12-2023

Enhanced Content-Based Fake News Detection Methods with Context-Labeled News Sources

Duncan Arnfield
East Tennessee State University

Enhanced Content-Based Fake News Detection Methods with Context-Labeled News Sources

---

A thesis

presented to

the faculty of the Department of Computer Science

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Information Systems,

Concentration in Cybersecurity Management

---

by

Duncan Arnfield

December 2023

Biju Bajracharya, Chair

Ghaith Husari

Phil Pfeiffer

Keywords: fake news, text classification, natural language processing, detection model

ABSTRACT

Enhanced Content-Based Fake News Detection Methods with Context-Labeled News Sources

by

Duncan Arnfield

This work examined the relative effectiveness of multilayer perceptron, random forest, and multinomial naïve Bayes classifiers, trained using bag of words and term frequency-inverse dense frequency transformations of documents in the Fake News Corpus and Fake and Real News Dataset. The goal of this work was to help meet the formidable challenges posed by proliferation of fake news to society, including the erosion of public trust, disruption of social harmony, and endangerment of lives. This training included the use of context-categorized fake news in an effort to enhance the tools' effectiveness. It was found that term frequency-inverse dense frequency provided more accurate results than bag of words across all evaluation metrics for identifying fake news instances, and that the Fake News Corpus provided much higher result metrics than the Fake and Real News Dataset. In comparison to state-of-the-art methods the models performed as expected.

DEDICATION

This thesis, or rather the work behind it, is dedicated to my friend Samuel Shafer, and my grandmother, Betty Stroup. While I wish it was better in so many ways, I think you'd both be proud of the work I put in despite wanting to roll over and give up several times. Wherever you two are, I miss you both.

ACKNOWLEDGEMENTS

I would like to thank Dr. Bajracharya for going above and beyond to help me figure out what, exactly, a thesis is, and helping me throughout the process. I'd also like to thank Dr. Pfeiffer and Dr. Husari for teaching me over the near decade I've been at ETSU, and for helping me in this process. Finally, I would like to thank my mother and father for giving me so much support over the last two years. I would rather literally not be here without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Overview

Falsity and deception have long been tools of control in wartime. In his *The Art of War*, Sun Tzu wrote that all warfare is based on deception (Tzu 2010). This potential for deception to thwart adversaries is the subject of multiple historical and literary accounts, that include (e.g.) Homer's account of the fall of Troy, the use of foliage to conceal MacDuff's armies during his overthrow of Macbeth, and, in 1943, the British use of the wonderfully named Operation Mincemeat to successfully direct Italian and German defenses away from the intended Allied landing sites of Sicily.

In the 20th and 21st centuries, uses of propaganda and disinformation to promote national interest have become commonplace. One such initiative, the KGB's department of 'dezinformatsiya' that Josef Stalin created in the 1920s to intentionally spread false information at home and abroad, is likely the beginning of what would come to be known as disinformation in the western world (Nunberg 2019).

More recently, the impact of disinformation on civilian populations has come under increased attention. The rapid advancement in technology throughout the 1990s and 2000s has significantly enhanced the ability of governments and other agents to deliver tailored content to specific audiences. While these advancements have brought numerous benefits, they have also raised concerns about the proliferation of misinformation and the challenges associated with distinguishing facts from fake news and disinformation. This increasingly personalized and targeted feed of content has also segmented the Internet, isolating individuals from opposing viewpoints or information disputing their own sources. Consequently, fake news contributes heavily to increasing polarization among populations, presenting a substantial obstacle in

maintaining the integrity of social networks and society at large. Addressing the issue of misinformation is paramount to restoring and maintaining trust in information and a shared understanding of reality. It necessitates concerted efforts to promote media literacy, critical thinking skills, and the widespread dissemination of accurate information.

Fake news can be understood as a variety of semantic attack that is outside the traditional avenues of phishing and scams. Rather than tricking people into disclosing sensitive information or acting on the attackers' behalf, fake news seeks to make its victims believe a specific, likely false or flawed, version of reality to open them to further exploitation. Factors that can convey a sense of false verisimilitude include the news's source, its visual appearance and content style, its method of distribution, and the reader's biases and opinions (confirmation bias): factors, that as with any semantic attack, reduce suspicion by matching the victim's expectations.

1.2. Statement of Research Problem

Information manipulation, a lack of trust in the media, and increasingly isolated information bubbles are critical issues that are increasingly affecting society. One challenge in combating these problems is the evolving and varied nature of fake news, which complicates detection for humans and algorithms alike. This lack of a consistent pattern in fake news creates a need for continuous innovation and adaptation in methods of detection.

Information spreads quickly on social media, especially news that can capture users' attention. In the case of breaking news, much of what is posted in the beginning stages of its propagation is unverified.

1.3. Purpose of the Study

This thesis explored the use of three common models for analyzing structured content in order to classify articles from public domain corpora as reliable or unreliable. These articles'

attributes were first extracted using the Bags of Words and Term-Frequency Inverse Document Frequency (TF-IDF) strategies for feature extraction, then analyzed models trained using Random Forest (RF), Multilayer Perceptron (MLP), and Multinomial Naïve Bayes (MNB) strategies for machine learning (ML). The resulting models' accuracy and reliability were then assessed with regard to their accuracy, precision, recall, and f1-scores. Conclusions were drawn from this assessment that support the presence of common stylistic features in fake news articles across subjects, distinct from more reliable news.

Apart from these results, this research provides insights into factors that contribute to the performance variations observed in false content detection. It offers a comprehensive understanding of the impact of various modeling choices, feature selection strategies, and pre-processing operations on the accuracy of the classification task. This knowledge can serve as a foundation for future research and the development of more robust and accurate systems for false content identification in news media.

2. LITERATURE REVIEW

2.1. Overview

This review of related literature focuses on the nature of fake news, previous work on fake news detection, approaches to classifying fake news, and representative frameworks and typologies for characterizing fake, deceptive, and non-factual news. Also covered will be some of the psychological principles involved in the spread and acceptance of fake news.

2.2. Defining Fake News

The study of fake news encompasses disciplines that include psychology, sociology, and rhetoric. This multidisciplinary effort has produced multiple definitions and theories of what qualifies as fake news, as opposed to simply inaccurate news, and of how fake news relates to other forms of disinformation. A European Union report (Directorate-General for Communications Networks, Content and Technology 2018) rejected the term 'fake news', preferring the more clinical term 'disinformation' and defining disinformation as "all forms of false, inaccurate, or misleading information designed, presented, and promoted to intentionally cause public harm or for profit." This definition identifies malicious or selfish intent as a necessary element in fake news.

As part of their review of 34 articles on fake news, Tandoc et al. (2017) identified satire, news parody, news fabrication, photo manipulation, advertising, and propaganda as cross-cutting categories for characterizing potentially deceptive content. According to Tandoc et al., these categories' content differs based on an article's deceptiveness and the accuracy of its content. News satire and parody were classified as having low intent to deceive, with satire and parody exhibiting high and low levels of accuracy, respectively. By contrast, advertising, manipulation, propaganda, and fabrication were characterized as having high intent to deceive, with

manipulation, propaganda, and fabrication exhibiting high and fabrication exhibiting low levels of accuracy.

Rubin et al. (2016) define fake news as simple deceptive content, whose primary subcategories were serious fabrications, large-scale hoaxes, and humorous fakes. Serious fabrications were published by individuals or organizations in a more formal style and over long periods, while large-scale hoaxes are perpetuated by simply false claims. 'Humorous Fakes' such as parody and satire were included with fake news, and used to stand in for more explicitly malicious or ignorant content in research due to their ready availability from sources such as The Onion (The Onion 2023) and The Beaverton (The Beaverton 2023). However, while the *content* of such publications is either fake or tinged with extreme absurdity, the *intent* is not consistent with the common understanding of fake news. These narratives are not meant to be taken as credible; rather, they are written to make their readers question the feasibility of the scenarios that the articles present.

Zhou and Zafarini (2020) define fake news as "intentionally false news published by a news outlet" (Zhou and Zafarani, A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities 2020). This captures the necessity of intention, factuality, and its authenticity, and differentiates it from other categories of potential deceptive news content.

Table 2-1: Features Assessed in Defining Fake News

Source	Factors Assessed	Labels Derived
Zhou & Zafarani 2020	Authenticity, Intention/Malice, Factual Content, Appearance of News	False news, fake news, disinformation
Tandoc et al. 2017	Intent to Deceive, Information Content	news satire, news parody, news fabrication, photo manipulation, advertising, propaganda
European HLEG on Fake News	Factual Content, Distribution, Intent	Disinformation
Rubin et al. 2016	Intent to Deceive, Scale, Manner of Falsification	Serious Fabrications, Large Scale Hoaxes, Humorous Fakes

For the purpose of this research, "fake news" was defined as "false news created with the intention to be received as reliable fact, containing heavily skewed or fabricated information distributed to influence, harm, or deceive the public."

2.3. Mechanisms of Influence

In order to achieve its purposes, fake news and other disinformation must target traits and biases in those who consume its content. Some of these traits and biases parallel aspects of human nature that are targeted by traditional phishing approaches. A small sample of theories at play in fake news consumption online are listed in Table 2-2 below. Understanding these theories can help in feature development and classification when dealing with fake news, enabling language targeting these biases to be recognized.

Table 2-2: Sample of Theories Involved in Fake News

Theory	**Description**
Dunning-Kreuger Effect (Dunning 2011)	An individual's lack of awareness of their own ignorance leads to their overestimating their own knowledge.
Confirmation Bias (Nickerson 1998)	Individuals will accept information which aligns to their existing viewpoints and opinions more readily.
Conservatism Bias (Equalture n.d.)	Individuals do not significantly revise their opinions when new or contradictory information is presented.
Echo Chamber Effect (Cinelli et al. 2021)	An individual's social group may exist in an information bubble, reiterating and reinforcing belief within itself and becoming resistant to challenge.

2.4. Other Works

The use of ML for classifying the trustworthiness of supposed news has become a popular subject for study. Blackledge et al. (2021) conducted a study on the ISOT and Combined Corpus datasets with a transformer-based approach, achieving accuracy of up to 77.5% without additional processing steps, and up to 80.8% with the addition of opinion-versus-fact identification. Perez-Rosas et al. (2018) hired individuals through Amazon's Mechanical Turk platform to generate fake news stories mimicking the writing style of journalists, as well as collecting 500 articles from online sources as a second dataset. Breaking down the corpora into features of n-grams, punctuation, psycholinguistic features, and syntax assessment, they achieved a combined accuracy score of 76% and an F1-score of 76%. Zhou, Jain, et al. (2020) developed a representative framework for fake news classification and used it to analyze datasets extracted from PolitiFact and BuzzFeed articles. They achieved 89.2% accuracy and an F1-score of 89.2%

against the PolitiFact dataset, and accuracy and F1 both scoring 87.9% against the BuzzFeed dataset.

2.5. Classification Approaches

Historically, sources have relied on human expert judgment to identify and repudiate fake news. This practice has been adopted by classic fact-checking sources, e.g., Snopes (2023) and PolitiFact (2023), and by news groups seeking to dispute the claims of others such as a journal evaluating the claims of a rival news organization. Human-based methods, however, are difficult to scale due to their requiring time, research, and effort to analyze content and are often focused on specific domains such as politics, science, military, or environmental news. They are typically concerned with critiquing falsehood and, where reporting is based in reality, providing a corresponding truth. These efforts to critique and correct falsehood exceed the scope of this research: they include repudiating disinformation as opposed to merely questioning an article's veracity.

Automated tools and processes are needed to detect fake news more quickly and to scale detection processes to match the current volume of information production. All current approaches to classifying fake news with machine tools and computer systems can be classified as either context-based, content-based approaches, or a hybrid of the two.

2.5.1. Context Based Approaches

Context-based approaches, or propagation-based approaches (Monti et al. 2019), assess factors such as a claim's original source (author and/or organization) and its historical trustworthiness, the actors furthering the spread of the claim, the timing of a claim's publication, and engagement with its reached audience. This approach yields attributes that can inform future models and decisions but cannot be applied in the very earliest stages of fake news propagation.

However, integrating the information generated in context-based methods can lead to strong hybrid approaches.

2.5.2. Content Based Approaches

Content-based methods rely on extracting features from a published claim and assessing these features with a pre-trained algorithm to determine their similarity to known unreliable and reliable news. Zhou and Zafarani (op. cit.) proposed that content-based detection can be categorized as style-based, relying on principles such as the Undeutsch hypothesis (Undeutsch 1967). This hypothesis asserts that statements rooted in fact have different linguistic qualities than falsified statements, or knowledge based (fact checking) as seen with groups such as PolitiFact, FactCheck, and Snopes. This manual fact checking can be either derived from expert opinion, or by public input and crowdsourcing.

2.5.3. Representation & Frameworks

The identification of fake news can be enhanced through the development and application of strong typology and models. Well understood methods of identifying key features such as domains of interest (e.g., science, politics, environmental, sports), types of fake news, and common syntactic features that delineate reliable and unreliable information can enhance the performance and reliability of machine tools. A general purpose, domain-insensitive detection model may make a less reliable determination than one that includes domain identification.

Zhou and Razafrani (2020) identified qualities identifying features at the lexicon, syntax, and semantic levels, decomposing each to account for attributes related to clickbait and disinformation. This led to a rich set of possible features to focus on in extracting features from potential sources of fake news.

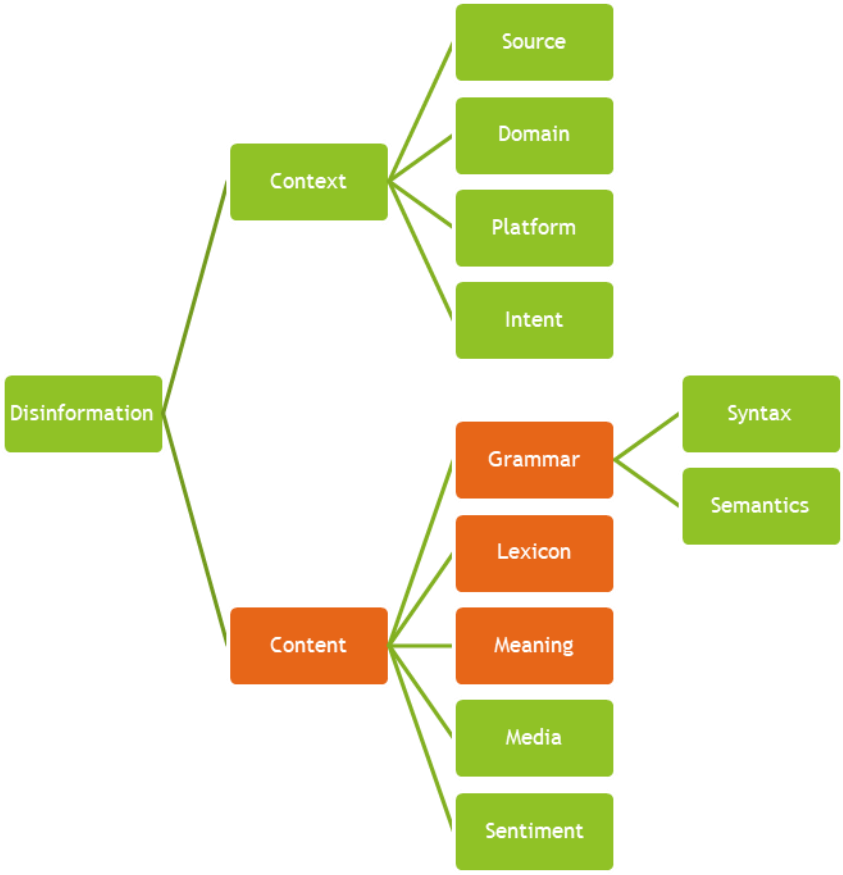

Figure 2-1: An example of a disinformation taxonomy

2.6. Natural Language & Preprocessing

The term *natural language* refers to language that promotes interpersonal communication (Cambridge Dictionary n.d.). Natural language poses well-known challenges to the interpretation of human speech. One is the extent to which local and cultural context can influence the meaning of natural language, resulting in meanings that can sharply diverge from phrases' denotations and confounding interpretation in ways that stymie even human comprehension. For example, the phrase 'bless your heart', which has positive implications throughout much of the United States, is a condescending insult in the Southeastern regions of the country. Another challenge is that natural language has large quantities of bloat. While grammatically necessary for human comprehension, bloat complicates attempts to work with large corpora. The context within which

utterances are formulated, moreover, must be assessed when attempting to determine their tone: e.g., to differentiate between straightforward and sarcastic observations about people, places, and events. As such, any large natural language corpus must be decomposed into features that ML tools can use as inputs.

Stemming and lemmatization are techniques for normalizing text data into their basic forms, suitable for direct interpretation or preprocessing for subsequent analysis. Stemming is the act of truncating words to their roots to regularize all occurrences of each word to a canonical form. Stemming is a rapid and efficient technique that can, however, sacrifice accuracy for simplicity. There is a higher chance that a simple stemming tool will incorrectly truncate a word.

Lemmatization is a refined stemming technique that attempts to map individual words to their context and place in a sentence to provide a more accurate stem. This may be done by first running the text body through a tool to identify each word's part of speech, or through direct application to a text's body. Lemmatization attempts to identify a stem that is contextually accurate for the base word's role in the text, such that a verb and noun version of the same word may be recorded differently.

3. METHODOLOGY

3.1. Overview

This section discusses this work's approach to testing for fake news, including its selection of fake news datasets, the subject corpora's scheme for data classification, the method for transforming these documents for analysis by vectorization tools, and the training and evaluation of fake news. This work can be reduced to four steps:

1) Data preprocessing and preparation

2) Selection of data classes and their specification

3) Final data preparation and vectorization to extract simple text features

4) Predictive classification of prepared data using selected ML tools

Two platforms were used for these tasks: a Windows 10 machine running Python 3.11.3 via Visual Studio Code, and Google Colaboratory running Python 3.11. Both were written and executed using Jupyter Notebooks (.ipynb) file format.

3.2. Datasets

To evaluate the effectiveness of fake news detection methods, researchers commonly assess how these methods classify test documents that have been preclassified as fake or valid. Creating a test dataset from scratch can be time consuming and laborious since each of the dataset's documents needs to be classified before use.

This work used two publicly available fake news datasets that were compiled for use in ML. These datasets were obtained from various platforms using common web crawling methods or APIs. Documents in these datasets whose veracity was not initially characterized were hand-labeled by professional journalists and experts. The number of sources, classifications, and

articles in these datasets vary, as do the standards of data cleaning. Table 3-1 details several fake news study datasets, as well as their labeling methods.

Table 3-1: Fake News Datasets used in Similar Work

Dataset	Volume	News Type	Labels	Accuracy
"**Liar, Liar Pants on Fire**" (Yang Wang 2017)	12800	News Articles & Statements (PolitiFact)	Pants-fire/ False/ Barelytrue/ Half-true/ Mostlytrue/ True	27.40%
Fake and real news dataset (Bisaillon 2017)	44898	News Articles	True/Fake	94%
Fake News (Dedhia 2022)	114061	News Articles	True/Fake	90.19%
BuzzFeed-Webis Fake News Corpus 2016 (Potthast et al. 2017)	1627	News articles (ABC News, CNN, Politico, Addicting Info, Occupy Democrats, The Other 98%, Eagle Rising, Freedom Daily, Right Wing News)	True/False/Mix	75%
Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. (Rubin, Conroy, et al. 2016)	360	News articles (The Onion, The Beaverton, The Toronto Star, The New York Times)	Satirical Online News/ Legitimate Online News	82%
Fake News Corpus (Szpakowski 2020)	9408908	NY Times, WebHose English News Articles	Fake/ Satire/ Bias/ Conspiracy/ State/ Junksci/ Hate/ Clickbait/ Unreliable/ Political/ Reliable	

Fake news datasets in the public domain often consist of less than a hundred thousand samples, limiting their usefulness for developing, training, and evaluating ML algorithms for the broad-based detection of fake news. This research took samples from the Fake News Corpus dataset for training models and initial testing, with the Fake and Real News dataset chosen for validation predictions after the work with the Fake News Corpus was complete.

3.2.1. Fake News Corpus Dataset

Maciej Szapowski's Fake News Corpus is an open-source dataset composed of 9,408,908 news articles and stories from 745 websites. Its content, which was mostly scraped from a curated list of 1001 domains from via the now defunct opensource.co, also contains articles from

the New York Times and WebHose.io English News Articles. Each document is associated with attributes that can include a statement identifier, domain, content, scraped time, inserted time, updated time, title, authors, keywords, meta key, meta descriptions, tags, summary, and one of eleven characteristic types. These types are presented in Table 3.1 and Figure 3.1.

Table 3-2: Listed Fake News Corpus Classification

Tag	**Count**	**Description**
fake	928,083	Sources that entirely fabricate information, disseminate deceptive content, or grossly distort actual news reports
satire	146,080	Sources that use humor, irony, exaggeration, ridicule, and false information to comment on current events.
bias	1,300,444	Sources that come from a particular point of view and may rely on propaganda, decontextualized information, and opinions distorted as facts.
conspiracy	905,981	Sources that are well-known promoters of kooky conspiracy theories.
state	0	Sources in repressive states operating under government sanction.
junksci	144,939	Sources that promote pseudoscience, metaphysics, naturalistic fallacies, and other scientifically dubious claims.
hate	117,374	Sources that actively promote racism, misogyny, homophobia, and other forms of discrimination.
clickbait	292,201	Sources that provide generally credible content, but use exaggerated, misleading, or questionable headlines, social media descriptions, and/or images.
unreliable	319,830	Sources that may be reliable but whose contents require further verification.
political	2,435,471	Sources that provide generally verifiable information in support of certain points of view or political orientations.
reliable	1,920,139	Sources that circulate news and information in a manner consistent with traditional and ethical practices in journalism (Remember: even credible sources may rely on clickbait-style headlines or occasionally make mistakes. A healthy news diet consists of multiple sources of information).

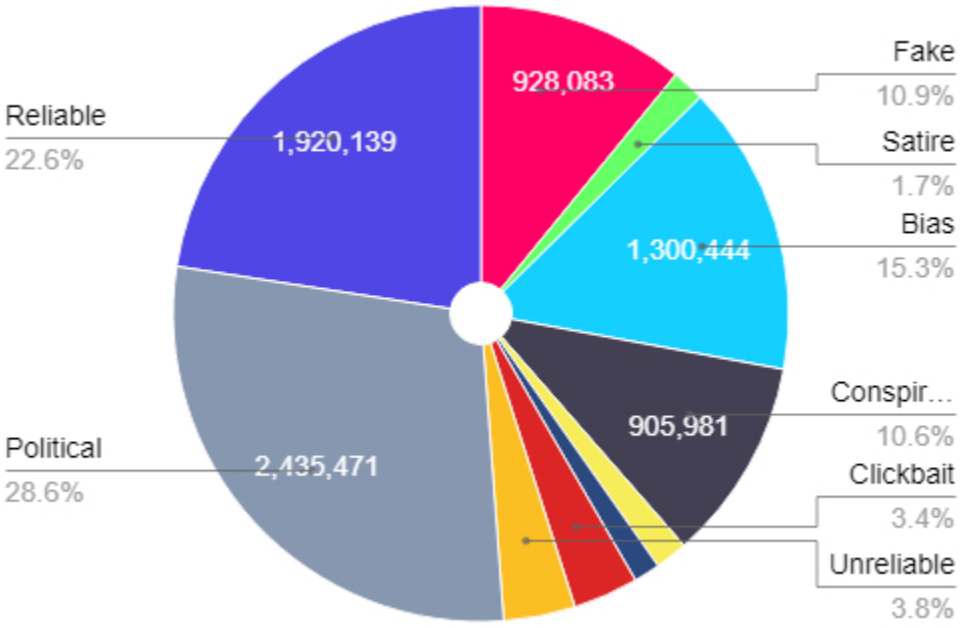

Figure 3-1: Fake News Corpus Original Data Classification

3.2.2. Fake and Real News Dataset

Clement Bisallon's Fake and Real News Dataset is a smaller dataset that identifies articles based on subject rather than source. It also assigns a binary True/False rating to each article. The data is comparable in content to the Fake News Corpus and consists of extracted article bodies. Identical data handling steps are possible, though its storage as two independent .csv files requires an additional step to create a single data source with a classification column. This column was labeled 'type', with the data from the 'true' file classed as 'reliable' and that from the 'fake' file classed as 'unreliable'.

3.3. Methodology

3.3.1. Data Pre-processing

Both corpora exhibit irregularities. Some entries are missing their 'type' class label. Some have erroneous entries, such as a timestamp in a non-time field. Some have repetitive records

that copy some or all content from another record. Finally, some include fragments of HTML in their article bodies.

The Fake News Corpus is organized by source and context rather than content. Each of its documents is categorized by its source according to a process determined by opensource.co's definitions of its eleven types of articles. These characterizations are an imperfect match for common categorizations of fake news and misinformation. They are, however, applied consistently across all of the corpus' sites and articles. This consistency of labeling allows the corpus' data to be regularized to obtain fewer, meaningful categories, and to select a smaller subset of information for processing. For this study, the corpus' fake and bias classes were merged under a new class of 'unreliable', reducing the problem to one of binary classification.

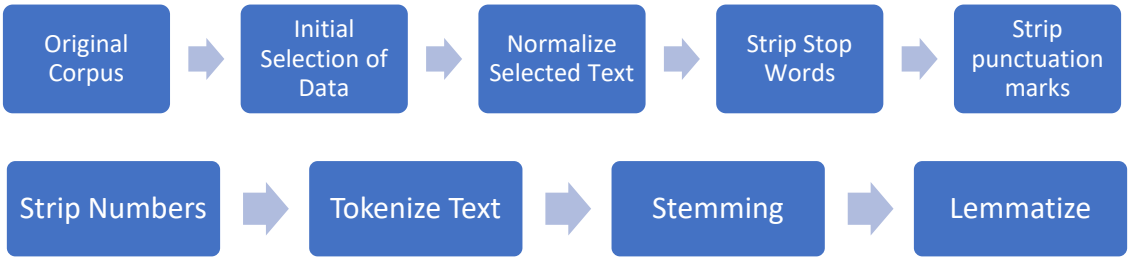

Figure 3-2: Workflow for Data Preprocessing

3.4. Data Feature Extraction

3.4.1. Bag of Words

The bag of words technique reduces a text to a list of that text's terms, together with a measure of each term's frequency. This approach ignores semantics such as a document's word order, context, and grammatical structure, focusing solely on developing a list of word counts. When used with ML models and a large corpus of information, the approach allows the model to link the rate of occurrence of certain terms to a text's outcome or class.

3.4.2. Term Frequency – Inverse Dense Frequency

Term Frequency-Inverse Document Frequency (TF-IDF) is a refinement of the bag of words technique applied across multiple documents. TF-IDF rates each of a document's terms according to its frequency, then adjusts that weight based on each term's presence across all documents. This adjustment presumes that a corpus' commonly repeated words are less important for conveying its articles' main points than the corpus' less frequently repeated words, and thus should be weighed less heavily in classification. Unlike the basic bag of words approach, TF-IDF reduces noise from stop words like articles and stock phrases.

3.5. Supervised Learning

Supervised learning, along with reinforcement and unsupervised learning, is one of three basic types of ML algorithms. Supervised learning algorithms like MLP, RF Classification, and MNB are trained using prepared datasets. The goal of this training is to produce a model that maps inputs to their proper characterizations.

3.5.1. Multi-Layer Perceptron (MLP)

MLP uses a neural network that includes one or more *hidden* layers: layers of *neurons* (nodes) between its input and output layers. In the hidden layers, each neuron updates the previous layer's nodes' activation weight values. An MLP network can learn non-linear models as well as real-time models. MLP validation accuracy depends on how its weights are initialized; it is sensitive to feature scaling, and hyperparameter tuning is also required.

3.5.2. Random Forest (RF) Classification

An RF Classifier is an ensemble-based method used for classification, anomaly detection, and regression problems. It uses a collection of decision trees, each of which is initialized using a bootstrap sample drawn from the training set. While constructing a tree, internal nodes are split

along input features to form a top-down flow of decisions that result in an output, or leaf, node. With large amounts of data some decision sequences, or branches, may have low usage, and be pruned. The RF model attempts to resolve issues of overfitting common with single decision trees by treating the result obtained from a majority of its decision trees as its final result.

3.5.3. Multinomial Naïve Bayes (MNB)

MNB is a type of Naïve Bayes model that is used to classify textual data that conforms to a multinomial distribution. It derives its probabilities by linking the number of a feature's occurrences to classes (scikit-learn n.d.). The model's advantages are its speed and well understood nature. Some of its shortcomings can be addressed using TF-IDF-generated features instead of bag of words preparations. MNB is a parametric learning method: it attempts to reduce its input features to a predefined number, regardless of count. This can impact its accuracy.

3.5.4. Loss Functions

A supervised learning model's accuracy during training is determined using loss functions, which measure the difference between a model's actual outcomes and its intended outcomes. These values are fed back into the model, which then attempts to minimize the loss value of subsequent runs in order to improve the accuracy of its predictions.

3.6. Implementation

Due to limitations in hardware and processing power, the research used a 3% sample of the Fake News Corpus to optimize hyperparameters for the MLP and RFC models, and a 15% sample of the Fake News Corpus to train MLP, RF, and MNB classifiers for evaluation. The samples were drawn with the dataframe sample function, initialized with random seeds of 69 and 42, respectively. Both dataset samples were split into training data and testing data, with 85% of

each sample being used to train and 15% to test. The entire Fake and Real News Dataset was then used for final testing and evaluation.

These split datasets were fed into two vectorizers from sklearn, a CountVectorizer. sklearn returned a bag of words characterization of the datasets, along with a TfidfVectorizer. This vectorizer was then used to obtain a TF-IDF matrix of characteristic weights and features. Each vectorizer was configured to produce a maximum of 20,000 features present between 1% and 85% of the input documents. This choice of ranges was motivated by the diverse nature of the articles' sources and platforms: i.e., any term that occurs in over 85% of documents would likely be too common to be of significance, while any that occurs in fewer than 1% would likely be too domain-specific. The data input to the vectorizers was identical: i.e., the same variables were fed into the CountVectorizer and TfidfVectorizer. This ensured that each feature extraction model would use an identical starting point. The training data was used to generate and fit the vectorizers' vocabularies prior to being transformed into counts of vocabulary features, while the testing data was transformed to counts of features without generating a new vocabulary.

The 3% dataset, along with the RFC and MLP models, were fed a GridSearchCV to identify optimal hyperparameters for initializing the final machine models. The hyperparameters used to tune these models are listed in Table 3-3 along with the output of recommended hyperparameters. While not factored into the hyperparameter recommendations of the GridSearchCV in Table 3-3, the time each combination of hyperparameters consumed was also used to select the most appropriate options. This is due to several higher setting attempts taking upwards of three hours to complete, with several resulting in crashes.

After confirming their suitability, these parameters were used to initialize the three models. The models were then trained with the training sample of the Fake News Corpus and

used to classify the 15% testing sample from the same corpus. Once these initial test runs completed, the Fake and Real News Dataset was vectorized via the transform functions and evaluated in the same way.

Table 3-3: Details of Hyperparameter Evaluation

Model	**Hyperparameters**	**Best Outcome**	**Hyperparameters Used**
RFC	Max depth: [14,20,25,30,35,40,45,50] N_estimators: [30,35,40,45,50]	Max depth: 50 N_estimators: 50	Max_depth: 45 N_estimators: 45
MLP	Hidden_layer_sizes: [10,10,10], [20,20,20], [40,40,40], [100,100,100]	[100,100,100]	[100,100,100]

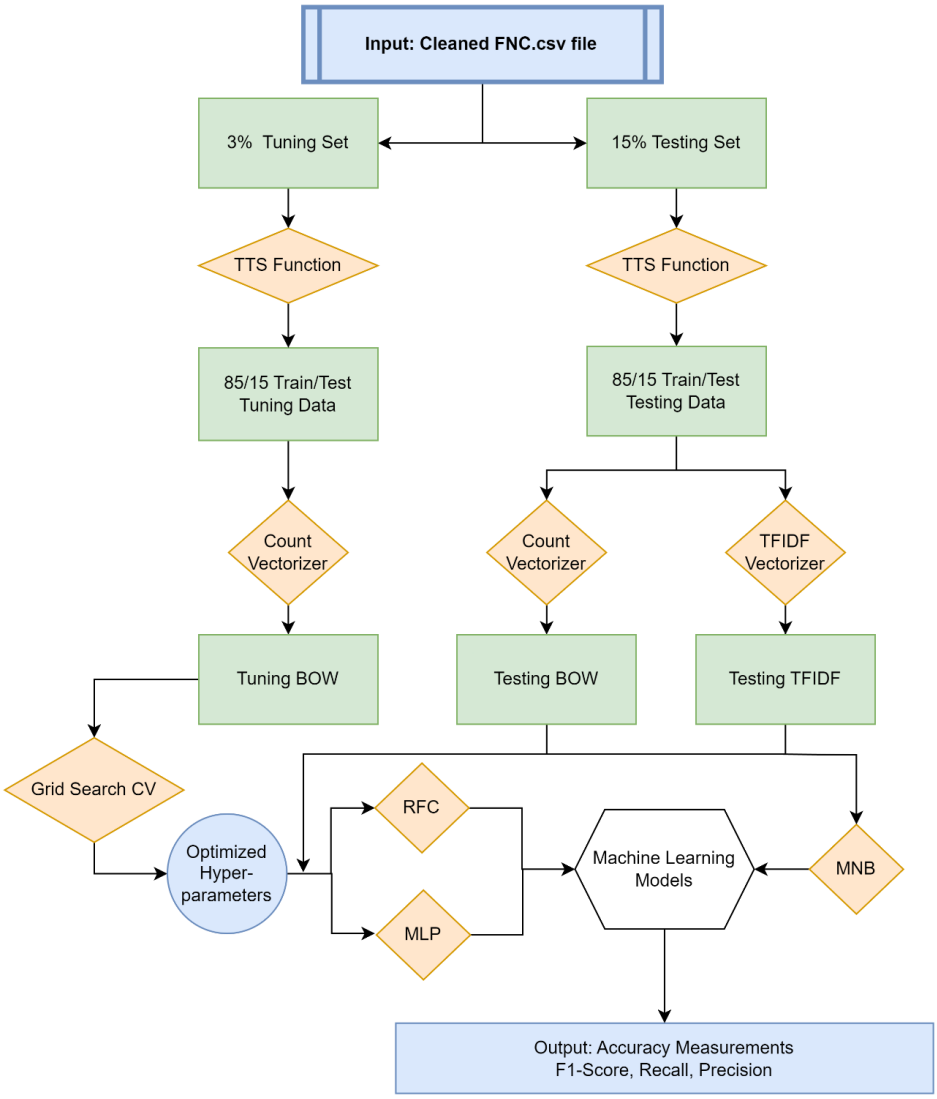

Figure 3-3: Workflow for the Fake News Corpus

4. DATA ANALYSIS & RESULTS

4.1. Evaluation Metrics

The six trained models—bag-of-words- and TF-IDF-trained instances of MLP, RF, and MBR—were used to classify the test datasets. The results of these trials were then used to evaluate the models' quality, based on the trials' precision, recall, and F1-scores.

Precision is the number of correctly predicted positive outcomes (TP – true positives) against the total number of predicted positive outcomes (TP + FP – the number of incorrectly predicted positive outcomes). Values range from 0 (worst) to 1 (ideal).

$$Precision = \frac{|TP|}{|TP| + |FP|} \tag{1}$$

Recall is the ratio of TP to the total number of positive instances in the data sample: i.e., TP + FN, where FN is the number of samples that were incorrectly predicted to be negative outcomes. Values range from 0 (worst) to 1 (ideal).

$$Recall = \frac{|TP|}{|TP| + |FN|} \tag{2}$$

F1-score is the harmonic mean of precision and recall. Being a harmonic mean, it is sensitive to low values, and thus will require both to perform well for the F1-score to increase.

$$f1 = 2 \times \frac{(Precision \times recall)}{(Precision + Recall)} \tag{3}$$

Additionally, a generalized accuracy score was used to compare predictions from each dataset with state-of-the-art results. After completing its run, each machine model generated a report containing a confusion matrix and set of numeric average scores for the prediction. The confusion matrix, a four-way characterization of these models' binary outputs, partitions a model's characterizations into TPs, FPs, FNs, and True Negatives (TNs) – TN being correct exclusions of an instance's membership from the class of interest.

4.2. Results

For the Fake News Corpus, the trained models yielded higher-than-expected evaluation metrics with simple bag of words and TF-IDF feature inputs across all models. The MNB model registered a precision of 86.09%, recall of 90.45%, and F1-score of 88.22% on detecting unreliable news bodies from bag of words features. The other scores from the Fake News Corpus sampled data are listed in Table 4-1, out to six significant figures.

Table 4-1: Unreliable News Detection from Fake News Corpus Sample

Model Used	Feature Preparation	Precision	Recall	F1-Score
Multinomial Naïve Bayes	Bag of Words	0.860949	0.904503	0.882189
	TF-IDF	0.890107	0.872649	0.881292
Random Forest Classifier	Bag of Words	0.914528	0.978127	0.945259
	TF-IDF	0.912075	0.978105	0.943937
Multilayer Perceptron	Bag of Words	0.965514	0.955729	0.960597
	TF-IDF	0.958356	0.959743	0.959049

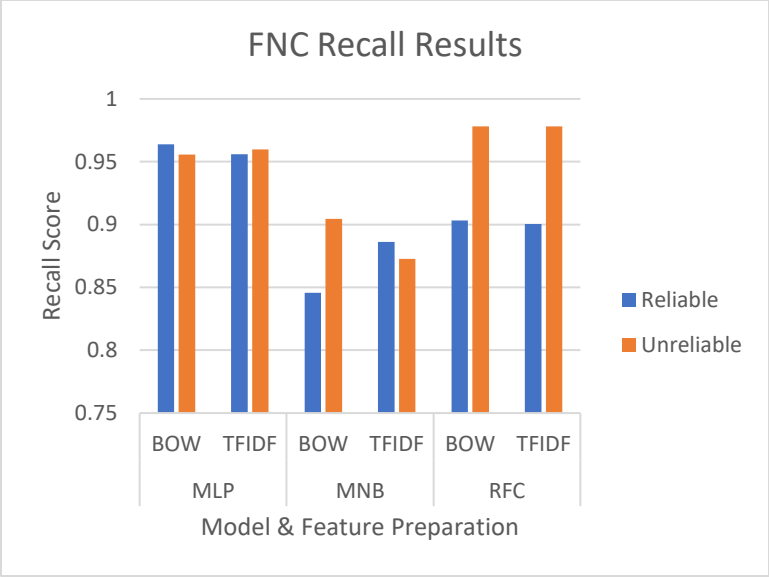

Figure 4-1: Recall Results of Fake News Corpus

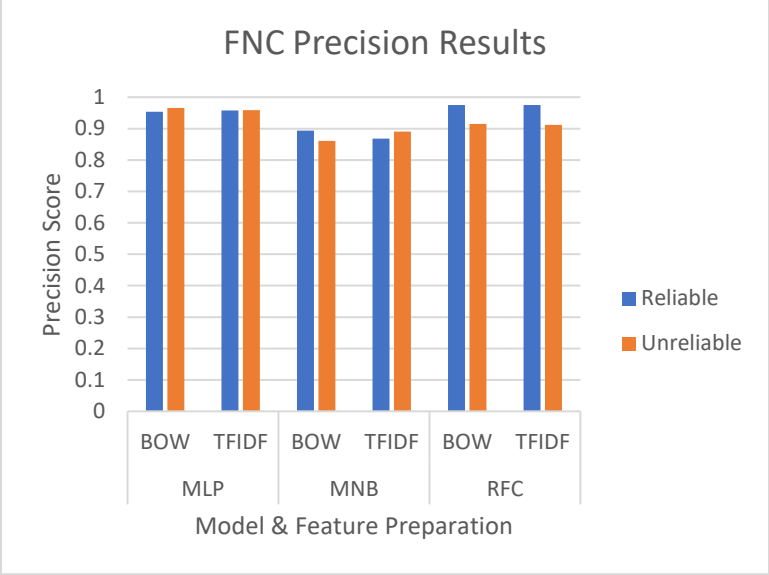

Figure 4-2: Precision Results of Fake News Corpus

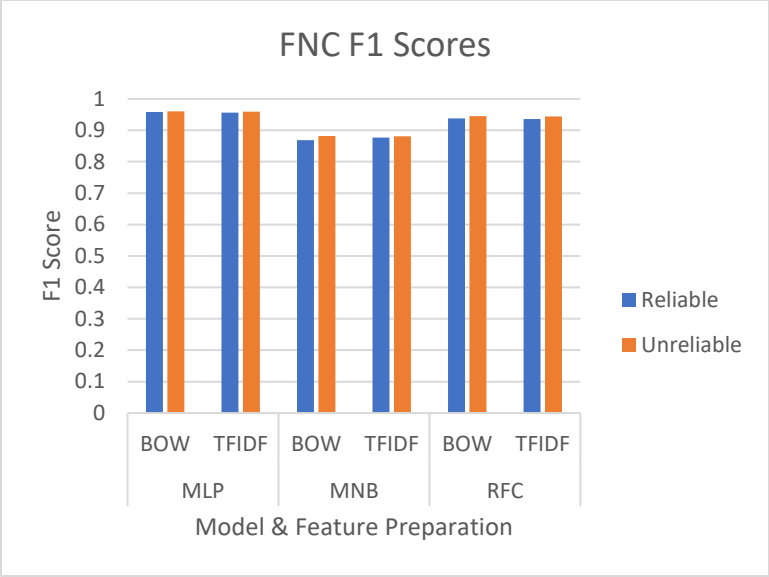

Figure 4-3: F1-Score Results of Fake News Corpus

The Fake News Corpus sample set's results were abnormally high, well above those of the chosen state-of-the-art methods used as comparison. The second prepared dataset from the - Fake and Real News Dataset was used to run a second classification trial for validation. The results from this were much lower than the initial testing set, performing much lower than the state-of-the-art methods and in line with initial expectations, as shown in Figures 4-4 through 4-6, and in Table 4-2.

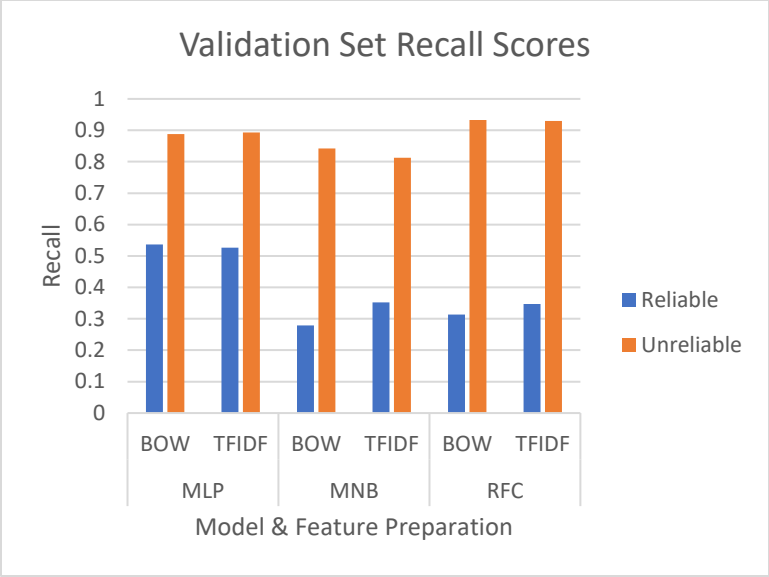

Figure 4-4: Recall Scores of Fake and Real News Dataset

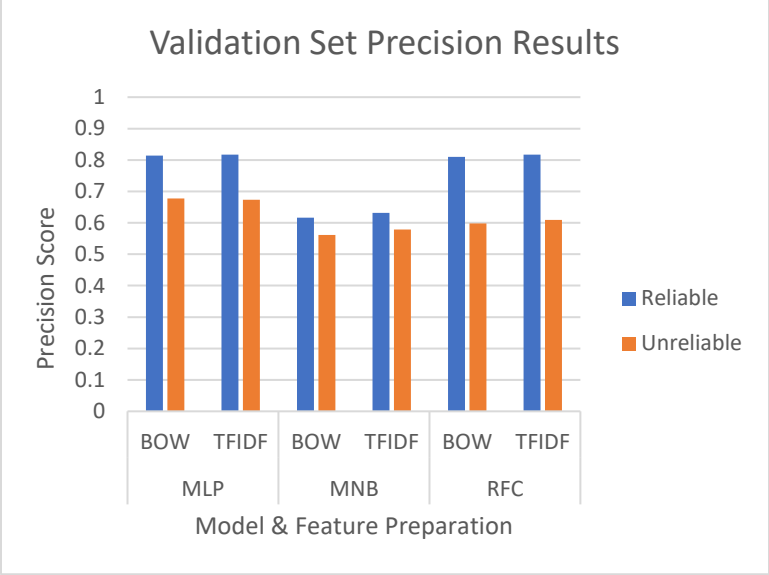

Figure 4-5: Precision Scores of Fake and Real News Dataset

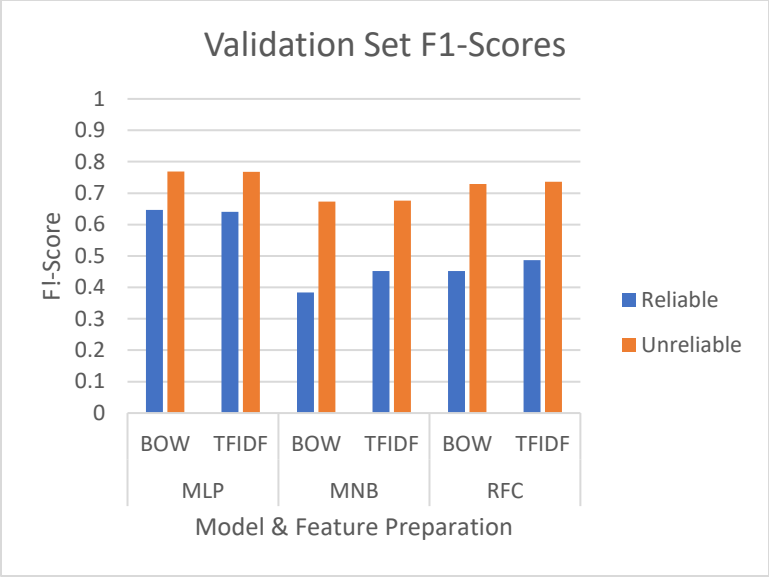

Figure 4-6: F1-Scores of Fake and Real News Dataset

Table 4-2: Unreliable News Detection Scores from Fake and Real News

Model Used	**Feature Preparation**	**Precision**	**Recall**	**F1-Score**
Multinomial Naïve Bayes	BOW	0.561302	0.841702	0.673482
	TFIDF	0.57911	0.812742	0.676318
Random Forest Classifier	BOW	0.677637	0.88825	0.76878
	TFIDF	0.673769	0.89285	0.767991
Multilayer Perceptron	BOW	0.598291	0.933095	0.729094
	TFIDF	0.609423	0.929305	0.736114

The extremely high scores generated by the Fake News Corpus testing set indicate an erroneous result. The source of this error was not determined. One possible source of error could have been an incorrect vectorization or sampling process, resulting in the reuse of articles in training and testing sets. A second could have been a training process that produced severe overfitting. In either case the error did not appear when evaluating the Fake and Real News

Dataset, resulting in expected scores compared to the state-of-the-art methods. As such, these results are accepted.

5. DISCUSSION AND CONCLUSION

5.1. Discussion

With the data drawn from the Fake News Corpus, the dataset prepared via TF-IDF vectorization was almost a direct inversion of the results obtained with the bag of words features. This, along with the results being well outside expectations, makes them highly suspect. The results from the independent Fake and Real News dataset were much more in line with expectations. For identifying fake news, vectorizing content to produce TF-IDF-weighted features produces better results than a bag of words feature set. The MLP and RF classifiers were comparable in performance, with both substantially better than the MNB model. The training time of both models increased exponentially with increases to the parameters beyond the selected, with minimal improvements in performance. Notably, while the source data differed from that used by Perez-Rosas (2018), the MLP and RF classifiers achieved similar results.

The quality of these results was severely limited by a lack of processing power, which limited the number of features that these analyses used. Attempting to extract more computationally intensive features such as POS tags, named entity recognition labels, and n-grams above one all resulted in crashes of the system or session. This was due to insufficient RAM and likely insufficiently optimized code. This limitation could likely have been addressed with more powerful resources and/or packages that support out-of-band computation. Selecting smaller datasets for use in training would be advisable in future iterations.

5.2. Conclusions

This thesis has demonstrated some of the problems in using purely context-determined, content-and-domain blind datasets to train content-oriented fake news detection tools for general purpose fake news detection. While with proper hyperparameter optimization and data treatment

reasonable results can be obtained with off the shelf tools and methods, the resulting models are unlikely to produce reliable determinations when accounting for the nuance and complexity of the problem space.

This thesis supports that the Undeutsch hypothesis applies when taking specific varieties of source-labeled unreliable news and extrapolating them to serve as a basis for general detection of largely unrelated forms of fake news. It substantiates work from previous studies. It also identifies and documents some of the complications that may arise when working with large volumes of data, and from potential tool misconfiguration.

5.3. Future Research

Future work could explore the accuracy of source classification datasets using more domain- and style-specific approaches to detecting fake news. This could further the understanding of features of specific types of fake news while refining source classification methods. Achieving a relatively neutral and unbiased mechanism to label sources such as organizations and authors, though likely to prove highly contentious, could prove useful for unbiased fake news repudiation and detection.

Future work should also account for the downsides of working with large corpus data on limited hardware resources and select working datasets to match available tools and timelines.

One possible approach for future testing would be to use context-classified sources of fake news to train machine detection tools to detect domain-specific trends. This could be done by selecting a specific format of fake news (e.g., Tweet, opinion piece, website article, news journal) related to a specific domain and topic and using similar text preprocessing methods to identify stories in that domain for further assessment: i.e.. constructing a two-phase process to determine a domain tag as well as likely classification as reliable or unreliable, and then using

that information for more detailed analysis. This approach would neatly reduce the problem space and allow for greater refinement in both tasks.

BIBLIOGRAPHY

Anaconda, Inc. *Dask.* 2014-2018. https://docs.dask.org/en/stable/.

Bisaillon, Clement. "Fake and real news dataset." *Kaggle.* 2017. https://www.kaggle.com/datasets/clmentbisaillon/fake-and-real-news-dataset.

Blackledge, Ciara, and Amir Atapour-Abarghouei. "Transforming Fake News: Robust Generalisable News Classification Using Transformers." *IEEE International Conference on Big Data.* IEEE, 2021.

Cambridge Dictionary. *NATURAL LANGUAGE.* Cambridge University. n.d. https://dictionary.cambridge.org/dictionary/english/natural-language (accessed June 2023).

Cinelli, Matteo, Gianmarco D. F. Morales, Alessandro Galeazzi, and Michele Starnini. "The echo chamber effect on social media." *PNAS*, 2021.

Colliander, Jonas. "“This is fake news”: Investigating the role of conformity to other users’ views when commenting on and spreading disinformation in social media." *Computers in Human Behavior* 97 (2019): 202-215.

de Beer, Dylan, and Machdel Matthee. "Approaches to Identify Fake News: A Systematic Literature Review." *Integrated Science in Digital Age 2020.* Kep, Cambodia, 2020.

Dedhia, Ronik. "Fake News." *Kaggle.* 2022. https://www.kaggle.com/datasets/ronikdedhia/fake-news.

Directorate-General for Communications Networks, Content and Technology. *A multi-dimensional approach to disinformation.* Government, Publications Office, 2018, 44.

Dunning, David. "Chapter five - The Dunning–Kruger Effect: On Being Ignorant of One's Own Ignorance." *Advances in Experimental Social Psychology* 44 (2011): 247-296.

Equalture. *What is Conservatism Bias - Definition & Examples in Recruitment.* n.d. https://www.equalture.com/bias-overview/conservatism-bias/ (accessed July 25, 2023).

George, Jordana, Natalie Gerhart, and Russell Torres. "Uncovering the Truth about Fake News: A Research Model Grounded in Multi-Disciplinary Literature." *Journal of Management Information Systems* 38, no. 4 (2021): 1067-1094.

IBM. "What is a Decision Tree." n.d. https://www.ibm.com/topics/decision-trees (accessed June 2023).

Internet Archive. *OpenSources.* March 6, 2019. https://web.archive.org/web/20190306114817/http://www.opensources.co/.

M. Imran, A. Qadir, and S. U. Khan. "Fake News: A Cybersecurity Threat." *IEEE Security & Privacy Magazine*, 2018.

Monti, Federico, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. "Fake News Detection on Social Media using Geometric Deep Learning." *arxiv.org.* February 10, 2019. https://arxiv.org/abs/1902.06673.

Nickerson, Raymond S. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of General Psychology*, 1998: 175-220.

Nunberg, Geoff. *'Disinformation' is the Word of the Year - And a Sign of What's To Come.* National Public Radio. December 30, 2019. https://www.npr.org/2019/12/30/790144099/disinformation-is-the-word-of-the-year-and-a-sign-of-what-s-to-come (accessed June 2023).

Perez-Rosas, Veronica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic Detection of Fake News." *Proceedings of the 27th International Conference on Computational Linguistics.* Santa Fe: Association for Computational Linguistics, 2018.

Politifact. *PolitiFact.* Politifact. 2023. https://www.politifact.com/article/2018/feb/12/principles-truth-o-meter-politifacts-methodology-i/.

Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Benedorff, and Benno Stein. "A Stylometric Inquiry into Hyperpartisan and Fake News." *ArXiv.* February 18, 2017. https://arxiv.org/abs/1702.05638.

Raza, Shaina, and Chen Ding. "Fake news detection based on news content and social contexts: a transformer-based approach." *International Journal of Data Science and Analytics* 13 (2022): 335-362.

Rubin, Victoria L., Niall J. Conroy, Yimin Chen, and Sarah Cornwell. "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News." *Proceedings of the Second Workshop on Computational Approaches to Deception Detection.* San Diego: Association for Computational Linguistics, 2016. 7-17.

Rubin, Victoria L., Yimin Chen, and Nadia K. Conroy. "Deception detection for news: Three types of fakes." *Proceedings of the Association for Information Science and Technology* 52, no. 1 (2016): 1-4.

scikit-learn. *1.9. Naive Bayes.* n.d. https://scikit-learn.org/stable/modules/naive_bayes.html.

—. "sklearn.ensemble.RandomForestClassifier." n.d. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html (accessed June 2023).

Snopes. *Snopes FAQs.* Snopes. 2023. https://www.snopes.com/faqs/.

Szpakowski, Maciej. "Fake News Corpus." *github.* January 24, 2020. https://github.com/several27/FakeNewsCorpus.

Tandoc Jr., Edson C, Zheng Wei Lim, and Richard Ling. "Defining 'Fake News'." *Digital Journalism* 6, no. 2 (2017): 137-153.

The Beaverton. *The Beaverton.* July 2023. https://www.thebeaverton.com/.

The Onion. July 2023. https://www.theonion.com/.

Tzu, Sun. In *The Art of War*. Chichester: Capstone Publishing, 2010.

Undeutsch, Udo. "Forensische Psychologie." In *Handbuch Der Psychologie*. Göttingen Hogrefe 1967, 1967.

Yang Wang, William. *"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection.* May 1, 2017.

Zhang, Xichen, and Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion." *Information Processing & Management* 57, no. 2 (2020).

Zhou, Xinyi, and Reza Zafarani. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities." *ACM Computing Surveys* 53, no. 5 (2020): 1-40.

Zhou, Xinyi, Atishay Jain, Vir V. Phoha, and Reza Zafarani. "Fake News Early Detection: A Theory-driven Model." *Digi. Threat.* 1, no. 2 (2020).

VITA

DUNCAN ARNFIELD

Education: M.S. Information Systems, East Tennessee State University, Johnson City, Tennessee, 2023

B.S. Information Technology, East Tennessee State University, Johnson City, Tennessee, 2020