THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF INFORMATION SCIENCES & TECHNOLOGY

Using COVID 19 Healthcare Misinformation to Develop a System to Detect Future Fake News

ALLISON DENENBERG
SPRING 2022

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Data Sciences
with honors in Data Sciences

Reviewed and approved* by the following:

Suhang Wang
Assistant Professor of Information Sciences & Technology
Thesis Supervisor

Marc Friedenberg
Assistant Professor of Information Sciences & Technology
Honors Adviser

* Electronic approvals are on file.

ABSTRACT

The purpose of this experimental study is to create a machine learning algorithm that can accurately classify tweets as fake or real news. For this experiment, I utilized CoAID (Covid-19 healthcare misinformation dataset). CoAID is a diverse COVID-19 healthcare misinformation dataset that includes fake news on websites and social platforms, along with users' social engagement from these news articles. Using this dataset enabled the gathering of tweets using Twitter's API. This dataset was processed and used to train a machine learning algorithm to automatically detect whether the tweets were real or fake news. The prevalence of misinformation posing as authentic news on social media platforms, especially Twitter, is a critical global issue. The goal of this study is to conduct further research in regards to detecting fake news on social media; specifically a machine learning model that can be trained to properly detect misinformation on social media. Due to incomplete features in the dataset, the model created for the user and social context attributes did not ultimately end up producing conclusive results. However, this research can serve as a foundation for further experiments which will minimize the amount of fake content people are being exposed to and therefore reduce harmful effects within years to come.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to acknowledge the people who helped me throughout my time as a Schreyers scholar. To Marc Friedenberg, my honors advisor and professor, thank you for your belief in my ability to succeed in the Schreyers Honors College. Without your recommendation and motivation to apply, I wouldn't be participating in this research process and enhancing my knowledge with real-world applications of data science. To Suhang Wang, my thesis supervisor and professor, I am grateful for your support, guidance and advice in this research process. To Tianxiang Zhao, a PhD student, thank you for assisting me throughout my research and helping me succeed.

Chapter 1

Introduction

People have increasingly used social media platforms to read the news in order to become accurately informed. Forbes magazine reports that as of 2018, “nearly 64.5 percent receive breaking news from Facebook, Twitter, YouTube, Snapchat and Instagram instead of traditional media” [1]. There is indication this number continues to rise. However, it can be challenging to distinguish what is authentic and accurate digital content.

The target of the fake news depends on the purpose of the dishonest information, but everyone can become a victim. Fake news content is easily generated and quickly spread making it a global issue with potentially harmful implications. The term and concept of fake news grew during the 2016 election in the United States, but ramifications have since been revealed in areas other than in politics. A recent example of this advance is misinformation regarding the COVID-19 pandemic. Statista reports, “A study in March 2020 showed that almost 80 percent of consumers in the United States reported having seen news about the coronavirus which proved to be completely false” [2]. This illustrates the extent to which the circulation of fake news can achieve, as well as the prevalence of untrustworthy sources seeking to capitalize on the public's need for news and information in a crisis situation. The spread of fake news has become a global challenge and threat to modern democracies. The purpose of this study is to apply machine learning models that can detect fake news and therefore reduce its harmful effects.

Interpreting the accuracy of news content is complex and human naivety makes us not particularly adept at consistently differentiating truth from misinformation [2]. An understanding

of the political or social context is often necessary to catch the subtle distinctions of truth. This is challenging for language processing algorithms as well as humans [3]. Sensational headlines will often be shared due to either the lack of readers' fact checking or the lack of reading more than just the headline. This results with fake news being more likely to spread than factual information.

Linguistic features extracted from the news content are a key identifying feature of fake news. The words and punctuation which compose the headlines and content include the opinions, emotions, attitudes and sentiment that the news creators wish to convey [4]. The author's expressions, including stance and sentiment polarity, contain the ideas, feelings and views that they intend to project on to the readers. Expressing powerful sentiment - positive or negative - will often increase the persuasiveness of writing. Therefore, it makes sense that this type of language would be present in fake news. However, it can be difficult to detect fake news strictly based on its language content. User and social context info can also be indicative of misinformation. My research will attempt to identify specific social context features and user analysis information that correlates to fake and real news stories, and then to train an algorithm to classify the news based on these features. Characteristics include user profiles, account integrity, responses, and networks including information spread by nonhuman bots as well as echo chambers. Spatiotemporal information can include user locations and timestamps of user engagement. My study has the potential to advance the detection of false or biased news stories and reduce their spread.

Chapter 2

Literature Review

The surge to research and develop data mining techniques as well as an algorithm for the detection of fake news is evolving. The authors of the reviewed studies all discuss fake news detection on social media, and their consensus reinforces the concept that the authenticity of news can be implied from both textual content as well as information from user engagement.

In [5], the authors Shu, Wang, and Liu modeled a tri-relationship (TriFN) among publishers, news pieces, and users in order to detect fake news. To effectively detect fake news, one must explore auxiliary information from different perspectives. This paper discussed five major components to the proposed framework: news contents embedding component, a user embedding component, a user-news interaction embedding component, a publisher-news relation embedding component, and a semi-supervised classification component. The independent variable for this research was the dataset and the dependent variable was the features extracted and the algorithms used. The experimental settings included feature extraction methods in comparison to their new framework in the algorithms. One only extracted features from news contents, another only constructed features from social context, and the last one considered both news content and social context. TriFN scored better than all other feature extraction methods when used in the algorithms and improved the classification results significantly. This research paper gave an example of approaches that can detect fake news. This research was very informative and it was interesting to read the unique variation of detecting fake news. This research paper broadened my knowledge about the realm of possibilities when looking to detect

fake news. The sampling strategy in [5] consisted of two pre-existing datasets, Buzzfeed and PolitiFact.

In [6], the authors Shu, Sliva, Wang, Tang, & Liu provide an overview of fake news detection and discuss the possible research directions. Fake news is now being found on social media through malicious accounts in addition to the echo chamber effect. Motivated by not having a set definition of fake news, this research focuses on characterization and detection. In the characterization phase, the authors introduced the basic concepts and principles of fake news in both traditional media and social media. In the detection phase, the authors reviewed existing fake news detection approaches from a data mining perspective, including feature extraction and model construction. We need to include supplementary information but exploiting this is challenging and produces "data that is big, incomplete, unstructured and noisy" [6].

This research paper enabled me to further my knowledge about all of the varied aspects of fake news detection. Throughout this paper, the authors discuss datasets that can be used as the dependent variable and that the features extracted can be used as the dependent variables in future research. The sampling strategy in Fake News Detection on Social Media: A Data Mining Perspective consisted of four pre-existing datasets, BuzzFeedNews, LIAR, BS Detector, and CREDBANK [6]. This paper is very informative in the background knowledge of fake news detection. The discussion and credible points of view will be helpful to me and other researchers that can use these details in our own fake news detection research

In [7], the authors Tacchini, Ballarin, Della Vedova, Moret, and de Alfaro propose a way to identify hoaxes on Social Network Sites (SNS) based on the users who interacted with them rather than their content. When looking at the users they divided them into three categories: those who liked hoax posts only, those who liked non-hoax posts only, and those who liked at least one

post belonging to a hoax page, and one belonging to a non-hoax page. The dependent variable used in their research was their Facebook dataset and the independent variable was the type of classification used, either logistic regression or harmonic boolean label crowdsourcing. These different classification models on the dataset will be used to determine whether the user interacts with posts that are hoaxes or non-hoaxes. The high accuracy score achieved by both logistic regression and the harmonic BLC algorithm confirmed the authors' basic hypothesis that the set of users that interacts with news posts in social network sites can be used to predict whether posts are hoaxes. This research paper gave a new approach to detecting fake news by specifically looking at the users and *not* the content. This focus on the users allowed me to gain insight to another perspective of accurately detecting a post that is a hoax.

The sampling strategy in [7] uses a dataset that consists of all the public posts and the posts' likes of a list of selected Facebook pages during the second semester of 2016: from July 1st, 2016 to Dec. 31st, 2016. They collected the data by means of the Facebook Graph API1 on Jan. 27th, 2017. The resulting dataset is composed of 15,500 posts from 32 pages (14 conspiracy and 18 scientific), with more than 2,300,00 likes by 900,000+ users. Among posts, 8,923 (57.6%) are hoaxes and 6,577 (42.4%) non-hoaxes [7].

In [4], the authors Guo, Cao, Zhang, Shu, and Yu created an emotion-based fake news detection framework (EFN). EFN uses a deep neural network to learn representations from publisher emotion, social emotion, and content simultaneously for fake news detection. "Emotion EFN has components: the content module mines the information from the publisher, including semantic and emotions information in news contents; the comment module captures semantic and emotion information from users; and the fake news prediction component fuses the high-level features from both news content and user comments and predicts fake news" [4]. This

researcher is strictly based on content: words and punctuation that indicate emotion and sentiment. The dependent variable was the dataset the authors constructed and the independent variable are the different fake news detection methods such as DTC, ML-GRU, Basic-GRU, and HSA-BLSTM. The EFN model achieved an overall accuracy of 87.2% and 87% of F1-Score on their datasets, which outperforms all the baseline models [4]. Using gates the authors observed the emotion of sentiment words such as ”scary” and punctuation such as ”!” and ”?” gain higher weight than semantic modality, while the sentiment words’ emotion modality obtain more attention than others words [4]. This paper proposed an end-to-end emotion-based fake news detection framework, EFN, which incorporates the publisher emotion and the social network user emotion in fake news detection simultaneously. This paper looked at fake news detection in a way I had heard of before but never associated it with fake news. This model was clearly definitive based on the accuracy it achieved. However, I think it could be difficult to exhibit an emotion from text. I will research sentiment analysis techniques and with more understanding and research on this way of detecting fake news could progress an enhanced version of the standard fake news detection processes. The sampling strategy in [4] constructed the dataset on Sina Weibo. This dataset includes 7880 pieces of fake news and 7907 pieces of real news, with nearly 160k comments. The fake news is collected from the official rumor debunking system of Weibo4 , and the real news is gathered from NewsVerify5, a real-time news certification system on Weibo which contains large-scale verified truth posts on Weibo [4].

In [3], the authors Monti, Frasca, Eynard, Mannion, and Bronstein used geometric deep learning. Geometric deep learning naturally deals with heterogeneous data (such as user demographic and activity, social network structure, news propagation and content), thus carrying the potential of being a unifying framework for content, social context, and propagation based

approaches. The independent variable used in this paper is the Twitter dataset the authors created and the dependent variable is the setting of fake news detection. The settings used were URL-wise classification and cascade-wise classification. URL-wise classification predicts the true/fake label of a URL containing a news story from all the Twitter cascades it generated. The cascade-wise classification assumes to be given only one cascade arising from a URL and attempts to predict the label associated with that URL. Using the area under the ROC curve, the authors were able to aggregate the accuracy. This was indicative that the neural network learning features that are useful for fake news classification and the results were highly accurate. A key advantage of using a deep learning approach (as opposed to 'handcrafted' features) is its ability to automatically learn task-specific features from the data [7]. The choice of geometric deep learning in this case is motivated by the graph-structured nature of the data. This research paper focuses on deep learning, which is something I have never used. From the reading, I was about to see specific fake news detection that I had not seen in other related works. The authors also incorporated an aging factor and tested data separated in time [3]. This experimental design was very interesting and I definitely want to look more into this type of research in comparison to the other classification algorithms used in other works.

The sampling strategy in [3] used four steps to create their dataset. First, they gathered the overall list of fact-checking articles from such archives and, for simplicity, discarded claims with ambiguous labels, such as 'mixed' or 'partially true/false'. Second, for each of the filtered articles they identified potentially related URLs referenced by the fact-checkers, filtering out all of the ones not mentioned at least once on Twitter. Third, the trained human annotators were employed to determine whether the web pages associated with the collected URLs were matching or denying the claim, or were unrelated to that claim. Finally, they retrieved Twitter

data related to the propagation of news associated with a particular URL. The news diffusion tree produced by a source tweet referenced a URL and all of its retweets. For each URL, they searched for all the related cascades and enriched their Twitter-based characterization (i.e. users and tweet data) by drawing edges among users according to Twitter's social network. This revised a pre-existing dataset [7].

The researchers of the above experiments all discuss fake news detection on social media, and their consensus reinforces the concept that the authenticity of news can be implied from both textual content as well as information from user engagement. Similar to the experiments reviewed, I will use a dataset from fact checking websites and use both content and supplemental data to distinguish fake news from real news. However, my data set does not include consistent user data. Additionally, my experiment will attempt to identify fake news based on classification models rather than deep neural networks or geometric deep learning.

Chapter 3

The Characteristics of Fake News

Online social media has changed the way people communicate, connect, share information, and stay informed. A consequence of this trend is the presence of fake news containing content specifically intended to mislead and confuse social media readers. Fake news can influence opinions and decisions, and also change the way that people interact with truthful news. None of this has positive outcomes. Therefore, the challenge of detecting and identifying fake news is crucial in order to reduce its potentially harmful effects.

A review of literature has revealed so many components which may be examined in order to determine whether news is actually misinformation. These features can be organized into three categories: content based, author or user profile based, and social context based as illustrated below.

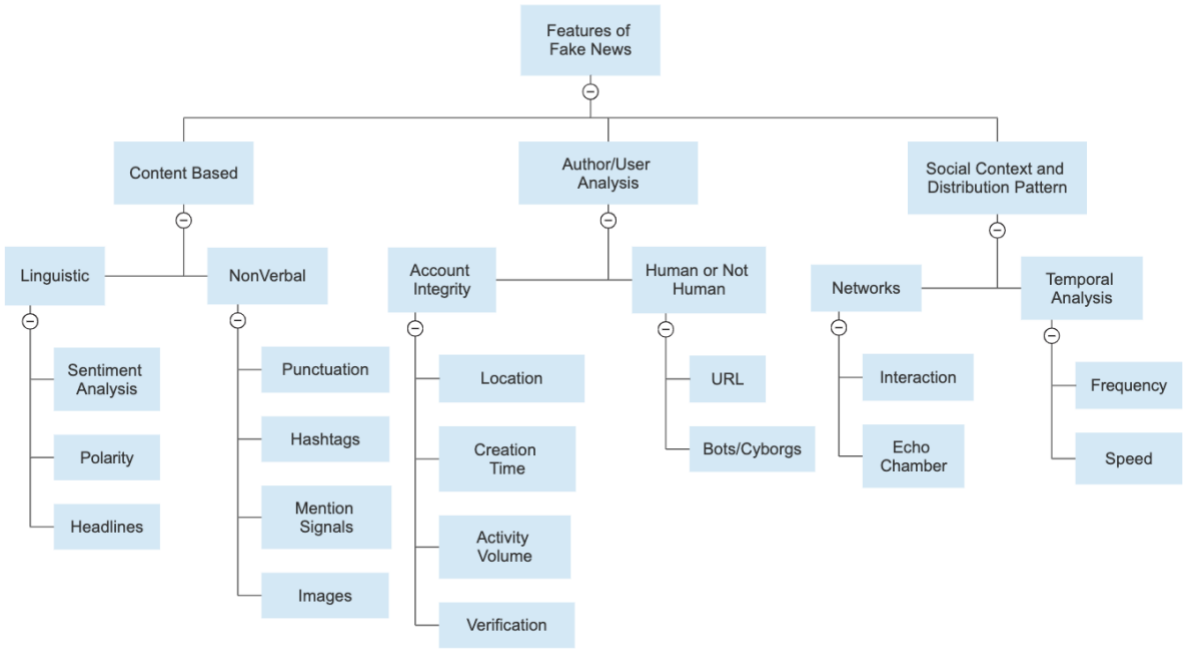

Figure 1. Characteristics to be examined in order to identify online fake news

Verification of truth in the news is not only dependent on the content. The creator as well as the social context can also be influential factors [5].

Content based characteristics include linguistic features as well as nonverbal attributes.

- **Linguistic features:** The specific words chosen to demonstrate the opinions, emotions, attitudes and sentiments that the creators want to convey can be an identifying feature of fake news. Sentiment analysis is a technique used to determine whether words can be classified as positive, neutral or negative. Words can be analyzed and used to infer the sentiment of the writer and what they intend to relay to the readers. Expressing extremely strong positive or negative points of view creates polarity and plays a key role in misinformation spreading on social media [4]. Tweets generate sentiment data by expressing feelings, ideas, and views through word choice. Malicious accounts can exaggerate the facts and mislead legitimate users. Such sentiment can be indicative of

human vs social bot accounts. Shu, Wang, and Liu wrote, “There is a direct connection between linguistic based indicators, such as negative sentiment, inflammatory language and unique words, and fake news” [5]. Fake news headlines also tend to be outrageous and attention grabbing in order to receive more clicks. They often contain exclamation points or question marks [4]. The subsequent news content may be unrelated to the headline or even conflict with the information in the headlines.

- **Nonverbal content** can also be used to signal fake news. Examples of nonverbal content features include excessive punctuation (!!!!,???), hashtags, emojis, a mention signal, pictures and videos. These elements can be indicators of suspicious content [4].

The author and the user analysis information can also be important to identify false news, as content analysis alone cannot always establish whether or not news is untruthful.

- **Account integrity**: Malicious social media accounts have different characteristics when compared to legitimate users. Examining user profile information including: the geographic location of the account, the account creation time, if the account is verified or not, the number of friends and followers, and how active/how many posts and tweets the account has are all worthy of review [8]. A large volume of replies and mentions might indicate a questionable account.
- **Real humans**: Real humans are the source of fake news. Social bots and cyborgs may be automated carriers and spreaders, but they are programmed to disperse falsehoods ultimately created by humans trying to affect beliefs and behavior of others. Information can also be shared by legitimate users, creating an echo chamber, which propagates its spread. Because the internet is anonymous, online users do not need to take responsibility for what they post, share and comment. Therefore, a source based approach focuses on

the origin of the author. Features that might spark suspicion include an unknown or abnormal domain or an unusual, “suspicious token” in the URL [9].

- **Non-humans**: Social bots and cyborgs are computer algorithms that interact with humans on social media. They are designed and programmed by humans to appear human and automatically produce and distribute content [10]. They participate with the social community and can be designed to distribute misinformation, becoming an efficient platform for the rapid and easy spreading of fake news. The numbers of friends and followers are good features for differentiating malicious accounts from legitimate human users. “The number of followers of a legitimate user is often close to its friends. However, social bots usually have much more friends than followers” [11].

Social context information can also provide relevant information on differentiating truth from misinformation online.

- **Networks**: Social context analysis examines the interactions between users, how the news is shared, and how quickly it spreads through the network. If a group continuously shares and forwards some information, an echo chamber is created which increases the influence of the message. This amplification is called the “echo chamber effect” and can enhance the spread of fake news by giving it an exaggerated presence online [6].
- **Temporal analysis** reveals the timing patterns of posts on an account including the frequency and speed of replying and sharing. Because they operate through timed programs, there are specific time periods during the day when social bots and cyborgs function most. This can flag suspicious accounts because real human interaction tends to have much more random temporal features [10].

Chapter 4

Hypothesis

Humans are vulnerable news consumers and the rapid increase of fake news has caused disruptions in society and weakened the trustability of news in general. Researchers are working to develop models to attempt to detect fake news on social media using data and machine learning techniques in order to diminish its negative effects. Because it is written to intentionally mislead, linguistics alone cannot be a reliable indication of false news content [5]. Previous work I have reviewed all reinforced the concept that the authenticity of news can also be implied from the context of its user engagement. Readers express their emotions or opinions toward fake news through social media posts, including skeptical opinions and sensational reactions. These features can be important signals to the study of fake news and disinformation [4]. Social context and user interaction will be integrated with a previous dataset with ground truths for both fake and real news in order to prove a correlation. Additionally, the speed and location of user profiles will attempt to explain how fake news spreads over time [7]. In order to look at the fake news detection performance, I will evaluate the accuracy and F1 score for each classifier. My intended outcome is to create a classification model that can recognize fake news content, and to learn how social context and spatiotemporal features are relevant in detection.

Machine learning models can be used to determine whether news is authentic or misinformation based on content, but classification can also be supplemented by social context and user engagement features.

Chapter 5

Dataset

The dataset used in my experiment, CoAID (Covid-19 Healthcare Misinformation Dataset), is a diverse COVID-19 healthcare misinformation dataset. CoAID includes both factual details and misinformation related to COVID-19, including news articles and claims. Table 1 summarizes the types of features in the CoAID dataset.

Table 1. Feature descriptions of CoAID [12]

Type	Features
Information on websites	ID, Fact-checking URL, Information URLs, Title, Article title, Content, Abstract, Publish date, Keywords
User engagement: Tweets	ID, Tweet ID
User engagement: Replies	ID, Tweet ID, Reply ID
Social Platform Posts	ID, Fact-checking URL, Post URLs, Title

Facts and Misinformation of Content on Websites

The truthful news articles used in this dataset have been taken from nine reliable media outlets and have been cross-checked and determined to be reliable. The true news articles include: Healthline, ScienceDaily, National Institutes of Health (NIH), Medical News Today (MNT), Mayo Clinic, Cleveland Clinic, WebMD, World Health Organization (WHO), and Centers for Disease Control and Prevention (CDC). In the CoAID dataset there were 4,532 real news articles.

The fake news articles came from several fact-checking websites, such as: LeadStores, PolitiFact, FactCheck.org, CheckYourFact, AFP Fact Check, and Health Feedback. In the CoAID dataset there were 925 fake news articles.

Claims were gathered from the WHO official website, WHO official Twitter account, and MNT. These were then separated into true and fake claims. In the CoAID dataset, there were 490 real claims and 28 fake claims [12].

User Engagement

Using a Twitter API, the CoAID dataset gathered user engagement data of both factual and misinformation.

Using the titles of the news articles as search queries and the specific dates of the COVID-19 pandemic, tweets and their replies were able to be gathered. The user engagement contained features such as user ID, tweets, and replies as shown in Table 2. There were 141,845 tweets about real news articles and 10,443 tweets about fake news articles. Following these tweets, there were 115,116 replies under real articles and 7,477 replies under fake news articles. There were 8,103 claim tweets for real news articles and 484 claim tweets for fake news articles. Following these tweets, there were 12,658 replies under the real claims and 626 replies under the fake claims.

Table 2. Statistics of CoAID

	Real	Fake	Total
News	4,532	925	5,457
Claims	490	28	968

News Tweets	141,845	10,443	152,288
Claim Tweets	8,103	484	8,587
News Replies	115,116	7,477	122,593
Claim Replies	12,658	626	13,284

The two master datasets I will be using in my experiment are real news articles and tweets and fake news articles and tweets. Each of these datasets combines the news data and their corresponding tweets. The real news dataset has 10,249 rows and 41 columns. The columns are type, fact_check_url, news_url, title, newstitle, content, abstract, publish_date, favorite_count, retweeted, possibly, sensitive, lang, extended_entities, quoted_status_id, quoted_status_id_str, quoted_status, withheld_in_countries, and ID.

Chapter 6

Methodology

Due to the absence of a regulatory authority on social media, the goal of this experiment was to create a machine learning classifier that could accurately identify if a tweet was real or fake. If the classifier was able to correctly classify tweets as real or fake, the expectation would be that it could be used to reduce the presence of these tweets from social media platforms and moderate the spread of fake news.

Feature Extraction

For feature extraction, my research used the features from the two news articles and tweets master datasets to be trained on multiple classifier algorithms. When deciding which features to use, I looked at data to see which features were consistent throughout both datasets. Most of the tweet data besides the text was incomplete and filled in with “NaN”. Additionally, the corresponding replies did not seem relevant enough to put in a classifier model due to the inconsistent amount of replies for both real and fake news. The features from the articles themselves were consistent except for the published date. Most of these values were “NaN”. The features I would be using for my experiment were the titles of the articles, the content of the articles, and the text of the tweets. Using the words from each of these features, I was able to create a list of all of the words in the real news articles and tweets, as well as a list of all the words in the fake news articles and tweets. With these two lists of words, I created word clouds for each real and fake news to see the most common words included. This would allow me to

visualize the common words in both lists and see if there were any commonalities or sentiment differences between the two. Linguistics have the ability to persuade people to think in a certain way, elicit emotion and bias, educate incorrectly, exaggerate, and create confusion.

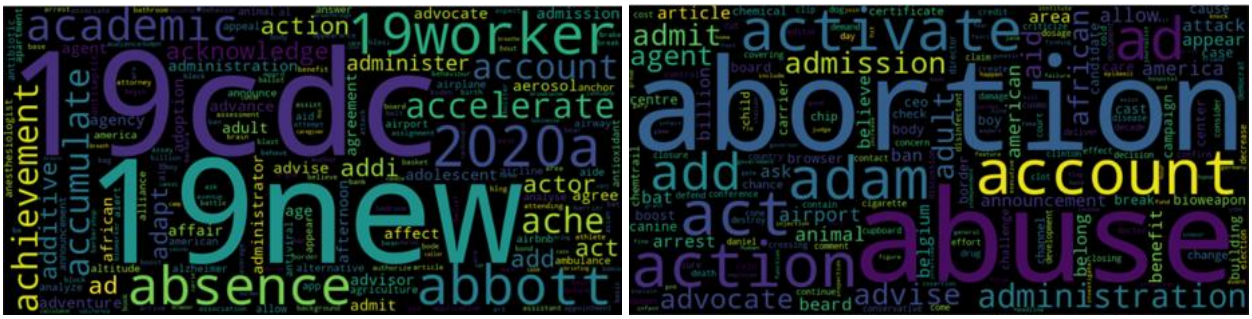

Figure 2. Word clouds of news titles, contents, & tweets for both real & fake news

As shown in the figure above, the words in each of the word clouds are very different and have contrasting sentiments. This divergence in wording between the datasets solidified my decision to further investigate linguistic features. Word sentiment was principal for my classification model, and therefore I wanted to determine the polarity scores of the datasets. The polarity could indicate whether positive or negative sentiment corresponds with the truthfulness (or lack thereof) within news that is being published.

Term Frequency Inverse Document Frequency (TF-IDF)

Term Frequency Inverse Document Frequency (TF-IDF) is a calculation of how relevant a word in a series is to the text. “The meaning increases proportionally to the number of times in the text a word appears – compensated by word frequency in the dataset” [13]. Using this array, I will be able to test multiple splits of training and testing data, oversampling techniques, and classification algorithms.

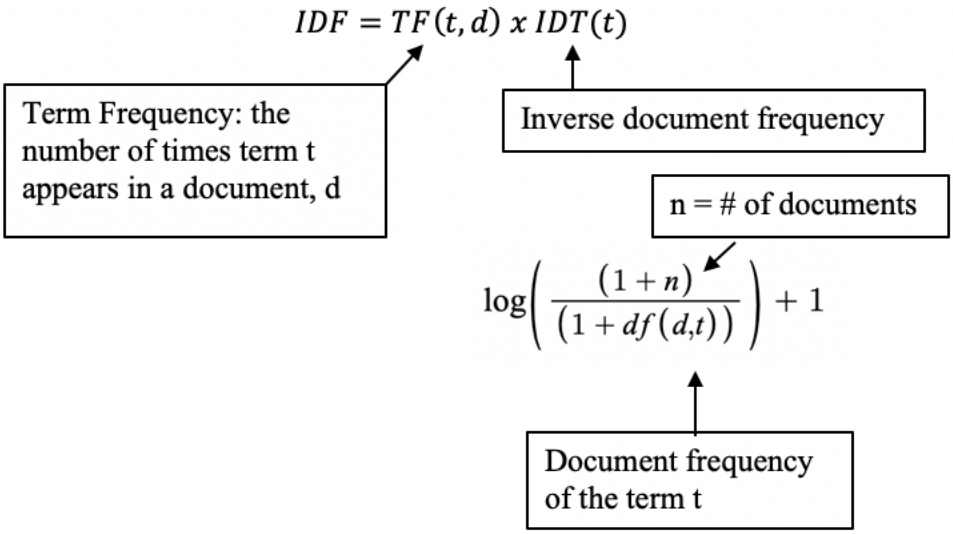

Figure 3. TF-IDF Formula [18]

Imbalanced Data

The COVID-19 classification dataset has an imbalance. When classifying news as real or fake, the data will be trained that needs to be balanced. The original COVID-19 dataset has more real news articles and tweets than it does fake. Working with a dataset that is this imbalanced will result in poor performance in fake news data, the minority class. One way to handle this imbalance dataset is to oversample the fake news class which is done by augmenting the data. This augmentation to the data is called Synthetic Minority Oversampling Technique (SMOTE).

"SMOTE is one way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model" [19].

"Classification is a technique that is used to categorize data into distinct classes" [14]. For this experiment I will be using multiple classifier algorithms to determine which results in the highest accuracy and F1 score.

Classification Models

Naive Bayes (NB) is a classifier that can calculate the probability of an event. It has a very low computation cost and it can efficiently work on a large dataset [15]. Naive Bayes proceeds with the following steps. First, it calculates the prior probability for given class labels. Next, it will find the likelihood probability with each attribute for each class. Following this, it will put these values in Bayes Formula and calculate posterior probability. Finally, we can see which class has a higher probability, given the input belongs to the higher probability class. It is not only a simple approach but also a fast and accurate method for prediction [16].

Logistic Regression is a classification algorithm that can be used to find the probability of an event success or event failure. A logistic regression model takes a linear equation as an input and uses a logistic function and log odds to perform a binary classification task.

Support Vector Machines (SVM) is a model that plots each data point in an n-dimensional space. It performs classification by finding the hyper-plane that separates the data. This hyperplane is adjusted to best split the data into two perfectly distinct classes. The plane is determined by margin. Margin measures the distance between data points, therefore maximizing the margin distance provides classification of future data points with more confidence [17].

The K-Nearest Neighbor Algorithm is a popular supervised machine learning algorithm that can solve both classification and regression problems. KNN uses distance measures to find k closest neighbors to a new, unlabeled data point to make a prediction. This algorithm will attempt to infer the new data point's class by looking at the classes of the majority of its k neighbors.

Decision trees are used to predict the value of a variable by producing if-then statements from data features. In other words, the final output is broken down by multiple inputs/stages. Each feature in the input becomes a node. The final decision becomes a leaf. The model observes

the decisions based on entropy. Entropy in this case means uncertainty. The goal is for entropy to be 0 so that the decision is 100% going to occur.

The final method stems directly from Decision Trees; Random Forest. Random Forest uses a similar structure to DT, but instead includes randomness. By including random results, the model can group outcomes based on the most likely responses called bags. It also tosses out outliers in the data, which help to avoid skewed results.

Chapter 7

Experimental Results

The results of this experiment show the different ways of training a classification model to improve the accuracy and f1 score. In order to analyze which was the best classification method for this dataset, I attempted a variety of methods to determine which would give me the best results.

Before implementing the datasets into the models, I calculated the polarity scores for the real and fake datasets. Unfortunately, the scores for each dataset were classified as neutral. This meant the polarity would not necessarily help in our classification methods.

Table 3. Polarity Scores

	Negative	Neutral	Positive
Real News Words	**0.05892962632885779**	**0.857879026354337**	**0.0831924692851252**
Fake News Words	**0.09052654232424621**	**0.8226499282639882**	**0.0868350071735995**

I was able to create multiple training datasets for the classification algorithms. To begin, I used a 70/30 training and testing split with no SMOTE oversampling or tweets included. Then 70/30 using SMOTE oversampling and no tweets. And then 70/30 using SMOTE oversampling and including the tweets. Following this, I did the 75/25 training and testing split with no SMOTE oversampling or tweets included. Then 75/25 using SMOTE oversampling and no tweets. And then 75/25 using SMOTE oversampling and including the tweets. I ran these classifiers multiple times to achieve a mean F1 and mean accuracy score for all the variations I tested.

In order to see the best results there were multiple variations of the models. I used multiple types of splits for the training and testing data, SMOTE for oversampling, and looking at the news articles with/without the tweets. In order to compare the results, I will be looking at the accuracy and F1 scores. Accuracy determines how many they got right in the testing set over the total number of test cases. Both the overall accuracy of the models and the accuracy by the class will be looked at. The F1 score is a measure that looks at the harmonic mean of the precision and recall of a model. This method has the following formula as shown in the figure below:

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Figure 4. F1 score formula [20]

When comparing the results for all of these methods, I was looking to see which classifier method could achieve the highest accuracy score. Overall, the 75/25 split using SMOTE oversampling and including the tweets resulted in the highest accuracy and f1 scores. Specifically, the SVM algorithm had the highest mean f1 score and accuracy score. SVM resulted in a 98.5% F1 score and 99.1% accuracy score.

Table 4. Performance without SMOTE

	Mean F1	Mean AUC
Naive Bayes	**0.6174142480211082**	**0.8439094650205762**
Logistic Regression	**0.7628865979381444**	**0.9318267154792077**
SVM	**0.8604651162790699**	**0.9412986613325127**
KNN	**0.7536945812807883**	**0.9002168578439764**
Decision Tree	**0.7876106194690267**	**0.8767738920782515**

Random Forest	**0.8195121951219513**	**0.93832826265492**

Table 5. Performance 70/30 Split – Title, Content

	Mean F1	Mean AUC
Naive Bayes	**0.5422386015**	**0.7448126829**
Logistic Regression	**0.8678621754**	**0.9151088454**
SVM	**0.873622122**	**0.929657034**
KNN	**0.4191114352**	**0.6318867666**
Decision Tree	**0.8108262461**	**0.8809746973**
Random Forest	**0.8641662965**	**0.9147064456**

Table 6. Performance 70/30 Split – Title, Content, Tweet

	Mean F1	Mean AUC
Naive Bayes	**0.9224590137**	**0.9786792062**
Logistic Regression	**0.9768274038**	**0.9820926854**
SVM	**0.984831234**	**0.9905195066**
KNN	**0.5913680602**	**0.7098748536**
Decision Tree	**0.9410483658**	**0.9519380836**
Random Forest	**0.9755638881**	**0.9791171589**

Table 7. Performance 75/25 Split – Title, Content

	Mean F1	Mean AUC
Naive Bayes	**0.564669937**	**0.7742293173**
Logistic Regression	**0.8672693023**	**0.9126879377**
SVM	**0.8763765311**	**0.9297601579**
KNN	**0.438007455**	**0.6394784528**
Decision Tree	**0.8118052783**	**0.8862331346**
Random Forest	**0.8597525355**	**0.9164609959**

Table 8. Performance 75/25 Split – Title, Content, Tweet

	Mean F1	Mean AUC
Naive Bayes	**0.9326420836**	**0.9787235416**
Logistic Regression	**0.9802310833**	**0.9851393073**
SVM	**0.9847835565**	**0.9912047626**
KNN	**0.6312250524**	**0.730558396**
Decision Tree	**0.9501612624**	**0.9595534983**
Random Forest	**0.9750821202**	**0.9810988774**

In order to see which words were the most important in the classification models, I was able to use the Random Forest model. Using the 75/25 split and SMOTE on the training data, I

fitted the Random Forest model and used the “feature_importances_” to output an array of the important features. Following this I was able to sort the importance and display the top 10 important words when classifying tweets as real or fake. These words included: “online”, “service”, “website”, “live”, “protect”, “security”, “facebook”, “19”, “covid”, “https”.

Chapter 8

Discussion

In the future, there are user engagement and social context features that should be further explored to further the research and detection of fake news. This is challenging because "auxiliary information is big, incomplete, unstructured, and noisy" [6].
Regardless, spatiotemporal features should be considered. Unfortunately with this dataset, the publish date and time was incomplete. This made it difficult to verify an echo chamber effect of news being posted. The geographic location and URL of the post, which can also indicate a malicious account, was also incomplete in the dataset. Additionally, looking at the timing of the corresponding tweets and replies could have helped in determining whether they were real or fake. The speed of responsiveness and the geographic location of the user account can imply fake news created by bots rather than humans [10].

The goal of this study was to be able to classify tweets as real or fake news. Applied machine learning models that can properly detect fake news can help reduce its harmful effects. The results display high accuracy when using the titles, content, and tweets of the news. In the future, incorporating additional attributes into the training dataset will make it more reliable for future detection of fake news.

BIBLIOGRAPHY

[1] Martin, N. (2018, November 30). *How Social Media Has Changed How We Consume News*. Forbes. https://www.forbes.com/sites/nicolemartin1/2018/11/30/how-social-media-has-changed-how-we-consume-news/

[2] *Statista—The Statistics Portal*. (2020, June). Statista. Retrieved March 3, 2022, from https://www.statista.com/aboutus/our-research-commitment/1429/amy-watson

[3] Monti, F., Frasca, F., Eynard, D., Mannion, D., & Bronstein, M. M. (2019). Fake news detection on social media using geometric deep learning. arXiv preprint arXiv:1902.06673.

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

[5] Shu, K., Wang, S., & Liu, H. (2019, January). Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining* (pp. 312-320).

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

[7] Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. *arXiv preprint arXiv:1704.07506*.

[8] Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, A., Epstein, D., Leggio, L., & Curtis, B. (2021, May 20). *Journal of Medical Internet Research—Bots and Misinformation Spread on Social Media: Implications for COVID-19.* https://www.jmir.org/2021/5/e26933/

[9] Zhang, X., Habibi Lashkari, A., & A. Ghorbani, A. (2017). A Lightweight Online Advertising Classification System using Lexical-based Features: *Proceedings of the 14th International Joint*

Conference on E-Business and Telecommunications, 486–494. https://doi.org/10.5220/0006459804860494

[10] Flammini, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016, July). *The Rise of Social Bots*. https://cacm.acm.org/magazines/2016/7/204021-the-rise-of-social-bots/fulltext?mobile=false

[11] Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., & Bhattacharjee, B. (2007). Measurement and analysis of online social networks. *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement - IMC '07*, 29. https://doi.org/10.1145/1298306.1298311

[12] Cui, L., & Lee, D. (2020). CoAID: COVID-19 Healthcare Misinformation Dataset. *ArXiv:2006.00885 [Cs]*. http://arxiv.org/abs/2006.00885

[13] *Understanding TF-IDF (Term Frequency-Inverse Document Frequency)—GeeksforGeeks*. (n.d.). Retrieved March 21, 2022, from https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/

[14] Classification Algorithms | 5 Amazing Types Of Classification Algorithms. (2019, September 8). *EDUCBA*. https://www.educba.com/classification-algorithms/

[15] *Naive Bayes Classifier Tutorial: With Python Scikit-learn*. (2018, December 4). DataCamp Community. https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn

[16] Zhang, Z. (2019, August 14). *Naive Bayes Explained*. Medium. https://towardsdatascience.com/naive-bayes-explained-9d2b96f4a9c0

[17] *Scikit-learn SVM Tutorial with Python (Support Vector Machines)*. (2019, December 27). DataCamp Community. https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

[18] Hamdaoui, Y. (2021, March 24). *TF(Term Frequency)-IDF(Inverse Document Frequency) from scratch in python* . Medium. https://towardsdatascience.com/tf-term-frequency-idf-inverse-document-frequency-from-scratch-in-python-6c2b61b78558

[19] Brownlee, J. (2020, January 16). SMOTE for Imbalanced Classification with Python. *Machine Learning Mastery*. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

[20] Korstanje, J. (2021, August 31). *The F1 score*. Medium. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6

Allison Denenberg

ald5806@psu.edu | https://www.linkedin.com/in/allison-denenberg-3ba24b171/

EDUCATION

The Pennsylvania State University: College of Information Sciences and Technology (IST)
BS: Applied Data Sciences *Expected Graduation: May 2022*
Minor: Security Risk Analysis
Certificates: National Security Agency, Business Administration

- Schreyer Honors College: Thesis - Fake News Detection using Twitter API
- Six Semesters Dean's List

PROFESSIONAL EXPERIENCE

Accenture (Virtual) **Florham Park, NJ**
Technology Summer Analyst *June-August 2021*

- Utilized PowerPoint to organize reporting for global meetings during ERP go-live/Hypercare
- Enhanced knowledge on SAP Master Data Governance for proposal content

Accenture (Virtual) **Philadelphia, PA**
Technology Summer Analyst *June-August 2020*

- Extracted text from invoices using OpenCV, Python, & Pytesseract to enhance organizational efficiency
- Facilitated the sales and solutioning process to improve client productivity

The Valley Bank **Wayne, NJ**
Project Management Intern *June-August 2019*

- Created a Robotic Process Automated (RPA) robot via UiPath to minimize repetitive bank processes

LEADERSHIP AND INVOLVEMENT

PSU IST Diplomat, *Student Director* *Spring 2019 – Present*

- Co-leader of the student ambassadors for the College of IST
- Organize and lead all events to enhance the Diplomat Team's leadership skills

Johnson & Johnson National Case Competition, *Second place of Seven teams* *Spring 2020*

- Team collaboration to explore and present an opportunity assessment
- Determined the risk mitigations and benefits for our mobile application solution

The Pennsylvania State Women's Club Basketball Team, *Treasurer* *Fall 2018 – Present*

IST Consulting Group, *Cybersecurity Group Member* *Fall 2018 – Fall 2020*

ACADEMIC PROJECT EXPERIENCE

The Pennsylvania State University

- Processed big data via Spark programming model *Spring 2021*
- Explored the correlation between diet, health, & happiness data using predictive models *Spring 2021*
- Utilized RStudio to combine datasets to determine optimal post grad salaries *Fall 2020*
- Examined key components of a baseball statistics dataset using SQL *Spring 2019*

SKILLS

Python, SQL, R, Excel, UiPath, D3, SAP MDG
Lean IT Foundation Training *August 2019*

COMMUNITY SERVICE

Boys & Girls Club *Summer 2020*

- Facilitated research initiative for the opening of three new location sites
- Cleaned location demographics datasets and created pivot table graphics in Excel

Penn State THON *Fall 2018 – Present*