake and real news. We conduct a series of experiments using varying feature sets and compare the results against state-of-the-art classification models.

- Our first set of experiments test our dataset's effectiveness by comparing the results of its classification task against state-of-the-art datasets. By performing baseline comparisons, we can obtain a better understanding of how psychological features and user comments affect fake news detection.

- Next, we test how various linguistics and psychological feature groups impact the performance of the classification task. With these experiments, we remove specific feature groups from the testing feature set to observe and report the results.

- Thirdly, we examine how adding a feed-forward linear layer to the classification

task impacts the model's performance when used with psychological features. This test determines whether the linear layer aids in our fake news classification task.

- Fourthly, we compare the performance using BERT-based embedding against bag-of-words (BOW) based text representation in our NCC classification model. This study helps determine the most effective text representation for our model's performance.

- Finally, we modify various hyperparameters of our model and identify parameters which significantly affect the NCC model's performance.

4.1.2 Evaluation metrics

The classification tasks addressed in this work are treated as a binary classification problem, where the resulting output P of the classification model is either an output of zero or one, indicating real or fake news, respectively. To measure the performance of both the NCC classification model and the PL-NCC dataset, we rely on standardized metrics [54, 61] such as accuracy and F1 Score as evaluation metrics. The confusion matrix in Table 4.1 provides details about the actual and predicted classifications used for our performance analysis.

Table 4.1: Confusion matrix

	Actual Fake	Actual Real
Predicted Fake	TP	FP
Predicted Real	FN	TN

We refer to true positive (TP) as the predicted fake news articles that are actually fake. False positive (FP) indicates predicted fake news articles which are actually true. False negative (FN) refers to predicted real news articles which are actually fake. True negatives (TN) indicate that predicted real news samples are real. For the F1 score (f)

and accuracy, we perform the specific calculations as follows:

$$f = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{4.1}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2}$$

4.1.3 Hyperparameters

When performing classification using the NCC model, the model is trained using one hundred epochs to minimize the training loss, and the model is optimized using the Adam optimizer. The linguistic features extracted in the PL-NCC dataset are embedded with a dimensionality size of 768 using the Scikit learn library, and the extracted psychological scores from LIWC are used as is during the classification task. Finally, the NCC model leverages a ReLU activation layer, a softmax output layer and utilizes the sparse categorical cross-entropy loss function. Table 4.2 shows a breakdown of the hyperparameters used in the NCC classification model.

4.2 Patterns of Linguistic and Psychological Features

Firstly, we determine how effective linguistic and psychological features are in the classification of fake news. We conduct a series of experiments to analyze patterns from the writing of real and fake news. The results of our analysis are presented in Table 4.3, which illustrates the score distribution of each linguistic and psychological feature obtained from the news corpus. Along with these results, examples of the score distribution between real and fake news as well as news articles and user comments for each feature are illustrated in Figures 4.1 to 4.4. For each figure, the Y-axis on the violin plot represents the feature scores obtained from LIWC for all articles in the testing corpus,

Table 4.2: Hyperparameters used for proposed NCC model

Hyperparameter	Optimal value
Dimensionality size	50
Feed-forward layer dimensionality	64
Feed-forward layer dimensionality	16
Activation Function	relu
Number of labels	2
Batch size	25
Epochs	100
Learning rate	0.001
L2 Regularization	0.0001
Dropout	0.1
Optimizer	Adam
Loss function	sparse categorical cross-entropy
Output layer	softmax

while the X-axis indicates whether the news is categorized as Fake or Real. Within each violin plot is a box plot indicated in black which identifies the median, upper and lower percentiles of the feature scores for each feature. Fake news is represented by the blue section, while real news is represented in orange.

We study the results of each feature in fake and real news and report these results. When analyzing patterns of Discrepancy in both the article's text and user comments for fake and real news, as illustrated in Figure 4.1, it is evident that the feature score obtained from LIWC is commonly higher in real news compared to fake news, as illustrated by the larger peak in the real news violin plot compared to the fake news plot. Similarly, there are more fake news articles with feature scores around one percent as indicated by the larger width of the violin plot at the one percent y-axis, with a smaller width in the related real news plot. When analyzing the feature scores obtained for related user comments with the Discrepancy feature, few fake news articles obtain a feature score of fourty percent, as indicated by the peak of the center black line of the fake news plot, which extends to the fourty percent mark on the y-axis, compared to the peak of twenty-five percent

on the real news plot. This analysis indicates that in fake news, user comments tend to have higher scores for disinformation-related attributes (DIA) compared to real news, such as discrepancy, as shown in Figure 4.1. However, patterns in the article's text differ for these features when analyzing both fake and real news. This signifies the importance of incorporating user comments alongside news content, as these unique patterns offer insights to improve the classification accuracy of fake news detection. Differentiation between the text of an article and its headline is evenly distributed between real and fake news, while certitude is more present in real news than in fake news, as shown in Figure 4.2.

Recent studies [19, 68] show cognitive NLP has promising outcomes in leveraging emotions from the article's text for fake news classification. Their research analyze emotional patterns in fake news detection and identify that fake news shows more emotional bias compared to real news. Our analysis confirms these findings, as indicated by Figure 4.3, where there is a greater prevalence of emotions in fake news, as shown by the higher scores compared to real news. Although real news articles also tend to exhibit higher negative emotional scores, the differences are less pronounced when compared to fake news.

Our study shows that fake news articles tend to exhibit more clickbait titles than real news articles. This is evident from the higher average feature scores among fake news articles, indicating their tendency to receive higher scores compared to real news articles.

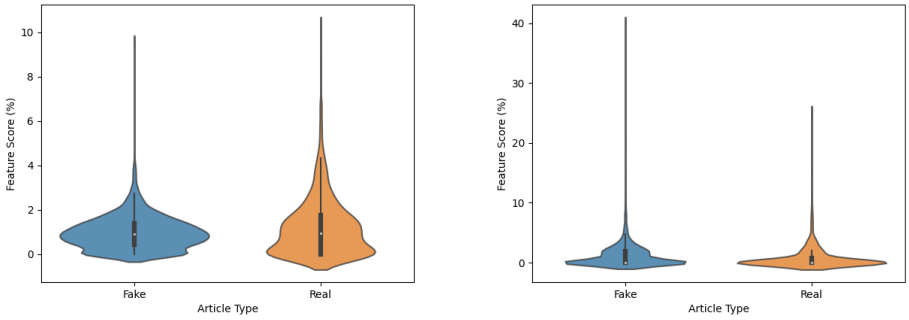
Moreover, user comments associated with fake news show increased emotional bias, whereas real news articles have a higher tendency to show negative emotions. In our analysis, we identify that real news articles tend to contain more toxic language, including swear words, when compared to fake news. When we inspect the real news articles with increased toxicity scores, we discover that specific news sources, such as Onion, are considered reputable true news sources; however, their writing style often includes offensive language in headlines, such as "depressed monkey throwing sh*t at himself".

Table 4.3: Feature Comparison between Fake and Real News (Higher Average Score means more prominent)

	Article				**Comments**			
	Fake News		**Real News**		**Fake News**		**Real News**	
Feature	Avg.	Range	Avg.	Range	Avg.	Range	Avg.	Range
Clickbait	0.008	0 - 1.0	0.024	0 - 1.0	-	-	-	-
Insight	1.738	0 - 7.7	2.160	0 - 18.1	1.580	0 - 33.3	1.234	0 - 50.0
Causation	1.387	0 - 6.0	1.194	0 - 11.1	1.088	0 - 20.0	0.795	0 - 25.0
Discrepancy	0.986	0 - 9.5	1.233	0 - 10.0	1.264	0 - 40.0	0.852	0 - 25.0
Tentative	1.348	0 - 10.3	1.678	0 - 9.1	1.892	0 - 33.3	1.109	0 - 33.3
Certitude	0.268	0 - 3.0	0.701	0 - 10.0	0.526	0 - 40.0	0.638	0 - 100.0
Differentiation	2.471	0 - 7.9	2.006	0 - 20.0	2.381	0 - 22.2	1.558	0 - 25.0
Toxicity	0.01	0 - 2.0	0.27	0 - 10.5	0.576	0 - 100.0	0.396	0 - 25.0
Pos. Emotion	0.292	0 - 7.1	0.498	0 - 13.3	0.519	0 - 33.3	0.415	0 - 20.0
Neg. Emotion	0.430	0 - 8.9	0.667	0 - 8.3	0.804	0 - 50.0	0.510	0 - 25.0
Soc. Behaviour	5.557	0 - 18.9	4.147	0 - 16.7	2.760	0 - 33.3	1.828	0 - 50.0
Pro Social	0.608	0 - 6.1	0.48	0 - 7.1	0.351	0 - 20.0	0.233	0 - 50.0
Politeness	0.344	0 - 4.9	0.157	0 - 4.8	0.288	0 - 16.7	0.129	0 - 10.0
Conflict	0.646	0 - 9.1	0.612	0 - 11.8	0.269	0 - 25.0	0.181	0 - 9.09
Morality	0.439	0 - 5.5	0.503	0 - 14.3	0.351	0 - 33.3	0.224	0 - 25.0
Communication	2.701	0 - 14.3	1.562	0 - 11.1	1.218	0 - 25.0	0.742	0 - 50.0

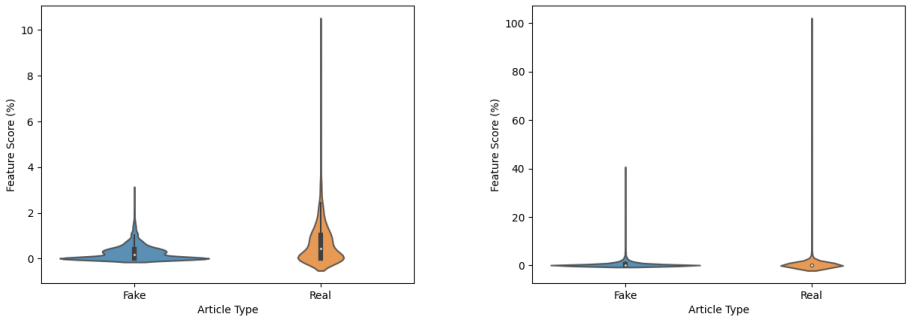
Average represents the **Average Embedding Score** obtained for each feature from LIWC. **Range** represents the lowest and highest **Feature Embedding Score** obtained for each feature from LIWC. **Clickbait attributes** relate to the news headlines only, thus, do not have a comment score.

Although these sources provide news from truthful, current events, their satirical approach in writing results in such articles being outliers in the dataset.



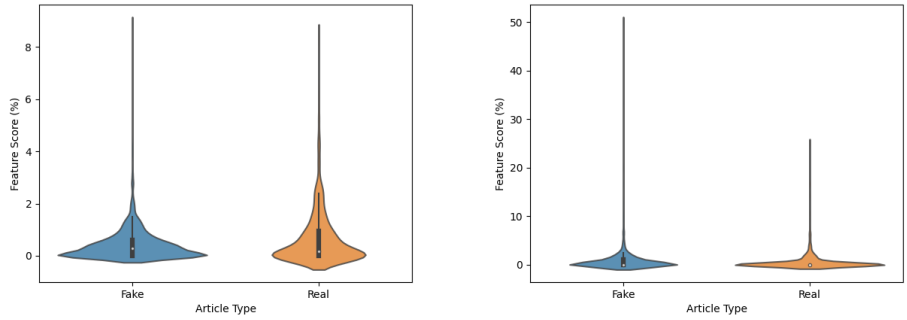
(a) Article (b) Comments

Figure 4.1: Score Distribution of **DIA Discrepancy** Feature for All News Articles and Comments



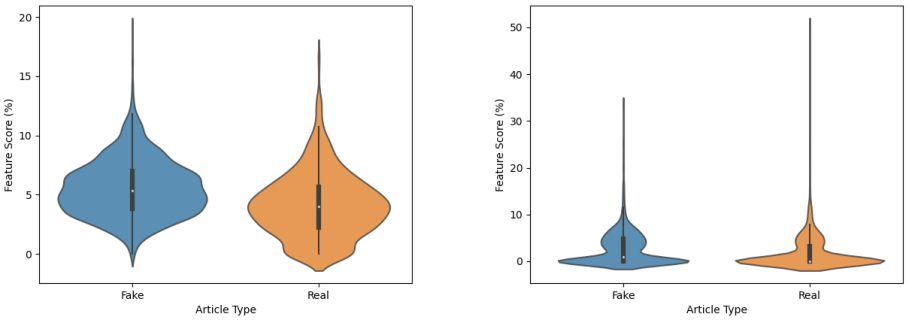
(a) Article (b) Comments

Figure 4.2: Score Distribution of **DIA Certitude** Feature for All News Articles and Comments



(a) Article (b) Comments

Figure 4.3: Score Distribution of **Negative Emotion** Feature for All News Articles and Comments



(a) Article (b) Comments

Figure 4.4: Score Distribution of **Social Behaviour** Feature for All News Articles and Comments

4.3 Effectiveness of User Comments

Text-based classification models typically focus on the writing styles of the news article; however, recent studies have shown that the inclusion of user engagement in the classification task has provided improved performance for the classification task. Unlike news articles, user comments provide details about the public opinion on the subject of the article, and allow classification models to analyze the stances of several individuals compared to only the author's writing perspective. Negative comments commonly reflect opposing views on an article's stance, while positive comments show support. When analyzing user comments and their writing style, we find a correlation between user comments with more toxic language and news articles which have higher negative emotional scores. Since fake news is commonly written to drive discussion on controversial topics [20] to generate user engagement and promote news propagation, we see a correlation with users expressing strong emotional opinions about the news, resulting in more toxic language in the comments. In our analysis, as indicated by Table 4.3, real news tends to have more negative emotions in its news content, as topics such as disasters, violence, and politics are common topics that are covered, resulting in a higher negative emotional score.

In addition, we explore how different social behaviours impact fake and real news, and present our findings in Table 4.3. In our analysis, we identify that fake news tends to exhibit increased scores in traits such as morality and interpersonal conflict in its news content compared to real news. In our analysis, we identify a morality score of 0.5 to suggest ethically just articles, while higher scores of conflict indicate news with more conflict-related information. Alternatively, real news shows higher scores for pro-social behaviour, politeness, and communication in both user comments and the article's text. Lower scores of these three traits indicate that the news is more pro-social, polite, and communicative.

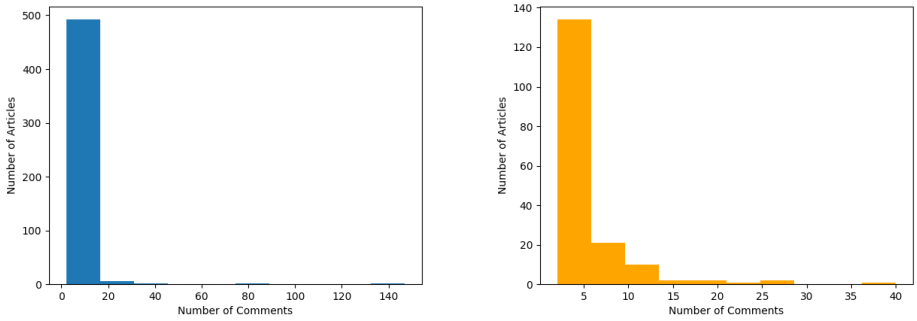
In Figures 4.5 to 4.7, we highlight the change in comments obtained over time since the initial publication of each news article in the news corpus. To enhance the clarity of our results, we omit news articles that received no comments within the first eight hours after publication from these figures, as many news articles in the news corpus do not receive any user comments during this time period. Our results indicate that fake news articles tend to attract more user comments during the initial four hours after publication compared to real news articles. However, as time progresses, the comment distribution tends to equalize between fake and real news, as seen in Figure 4.7, where both types of articles exhibit similar comment distribution patterns.

With our analysis, we highlight the value of using linguistic and psychological features along with user comments to enhance classification tasks and improve the performance of existing detection models. Leveraging these patterns allows classification models to more effectively distinguish between fake and real news.

4.4 Effectiveness of PL-NCC Dataset Against Baseline Models

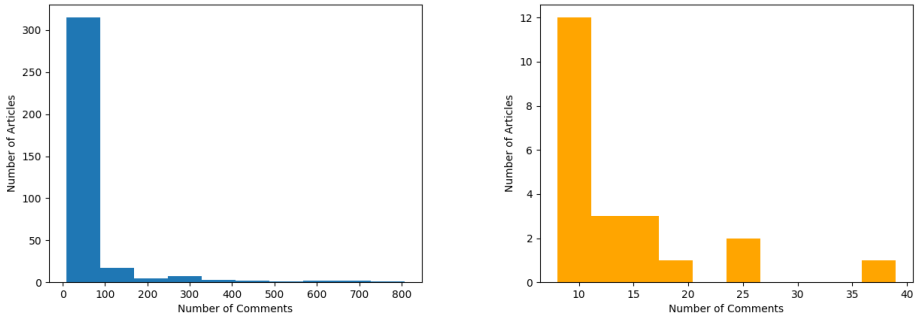
After a thorough analysis of each feature's behaviour in fake and real news, we aim to determine the efficacy of our proposed PL-NCC dataset by comparing its performance against leading datasets such as NELA-GT and Fakeddit. We compare each dataset against a series of state-of-the-art classification models, as outlined in Table 4.3. To determine the effectiveness of our proposed features, we compare the performance of only the article's text and headline, followed by another set of experiments using all features in our PL-NCC dataset.

To test the performance of our PL-NCC dataset and NCC classification model, we execute the detection task using four deep-learning models commonly used for the clas-



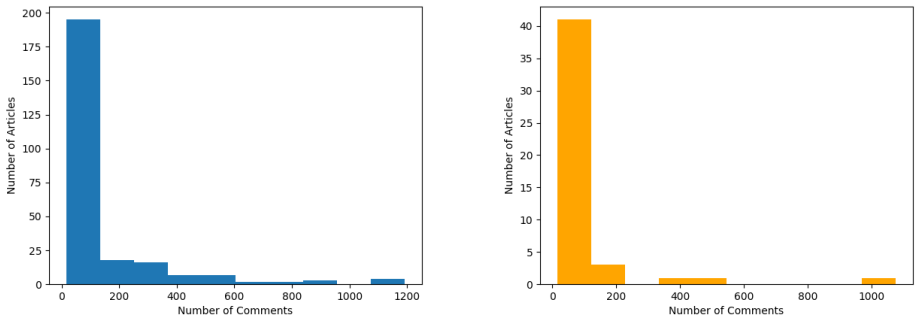
(a) Fake News (b) Real News

Figure 4.5: Number of Comments Collected in the First One Hour of Propagation



(a) Fake News (b) Real News

Figure 4.6: Number of Comments Collected in the First Four Hours of Propagation



(a) Fake News (b) Real News

Figure 4.7: Number of Comments Collected in the First Eight Hours of Propagation

Table 4.4: Fake news detection accuracy on our dataset combined over 5-folds

PL-NCC Dataset						
	Text-Only Input		*With Proposed Features*		**Change in Accuracy**	
Model	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
BERT	0.868	0.908	0.972	0.980	+ 10.4%	+ 7.2%
DistilBERT	0.871	0.908	0.972	0.979	+ 10.1%	+ 7.1%
MLP	0.957	0.969	0.975	0.982	+ 1.8%	+ 1.3%
XGBoost	0.945	0.961	0.975	0.982	+ 3.0%	+ 2.1%
CNN	0.949	0.963	0.958	0.964	+ 0.9%	+ 0.1%
DTC	0.855	0.897	0.917	0.941	+ 6.2%	+ 4.4%
NCC	**0.957**	**0.969**	**0.978**	**0.984**	+ 2.1%	+ 1.5%

Proposed Features include User Comments, Linguistic and Psychological Features.
BOW embedding of article headline and texts are the input to the models.
The text-only input excludes user comments, linguistic, and psychological features.

Table 4.5: Fake news detection accuracy on original dataset combined over 5-folds

	NELA-GT Dataset		**Fakeddit Dataset**	
Model	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
BERT	0.920	0.942	0.789	0.847
DistilBERT	0.948	0.962	0.767	0.833
MLP	0.951	0.965	0.656	0.754
XGBoost	0.952	0.966	0.706	0.811
CNN	0.948	0.965	0.679	0.779
DTC	0.853	0.894	0.673	0.770

sification task of fake news [27, 28, 46, 56], including Multilayer Perception (MLP) [25], Convolutional Neural Networks (CNN) [29], XGBoost and a standard Decision Tree (DTC) based classifier, and BERT learning models. The results of our experiments are shown in Tables 4.4 and 4.5.

4.4.1 Performance comparison

Our results showcase that the PL-NCC dataset and NCC classification model outperform baseline models by up to twelve percent, as shown in Tables 4.4 and 4.5. For our work, we exclude the performance of our NCC classification model in the baseline comparisons illustrated in Table 4.4, as these tests only use the basic text from each respective dataset, and excludes the proposed linear layer of our NCC classification model. Thus, the results for this experiment are identical to the basic MLP classification model as illustrated in the table. In our analysis, the decision tree model is the least performant compared to all other baselines, and our model outperforms the leading CNN classification model by one percent. In our analysis, we also compare the baseline and PL-NCC datasets against different language learning models, including BERT embedding. In our comparison, we include both BERT and DistilBERT embedding models to test our model's performance. As discussed in Section 3.5.1, we compare two different BERT language models, BERT and DistilBERT, to determine which model is more efficient when creating the respective BERT linguistic features. As illustrated in our baseline comparison, the DistilBERT language model performs better using all three testing datasets compared to the BERT language model. Based on this set of experiments, we conclude that the proposed PL-NCC dataset and NCC classification model perform better than baseline models due to the inclusion of both linguistic and psychological features, user comments and our proposed linear feed-forward layer. When comparing each baseline model, the performance of each baseline model is succeeded by the performance accuracy of our NCC classification model when using the PL-NCC dataset. In addition, with the same feature sets, we notice better results when utilizing the article's text and user comments from the PL-NCC dataset, compared to relying solely on NELA-GT's article content or Fakeddit's comment data. The results of our baseline dataset comparison, as illustrated in Table 4.4 indicates that the Fakeddit dataset performs worse for the classification tasks compared to the NELA-GT dataset. Analysis of the results indicates that

the performance drop is a result of the missing news article content, as Fakeddit only provides the raw user comments related to each news article. However, the results of our experiments show that our proposed NCC classification model outperforms both benchmark datasets when using the proposed linguistic and psychological features included in the PL-NCC dataset.

4.5 Effectiveness of Linguistic and Psychological Features

In this set of experiments, we determine how different permutations of linguistic and psychological features affect the performance of the classification task. Our goal is to determine the best feature set using the PL-NCC dataset to perform fake news classification. We utilize our NCC classification model and keep all hyperparameters constant while only changing the feature sets used. When analyzing the results, we see up to a twenty percent improvement from the final experiment as more proposed features are included in the classification task. The results of these experiments indicate that when including user comments and linguistic and psychological traits in our dataset, fake news classification models become more effective at detecting fake news compared to conventional text-based models and datasets.

4.5.1 Experiment one

We present the results of our complete model's performance in Table 4.6. The proposed NCC classification model performs best when utilizing a combination of all features proposed in this work, including user comments, and linguistic and psychological features.

Table 4.6: Effectiveness of complete model using NCC Classification Model

Note: Bold text indicates best performing feature set.

Feature groups	Accuracy	F1 score
BERT + POS + CFG + DIA + CBA + E + SW + SB	**0.97838**	**0.98431**

Legend: BERT embedding (BERT), parts of speech (POS), context-free grammar (CFG), disinformation-related attributes (DIA), clickbait related attributes (CBA), emotions (E), swear words (SW), social behaviour (SB)

4.5.2 Experiment two

The next experiment we conduct examines the effects of the classification task when one psychological feature group is removed from the feature set. The results of these experiments are shown in Table 4.7. Our top-performing feature sets include BERT, emotion, and social behaviour, as well as BERT, disinformation-related attributes, clickbait-related attributes, emotion, and social behaviour. These feature combinations achieve a ninety-seven percent accuracy in our model's performance. When we remove social behaviour and BERT from the testing feature set, the performance of the model decreases by up to ten percent accuracy compared to the better-performing combination. These results indicate that social behavioural features and BERT contribute positively to fake news classification.

Table 4.7: Effectiveness of linguistic and two psychological features using NCC Classification Model

Note: Bold text indicates best performing feature set.

Feature groups	Accuracy	F1 Score
BERT + E + SW	0.96018	0.97138
POS + E + SW	0.97383	0.98110
CFG + E + SW	0.92491	0.94554
DIA + E + SW	0.88282	0.89401
CBA + E + SW	0.84300	0.89401
BERT + POS + CFG + E + SW	0.97042	0.97862
BERT + DIA + CBA + E + SW	0.96018	0.97129
POS + CFG + DIA + CBA + E + SW	0.97042	0.98765
BERT + POS + CFG + DIA + CBA + E + SW	0.96638	0.98441
BERT + E + SB	0.95449	0.96721
POS + E + SB	0.96701	0.97617
CFG + E + SB	0.92264	0.94525
DIA + E + SB	0.93970	0.95709
CBA + E + SB	0.86007	0.90277
BERT + POS + CFG + E + SB	0.97538	0.97439
BERT + DIA + CBA + E + SB	0.96473	0.97461
POS + CFG + DIA + CBA + E + SB	0.97538	0.97439
BERT + POS + CFG + DIA + CBA + E + SB	**0.97611**	**0.98274**
BERT + SW + SB	0.95222	0.96569

POS + SW + SB	0.97497	0.98191
CFG + SW + SB	0.92150	0.94330
DIA + SW + SB	0.89534	0.92879
CBA + SW + SB	0.85779	0.89712
BERT + POS + CFG + SW + SB	0.97270	0.98020
BERT + DIA + CBA + SW + SB	0.96815	0.97701
POS + CFG + DIA + CBA + SW + SB	0.96294	0.98765
BERT + POS + CFG + DIA + CBA + SW + SB	0.96066	0.97439

Legend: BERT embedding (BERT), parts of speech (POS), context-free grammar (CFG), disinformation-related attributes (DIA), clickbait related attributes (CBA), emotions (E), swear words (SW), social behaviour (SB)

4.5.3 Experiment three

We then expand on the previous experiment by removing two psychological feature groups from the testing feature sets. The results are presented in Table 4.8, which indicates that the best-performing feature set consists of all linguistic features and emotions.

While the best-performing model of this experiment includes emotions, compared to the inclusion of social behavioural and word toxicity features, the results show a higher variability in terms of model accuracy when different linguistic feature combinations are used. The feature sets containing emotions exhibit an average performance of around ninety-two percent, while feature groups including social behavioural features consistently outperform both word toxicity and emotions, maintaining an average performance of ninety-five percent. Through analysis, we identify that the link to this performance difference is a result of the quantity of individual features within each psychological feature group. Emotions consists of fewer individual features compared to the social

behavioural group. From this set of experiments, we conclude that adding more features to the testing feature set provides the model with more training information, leading to more consistent performance in terms of model accuracy.

Table 4.8: Effectiveness of linguistic and one psychological features using NCC Classification Model

Note: Bold text indicates best performing feature set.

Feature groups	Accuracy	F1 Score
BERT + E	0.96132	0.97218
POS + E	0.97156	0.97942
CFG + E	0.91354	0.93740
DIA + E	0.82253	0.88563
CBA + E	0.80546	0.86567
BERT + POS + CFG + E	0.97156	0.97939
BERT + DIA + CBA + E	0.96359	0.97373
POS + CFG + DIA + CBA + E	0.96701	0.97633
BERT + POS + CFG + DIA + CBA + E	**0.97638**	**0.96439**
BERT + SW	0.95563	0.96811
POS + SW	0.97042	0.97862
CFG + SW	0.92605	0.94650
DIA + SW	0.84187	0.89557
CBA + SW	0.81456	0.88041
BERT + POS + CFG + SW	0.96952	0.97515
BERT + DIA + CBA + SW	0.95563	0.96790
POS + CFG + DIA + CBA + SW	0.96701	0.97625

BERT + POS + CFG + DIA + CBA + SW	0.96725	0.97358
BERT + SB	0.96359	0.97381
POS + SB	0.96815	0.97697
CFG + SB	0.92150	0.94321
DIA + SB	0.91923	0.94195
CBA + SB	0.86803	0.91036
BERT + POS + CFG + SB	0.96928	0.97778
BERT + DIA + CBA + SB	0.95791	0.96960
POS + CFG + DIA + CBA + SB	0.96928	0.97778
BERT + POS + CFG + DIA + CBA + SB	0.97511	0.98274

Legend: BERT embedding (BERT), parts of speech (POS), context-free grammar (CFG), disinformation-related attributes (DIA), clickbait related attributes (CBA), emotions (E), swear words (SW), social behaviour (SB)

4.5.4 Experiment four

For the following experiments, we test the unique properties of both linguistic and psychological features individually to determine each's contribution to fake news classification and obtain a baseline for our results. In experiment four, we exclude all psychological features from the testing set, and only showcase the performance of the proposed NCC classification model using linguistic features. We then compare the results of this experiment against our previous experiments to analyze how the inclusion of psychological features affects the performance of the classification task. The results of this experiment are shown in Table 4.9.

Similar to our previous set of experiments, we identify that feature sets including BERT embedding show better performance. In this set of experiments, the best-

performing feature set includes all linguistic features by achieving an ninety-seven percent accuracy and a ninety-eight percent F1 score.

In this set of experiments, we observe that disinformation (DIA) and clickbait-related attributes (CBA) perform poorly when used alone for the classification task. However, when DIA is used in conjunction with other linguistic features, an improvement in performance is evident. Similar to our conclusion from experiment three, disinformation and click-bait related attributes have fewer individual features extracted from the text, compared to the extracted BERT, POS, and CFG features. While linguistic features such as BERT, POS, and CFG generate 768 unique numeric embeddings from the input text, disinformation and clickbait-related attributes have six and forty-seven main feature groups respectively. The use of DIA and CBA alone causes the classification model to struggle in identifying distinct traits unique to fake news, while introducing training noise, thus, reducing the model's performance. However, when used in conjunction with other linguistic features, the model is better capable of identifying the unique characteristics of disinformation-related and clickbait attributes in fake news articles to supplement the trained linguistic features. Thus, we conclude that using a combination of various linguistic feature groups allows the classification model to accurately classify fake news better.

Table 4.9: Effectiveness of linguistic features only using NCC Classification Model

Note: Bold text indicates best performing feature set.

Feature groups	**Accuracy**	**F1 Score**
BERT	0.95791	0.96955
POS	0.97270	0.98030
CFG	0.90216	0.92797

DIA	0.83618	0.89318
CBA	0.79295	0.86888
DIA + CBA	0.88055	0.91879
BERT + POS	0.97383	0.98107
BERT + CFG	0.94767	0.96230
BERT + POS + CFG	0.97270	0.98023
BERT + DIA + CBA	0.96246	0.97302
POS + CFG + DIA + CBA	0.96701	0.97625
BERT + POS + CFG + DIA + CBA	**0.97611**	**0.98274**

Legend: BERT embedding (BERT), parts of speech (POS), context-free grammar (CFG), disinformation-related attributes (DIA), clickbait related attributes (CBA)

4.5.5 Experiment five

The next experiment is conducted to determine the importance of linguistic features in fake news detection. To illustrate these results, we exclude all linguistic features from the model, and only perform our classification tasks using the psychological features extracted in our PL-NCC dataset. The results of our experiments are shown in Table 4.10. After analysis, we notice poorer performance when using only psychological features. However, in these experiments, we continue to see that feature sets including social behavioural traits exhibit better performance compared to emotions and swear words. We conclude that although psychological features are not as effective when used in isolation, they enhance the performance of the classification task when used in conjunction with linguistic features. Particularly, social behavioural traits provide the largest and most consistent improvement as a psychological feature group for the classification task.

Table 4.10: Effectiveness of psychological feature groups only using NCC Classification Model

Note: Bold text indicates best performing feature set.

Feature groups	Accuracy	F1 Score
Emotions	0.83049	0.88772
Swear words	0.80774	0.87745
Social behaviour	0.82025	0.88450
Swear words	0.92378	0.94495
Emotions + Social behaviour	0.87372	0.91375
Swear words + Social behaviour	0.93402	0.95345
Emotions + Swear words + Social behaviour	**0.93588**	**0.95336**

4.5.6 Experiment six

Finally, we conduct an experiment to analyze the effectiveness of using user comments within the classification task. For this set of experiments, we analyze how user comments affect the performance of our classification model using all features of our PL-NCC dataset. We test our NCC classification model on the first eight hours of user comments and report our results in Table 4.11. When no comments are present, we see a classification accuracy of about ninety-five percent; however, once user comments are introduced after one hour, we see a dip of two percent in the model's accuracy. After analyzing the data, we relate this dip in performance to the few user comments that are received compared to the multiple news articles without user comments. Studies have shown that during the propagation of news, user engagement is scarce during its initial propagation, and this data may be unavailable during the first hour of classification, and

we see similar results as shown in Table 4.11. However, after more user comments are collected as time progresses, we eventually see our model perform better than without user comments, exhibiting a classification score of about ninety-seven percent accuracy, which is about two percent higher than the initial results. We conclude that user comments have positive merits for the classification of fake news.

Table 4.11: Effectiveness of user comments using NCC Classification Model

Note: Bold text indicates best performing feature set.

Propagation Time	**Accuracy**	**F1 Score**	**Total**	**Min.**	**Avg.**	**Max.**
No Comments	0.95355	0.96661	0	0	0	0
One Hour	0.94866	0.96309	3,299	0	1.13	147
Two Hour	0.95844	0.97012	7,799	0	2.66	382
Three Hour	0.94621	0.96071	12,505	0	4.26	424
Four Hour	0.95110	0.96429	20,606	0	7.01	2,001
Five Hour	0.94866	0.96257	26,447	0	9.00	2,001
Six Hour	0.95599	0.96786	31,123	0	10.59	2,001
Seven Hour	0.96577	0.97491	35,655	0	12.12	2,001
Eight Hour	**0.97311**	**0.98032**	**43,005**	**0**	**14.59**	**2,014**

Total represents the total number of comments collected in the propagation period. **Min.** represents the minimum number of comments collected. **Avg.** represents the average number of comments collected. **Max.** represents the maximum number of comments collected.

Table 4.12: Effectiveness of linear layer using NCC Classification Model

Has Linear Layer	(Complete Model) Accuracy	F1 Score
Yes	**0.978**	**0.984**
No	0.957	0.969

4.6 Effectiveness of linear layer

In our work, we include a feed-forward linear layer to the NCC classification model to improve the accuracy of classification for the psychological features extracted. To test the efficacy of this linear layer, we run the classification task against the PL-NCC dataset with and without the linear layer. The results of these experiments are shown in Table 4.12. The standard MLP model without the linear layer performs similarly to the baseline MLP models presented in Tables 4.4 and 4.5. The results of this experiment indicate that there is better performance with the use of a linear layer, as we see up to a two percent increase in performance accuracy.

The linear layer amplifies the weights of psychological features often found in fake news, such as word toxicity and pro-social behaviour, whilst reducing the significance of psychological features common in true news. By emphasizing the psychological traits associated with fake news, our NCC model can more accurately identify the veracity of fake news compared to the baseline MLP models.

4.7 Effectiveness of neural representation with linguistic features

Recent work on fake news detection [16, 56] has focused on using deep learning techniques to improve the classification of fake news detection. As part of our research, we

aim to identify the significance of deep learning language models, such as BERT, and how they compare to traditional linguistic features such as bag-of-words (BOW). Work by Dacrema et al. showcases how BERT embeddings may not be as effective in the task of fake news classification; however, we present the results of our experiments in Table 4.13, which showcases the unique patterns obtained from BERT and BOW features.

While both BERT and BOW produce numerical embeddings of the input text, BERT performs its prediction using masked words within the article to create its embeddings. However, since BERT is a large model with numerous parameters, it does not perform well when using smaller datasets, such as PolitiFact, as seen in our results. We conclude this is a result of overfitting done within the embedding model. In contrast, BOW generates embeddings by counting word occurrences and does not perform its own classification of the input text. However, when we analyze the results of BERT embedding using a larger dataset, such as our PL-NCC dataset, we see much better performance using BERT embeddings compared to BOW. Our analysis showcases that BERT embedding performs much better with larger datasets as there is a larger vocabulary from the input dataset for the BERT embedding model to train on. Thus, although BERT may underperform with smaller datasets compared to BOW, it still proves to be a novel tool in the classification of fake news.

Table 4.13: Effectiveness of neural representation with linguistic features (BOW/BERT) using NCC Classification Model

Neural representation	**BOW**		**BERT**	
Dataset	**Accuracy**	**F1 Score**	**Accuracy**	**F1 Score**
PolitiFact	**0.9028**	**0.9091**	0.8333	0.8462
PL-NCC	0.9682	0.97674	**0.9731**	**0.9803**

Chapter 5

Conclusion and Future Work

As fake news becomes more prominent in modern day, it is imperative that research is focused on the mitigation of fake news propagation, both through the development of robust news datasets as well as accurate and efficient classification models. In this work, we proposed a novel approach to consolidate two state-of-the-art fake news datasets NELA-GT and Fakeddit, by introducing a set of linguistic and psychological features extracted from the news dataset, as well as creating a consolidated dataset to include both news articles and their respective user comments. We classify this dataset as the Psycho-Linguistic News Content and Comments (PL-NCC) dataset, which we leverage to develop an improved MLP classification model, which introduces a feed-forward linear layer to enhance the efficacy of psychological characteristics extracted from the news text and user comments. We name this new approach as the News Content and Comments (NCC) classification model.

In this work, we make contributions to fake news research by: 1) assessing the linguistic and psychological characteristics within news articles to better classify fake news; 2) leveraging the unique linguistic and psychological characteristics from the news text and user comments to include in a consolidated dataset to better improve the classification task of fake news; 3) exploring the benefits of user engagement in the classification tasks;

4) exploring different language learning models, including BERT embeddings and bag-of-word models to access the benefits of each method; and 5) determining the effectiveness of our proposed PL-NCC dataset using our NCC classification model, which introduces a feed-forward linear layer into the MLP classifier to better identify the veracity of fake news. We execute a series of experiments to demonstrate the effectiveness of our PL-NCC dataset and NCC classification model, and our results indicate that the proposed models provide significant improvements over existing state-of-the-art classification models and datasets, showing upwards to a twelve percent increase in classification accuracy.

Analysis of our experiments indicates significant benefits to the user of linguistic and psychological features in fake news classification when used in conjunction with user comments and the article text. While psychological features are not as effective for the classification task when used in isolation, we identify that these features excel when used in conjunction with other linguistic features. Specifically, our experiments showcase that social behavioural traits in particular demonstrate the best performance for the classification task. In our set of experiments, the psychological features improve the performance of the classification task. Thus, We include these features as part of our dataset so they can be integrated into newer classification models as deemed fit. In addition, we identify that the inclusion of the feed-forward linear layer in our work enhances the efficacy of the proposed psychological features, improving the classification of fake news even further.

5.1 Limitations and Future Work

Due to restrictions made by the Reddit API, our work only includes the first eight hours of user comments for each news post in the PL-NCC dataset. While sufficient for early fake news detection models, this restriction limits general text-based classification models. Future work can include extracting more user comments using Reddit's API

to expand the time window available in the PL-NCC dataset. Although possible, this process is a tedious task and will require more effort to implement. When labelling our PL-NCC dataset, we rely on NELA-GT's and Fakeddit's source-based labels. Although effective for the classification task, there are better solutions available to label our dataset, including the process of labelling the news corpus on a per-article basis instead of at the source level. Per-article labelling methods will better improve the accuracy of our PL-NCC dataset in relation to the actual news content, better assisting the classification task of detection models. In addition, we plan to extend our research to study other linguistic and psychological features which can be extracted from the article's text and user comments. Our research has shown to be effective with the contribution of the proposed features, and we are confident that the model can be improved further by identifying additional characteristics in the article's text and user comments to better distinguish fake news from real.

Appendix A

Examples of Fake News Stories

A.1 Trump declares he is having a good day as redacted Mueller report is released

"They're having a good day. I'm having a good day, too. It was called no collusion. No obstruction," Trump said to cheers at a Wounded Warriors event at the White House. "There never was by the way and there never will be. And we do have to get to the bottom of these things I will say. This should've never happened ... I say this in front of my friends, this should never happen to another president again. This hoax – it should never ... The special counsel's investigation into possible collusion found that members of the Trump campaign knew they would benefit from Russia's illegal actions to influence the election, but didn't take criminal steps to help, the Mueller report said. Mueller also makes clear Congress can continue to investigate Trump."

With respect to whether the President can be found to have obstructed justice by exercising his powers under Article II of the Constitution, we concluded that Congress has the authority to prohibit a President's corrupt use of the integrity of the administration of justice. Attorney General William Barr released the report after holding a press

conference where he defended his conclusion there wasn't sufficient evidence to prosecute an obstruction case. After nearly two years of investigation, thousands of subpoenas, and hundreds of warrants and witness interviews, the special counsel confirmed that the Russian government sponsored efforts to illegally interfere with the 2016 presidential election, but did not find that the Trump campaign or other Americans colluded in those schemes,

A.2 California becomes first state to ban fur trapping after gov Newsom signs law

California has enacted a new ban on fur trapping for animal pelts, making it the first state to outlaw a centuries-old livelihood that was intertwined with the rise of the Western frontier. The Wildlife Protection Act of 2019, signed into law by Gov. Gavin Newsom on Wednesday, prohibits commercial or recreational trapping on both public and private lands. Assemblywoman Lorena Gonzalez (D-San Diego), who introduced the legislation, said it was time to end fur trapping.

"It seems especially cruel, obviously, and it's just unnecessary." Although commercial trapping was an early part of California's economy, opening the San Francisco Bay Area to international commerce even before the 1848 California Gold Rush, its fortunes have waned over many decades. Gonzalez said that the roughly six dozen trappers still working in the state, down from more than 5,000 a century ago, can not afford to pay the full cost of implementing and regulating their industry.

The ban also comes as California lawmakers consider more aggressive measures to protect animals and wildlife, considering proposals to ban the sale of all fur products, including fur coats, and to outlaw the use of animals in any circus in the state, with the exception of domesticated horses, dogs and cats. "There's been a real change in attitudes about how we treat animals," Gonzalez said.

A total of 68 trappers reported killing 1,568 animals statewide in 2017, according to the California Department of Fish and Wildlife. Among the 10 species reported taken were coyote. Trapped animals are strangled, shot or beaten to death, with care taken not to damage pelts before skinning them. Under the law, using traps to catch gophers, house mice, rats, moles and voles would still be permitted. The law followed a 2013 public outcry when conservationist Tom O'Key in 2013 discovered a bobcat trap illegally set on his property near the edge of Joshua Tree National Park. O'Key stumbled upon the trap chained to a jojoba bush and camouflaged 720,000-acre park, where the big cats are a dominant force in the ecosystem.

He immediately alerted neighbors and contacted the San Bernardino County Sheriff's Department and Hi-Desert Star newspaper, triggering an angry tide of complaints that put a spotlight on the practice of trapping, killing and skinning bobcats to supply fur markets in China, Russia and Greece. "I could not have guessed in a million years," O'Key said in an interview, that trap would spark an unstoppable movement. Assemblyman Richard Bloom (D-Santa Monica) pushed through his Bobcat Protection Act of 2013, which was in response to petition drives, social media campaigns and telephone calls to lawmakers from wildlife advocates who decried trapping and killing as a cruel trade.

Eight months after O'Key sounded the alarm in Joshua Tree, the California Fish and Game Commission voted 3 to 2 to ban commercial bobcat trapping statewide. The Wildlife Protection Act of 2019 argues that the small number of active trappers in the state cost of implementing and regulating their industry as required by law. It was backed by the Center for Biological Diversity, and the nonprofit group Social Compassion in Legislation, which spearheaded a recent bill that put an end to the sale of mill-bred dogs, cats and rabbits.

Opponents included the California Farm Bureau Federation, which warned that the bill if passed, could have significant economic consequences for the agriculture industry.

The trapping industry declined over decades in California. Before California's population ballooned, trapping played a significant role in the extirpation of wolves and wolverines and the severe declines of sea otters, fishers, martens, beavers and other fur-bearing species.

Over the last two decades, animal protectionists have partnered with mainstream environmental groups to put pressure on state and federal wildlife authorities, and to take their animal-cruelty concerns to the voters. "Trappers are anachronistic", they said, "and their snares subject wildlife to horrific suffering".

"The signing of this bill into law is the result of compelling public opinion regarding animal cruelty", said Judie Mancuso, founder and president of Social Compassion in Legislation.

Appendix B

Examples of True News Stories

B.1 James Harden Chris Paul deny rumors of discord say they are fully committed to team at State Farm

Houston attempting to set the record straight about reports of a toxic work relationship between the two players, James Harden and Chris Paul publicly denied rumors of discord Friday and reiterated that they remain fully committed to the team at State Farm Insurance. "Things may get heated from time to time, but at the end of the day, we both know all we want success for State Farm," said Harden in a statement to the press, hoping to assure fans that they would be seeing the pair bouncing jokes off each other.

"Sure, sometimes we might clash over the artistic direction of a commercial or if someone accidentally steps on someone else's line, but it's all done out of love for State Farm. Even though we may lose our heads now and then, we're still very passionate about our performance in these commercials. We're both in this for the long haul and want to build a real legacy here." Harden and Paul both agreed that with a little more could end up as the number-one insurance company in the country.

B.2 Tyson holds contest to let fans submit new ideas for torturing chicken to death

SPRINGDALE are announcing that the winner would receive a year's supply of their frozen poultry products killed in their method of choosing. Tyson Foods unveiled a contest Thursday to let fans submit new ideas for torturing chickens to death.

"We know our fans love expressing themselves as much as they love chicken nuggets, which is why we're asking you to send us your most creative ideas for brutally slaughtering chickens at our processing plant, no matter how outlandish, disgusting, or painful," said Tyson spokesperson John Jaworski, for lethally mistreating the animals using the hashtag FowlPlay, encouraging them to upload videos of themselves trying out their idea on a live chicken.

"We're looking for any and all concepts for chicken torture, whether that's firing a chicken out of a cannon into a brick wall, smashing it with a hammer, or slowly cutting its throat with a dull knife" those are just a few of our current processes to get you started. "Our winners will get a chance for in Arkansas, where we'll let you execute hundreds of chickens yourself using your suggestion. Bonus points are awarded to any idea that can kill over a thousand chickens simultaneously or cause them to squawk in immense pain for over 24 hours. All right, time to get creative!"

At press time, Tyson Foods announced that the first winner, a Twitter user who submitted the idea of cramming dozens of chickens into a tank and pouring chicken blood into it until they all gradually suffocate at company headquarters getting a photo op with the CEO.

Appendix C

Examples of User Comments

C.1 California becomes first state to ban fur trapping after gov Newsom signs law

UC 1: "French Canadians hate California State Government for this one simple trick!"

UC 2: "under current law that would be no more than 24 hours. Many people who support this law would consider this cruel. How do you feel about it and how do you feel about kill traps as an alternative? As someone who has spent a great deal of time in the CA wilderness"

UC 3: "I can assure you that traps don't always work as intended. I saw a rabbit that had a leg caught in a snare and had been dead for a few days. The snare was obviously lost or forgotten about. Whenever trapping it is of course the responsibility of the trapper to not cause undue stress. We have had trap lines that span up to 500 miles that we would check weekly. Some times you get something some times you do not. The unfortunate reality is that some trappers are not responsible"

UC 4: "Now when it comes to unintentional trapping of stuff. This was and still is a big

issue in Alaska where I trap. We would be going after wolves that are encroaching on cities or settlements and occasional catch a pet dog. I would use signage notifying locals and trail goers that there are active wolf traps"

UC 5: "Laws and regulations control populations. Anyone who was going to poach and take endangered or threatened animals illegally will continue to do so. This law only hurts lawful trappers and for no good reason"

C.2 Tyson holds contest to let fans submit new ideas for torturing chicken to death

UC 1: "This made me sad...."

UC 2: "/r/morbidreality ?"

UC 3: "KFC wants to: know your location"

UC 4: "Someone's going to eat this up."

UC 5: "Oh look! The preserved bug candy is in stock!"

List of Appendices

References

[1] Acerbi, A. Cognitive attraction and online misinformation. *Palgrave Communications 5*, 1 (2019).

[2] Ahmed, S., Hinkelmann, K., and Corradini, F. Development of fake news model using machine learning through natural language processing. *arXiv preprint arXiv:2201.07489* (2022).

[3] Allcott, H., and Gentzkow, M. Social media and fake news in the 2016 election. *Journal of economic perspectives 31*, 2 (2017), 211–236.

[4] Balcar, S., Skrhak, V., and Peska, L. Rank-sensitive proportional aggregations in dynamic recommendation scenarios. *User Modeling and User-Adapted Interaction* (2022), 1–62.

[5] Baly, R., Karadzhov, G., Alexandrov, D., Glass, J., and Nakov, P. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765* (2018).

[6] Bierhoff, H. W. *Prosocial behaviour*. Psychology Press, 2002.

[7] Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin* (2022).

[8] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (2016), pp. 785–794.

[9] Cheng, L., Guo, R., Shu, K., and Liu, H. Causal understanding of fake news dissemination on social media. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021), pp. 148–157.

[10] Dadkhah, S., Zhang, X., Weismann, A. G., Firouzi, A., and Ghorbani, A. A. Truthseeker: The largest social media ground-truth dataset for real/fake content.

[11] de Oliveira, N. R., Pisa, P. S., Lopez, M. A., de Medeiros, D. S. V., and Mattos, D. M. Identifying fake news on social networks based on natural language processing: trends and challenges. *Information 12*, 1 (2021), 38.

[12] Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Hoi, G. W. S., and Zubiaga, A. Semeval-2017 task 8: Rumoureval: Determining rumour veracity and support for rumours. *arXiv preprint arXiv:1704.05972* (2017).

[13] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Effron, D. A., and Raj, M. Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological science 31*, 1 (2020), 75–87.

[15] Emerson, R. M. *Everyday troubles: The micro-politics of interpersonal conflict*. University of Chicago Press, 2015.

[16] Farokhian, M., Rafe, V., and Veisi, H. Fake news detection using parallel bert deep neural networks. *arXiv preprint arXiv:2204.04793* (2022).

[17] Gaillard, S., Oláh, Z. A., Venmans, S., and Burke, M. Countering the cognitive, linguistic, and psychological underpinnings behind susceptibility to fake news: A review of current literature with special focus on the role of age and digital literacy. *Frontiers in Communication 6* (2021), 661801.

[18] Gruppi, M., Horne, B. D., and Adali, S. Nela-gt-2019: A large multi-labelled news dataset for the study of misinformation in news articles, 2020.

[19] Guo, C., Cao, J., Zhang, X., Shu, K., and Yu, M. Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728* (2019).

[20] Horne, B. D., and Adali, S. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Eleventh international AAAI conference on web and social media* (2017).

[21] Horne, B. D., Dron, W., Khedr, S., and Adali, S. Assessing the News Landscape: A Multi-Module Toolkit for Evaluating the Credibility of News. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018* (2018), pp. 235–238.

[22] Horne, B. D., Nørregaard, J., and Adali, S. Robust fake news detection over time and attack. *ACM Transactions on Intelligent Systems and Technology (TIST) 11*, 1 (2019), 1–23.

[23] Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., and Zhou, M. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021), pp. 754–763.

[24] Huang, Q., Zhou, C., Wu, J., Liu, L., and Wang, B. Deep spatial–temporal structure learning for rumor detection on Twitter. *Neural Computing and Applications*, August (2020).

[25] Jehad, R., and Yousif, S. A. Classification of fake news using multi-layer perceptron. In *AIP Conference Proceedings* (2021), vol. 2334, AIP Publishing LLC, p. 070004.

[26] Jiang, S., Chen, X., Zhang, L., Chen, S., and Liu, H. User-characteristic enhanced model for fake news detection in social media. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8* (2019), Springer, pp. 634–646.

[27] Jwa, H., Oh, D., Park, K., Kang, J. M., and Lim, H. exBAKE: Automatic fake news detection model based on Bidirectional Encoder Representations from Transformers (BERT). *Applied Sciences (Switzerland) 9*, 19 (oct 2019), 4062.

[28] Kaliyar, R. K., Goswami, A., and Narang, P. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia tools and applications 80*, 8 (2021), 11765–11788.

[29] Kaliyar, R. K., Goswami, A., Narang, P., and Sinha, S. Fndnet–a deep convolutional neural network for fake news detection. *Cognitive Systems Research 61* (2020), 32–44.

[30] Ketkar, N. Introduction to keras. In *Deep learning with Python*. Springer, 2017, pp. 97–111.

[31] Kochkina, E., Liakata, M., and Zubiaga, A. All-in-one: Multi-task learning for rumour verification. *arXiv preprint arXiv:1806.03713* (2018).

[32] Liu, C., Wu, X., Yu, M., Li, G., Jiang, J., Huang, W., and Lu, X. A Two-Stage Model Based on BERT for Short Fake News Detection. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 11776 LNAI* (2019), 172–183.

[33] Liu, Y., and Wu, Y.-F. B. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS) 38*, 3 (2020), 1–33.

[34] Loper, E., and Bird, S. Nltk: The natural language toolkit. *arXiv preprint cs/0205028* (2002).

[35] Loshchilov, I., and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[36] Mitra, T., and Gilbert, E. Credbank: A large-scale social media corpus with associated credibility annotations. In *Proceedings of the international AAAI conference on web and social media* (2015), vol. 9, pp. 258–267.

[37] Nakamura, K., Levy, S., and Wang, W. Y. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854* (2019).

[38] Nguyen, V. H., Sugiyama, K., Nakov, P., and Kan, M. Y. FANG: Leveraging Social Context for Fake News Detection Using Graph Representation. *International Conference on Information and Knowledge Management, Proceedings* (2020), 1165–1174.

[39] Nørregaard, J., Horne, B. D., and Adali, S. Nela-gt-2018: A large multi-labelled news dataset for the study of misinformation in news articles. In *Proceedings of the international AAAI conference on web and social media* (2019), vol. 13, pp. 630–638.

[40] O'Shea, K., and Nash, R. An introduction to convolutional neural networks. *arXiv preprint arXiv:1511.08458* (2015).

[41] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research 12* (2011), 2825–2830.

[42] Pennebaker, J. W., Francis, M. E., and Booth, R. J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates 71*, 2001 (2001), 2001.

[43] Pennycook, G., and Rand, D. G. The psychology of fake news. *Trends in cognitive sciences 25*, 5 (2021), 388–402.

[44] Qian, F., Gong, C., Sharma, K., and Liu, Y. Neural user response generator: Fake news detection with collective user intelligence. In *IJCAI International Joint Conference on Artificial Intelligence* (2018), vol. 2018-July, pp. 3834–3840.

[45] Ramchoun, H., Ghanou, Y., Ettaouil, M., and Janati Idrissi, M. A. Multilayer perceptron: Architecture optimization and training.

[46] Raza, S., and Ding, C. Fake news detection based on news content and social contexts: a transformer-based approach. *International Journal of Data Science and Analytics 13*, 4 (2022), 335–362.

[47] Ruchansky, N., Seo, S., and Liu, Y. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), pp. 797–806.

[48] Schmitt, M. F., and Spinosa, E. J. Scalable stream-based recommendations with random walks on incremental graph of sequential interactions with implicit feedback. *User Modeling and User-Adapted Interaction* (2022), 1–31.

[49] Shu, K., Bernard, H. R., and Liu, H. Studying fake news via network analysis: detection and mitigation. *Emerging research challenges and opportunities in computational social network analysis and mining* (2019), 43–65.

[50] Shu, K., Cui, L., Wang, S., Lee, D., and Liu, H. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (2019), pp. 395–405.

[51] Shu, K., Mahudeswaran, D., Wang, S., Lee, D., and Liu, H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data 8*, 3 (2020), 171–188.

[52] Shu, K., Mahudeswaran, D., Wang, S., and Liu, H. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media* (2020), vol. 14, pp. 626–637.

[53] Shu, K., Sliva, A., Wang, S., Tang, J., and Liu, H. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter 19*, 1 (2017), 22–36.

[54] Shu, K., Wang, S., and Liu, H. Beyond news contents: The role of social context for fake news detection. *WSDM 2019 - Proceedings of the 12th ACM International Conference on Web Search and Data Mining 9* (2019), 312–320.

[55] Shu, K., Zheng, G., Li, Y., Mukherjee, S., Awadallah, A. H., Ruston, S., and Liu, H. Early detection of fake news with multi-source weak social supervision. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part III* (2021), Springer, pp. 650–666.

[56] Szczepański, M., Pawlicki, M., Kozik, R., and Choraś, M. New explainability method for bert-based model in fake news detection. *Scientific Reports 11*, 1 (2021), 1–13.

[57] Tandoc Jr, E. C., Lim, Z. W., and Ling, R. Defining "fake news" a typology of scholarly definitions. *Digital journalism 6*, 2 (2018), 137–153.

[58] Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355* (2018).

[59] Verma, P. K., Agrawal, P., Amorim, I., and Prodan, R. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems 8*, 4 (2021), 881–893.

[60] Vijjali, R., Potluri, P., Kumar, S., and Teki, S. Two stage transformer model for covid-19 fake news detection and fact checking. *arXiv preprint arXiv:2011.13253* (2020).

[61] Vo, N., and Lee, K. Learning from fact-checkers: Analysis and generation of fact-checking language. *SIGIR 2019 - Proceedings of the 42nd International ACM*

SIGIR Conference on Research and Development in Information Retrieval (2019), 335–344.

[62] Vosoughi, S., Roy, D., and Aral, S. The spread of true and false news online. *science 359*, 6380 (2018), 1146–1151.

[63] Wang, W. Y. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).

[64] Wu, Y., Zhan, P., Zhang, Y., Wang, L., and Xu, Z. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (2021), pp. 2560–2569.

[65] Yang, S., Shu, K., Wang, S., Gu, R., Wu, F., and Liu, H. Unsupervised fake news detection on social media: A generative approach. In *Proceedings of the AAAI conference on artificial intelligence* (2019), vol. 33, pp. 5644–5651.

[66] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. *Advances in neural information processing systems 32* (2019).

[67] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending against neural fake news. *Neurips* (2020).

[68] Zhang, X., Cao, J., Li, X., Sheng, Q., Zhong, L., and Shu, K. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021* (2021), pp. 3465–3476.

[69] Zhang, X., and Ghorbani, A. A. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management 57*, 2 (2020), 102025.

[70] Zhou, X., Jain, A., Phoha, V. V., and Zafarani, R. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice 1*, 2 (2020), 1–25.

[71] Zhou, X., Wu, J., and Zafarani, R. SAFE: Similarity-Aware Multi-modal Fake News Detection. Tech. rep., 2020.

[72] Zhou, X., and Zafarani, R. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter 21*, 2 (2019), 48–60.

[73] Zhou, X., and Zafarani, R. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. Tech. Rep. 5, 2020.