STEM Fellowship Big Data Challenge 2021

Characterization and Detection of Misinformation to Inform Healthcare Policies

Amin Abdolkhani[1], Umar Ali[1], Hamza Rasheed[1], and Johnny Zhao[1]

[1]The University of British Columbia

June 1, 2021

Abstract

The rampant spread of misinformation regarding COVID-19 is abundant on social media platforms. There is a clear need to investigate the characteristics of misinformation as it relates to healthcare in order to create effective policies. In this paper, we explore the primary characteristics of fake Twitter posts by examining the frequency, length, and vocabulary of their words used. Additionally, we use a logistic regression model, reaching an accuracy of 93.6%, to verify the veracity of Twitter posts relating to COVID-19. We propose a comprehensive list of guidelines for healthcare workers to consider when consulting patients; the aim of this list is to address the immediate characteristics of social media misinformation. By informing patients of the characteristics of social media misinformation, we hope to mitigate the perception of fake COVID-19 news as fact.

Keywords
Misinformation, COVID-19, Infodemiology, Informatics

1 Introduction

The COVID-19 pandemic is the first global outbreak to occur in the era of major information sharing websites such as Twitter and Instagram. As a result, it has been much easier for misinformation about the pandemic to spread easily and cause harm. The spread of false information has proven to create several additional issues that can intensify the severity of the pandemic and overall have a damaging effect on the public as a whole. Some outrageous examples of misinformation have caused issues such as mass panic, toilet paper hoarding, and drinking bleach as a cure for the virus. We hypothesize that there are certain characteristics of fake COVID-19 statements that delineate itself from real and truthful information. To combat the spread of misinformation, our team has developed a machine learning algorithm to detect untrue statements about the COVID-19 pandemic. Our algorithm is able to analyze tweets written about COVID-19 and predict whether or not it is likely to be incorrect. We hope to understand the characteristics that define misinformation and, using our algorithm, validate existing fake news detection systems already implemented in social media websites.

2 Materials & Methods

Our research was comprised of two main components: exploration and detection.

2.1 Data Exploration

In this paper, two major data sets were utilized: the COVID-19 Fake News data set (CFND) [1] and the COVID-19 healthcare misinformation Data set (CoAID) [2]. Specifically, in the CoAID data set, only the non-twitter information was explored. A preliminary investigation of both data sets was conducted by cleaning the text information. The following augmentations were made to the text data:

- Conversion to all lower case letters
- Removal of any numerical values and whitespace
- Removal of common English stop words (eg. and, the, etc)

- Removal of stop words related to the scenario (specifying them instead as a character vector)

Finally, one had also used Porter's stemming algorithm [3] in order to reduce words to their root form and reduce overall redundancy. In order to process and plot the text data, tokens in the CFND and the CoAID data were turned into vectors and stored as a corpus. The whole investigation was done using R in RStudio.

2.2 Misinformation Detection

Through the use of a logistic regression learning algorithm implemented using Sckit Learn and a data set of tweets contained in the CFND [1] labeled as containing true or false information, we trained a model to detect tweets containing misinformation. Our data set consisted of 7,200 COVID related tweets, 80% of which were used for training our model and 20% of which were used for testing its performance. The algorithm characterized tweets by what words they consisted of and in what quantity using Word Vectorization.

In addition to the logistic regression learning algorithm, we trained a K-means clustering algorithm using the same data set. As a redundant method of misinformation detection. We preformed dimensional reduction using Principle Component Analysis (PCA) to allow for a better visualization and identification of clusters.

3 Results

3.1 Data Discoveries

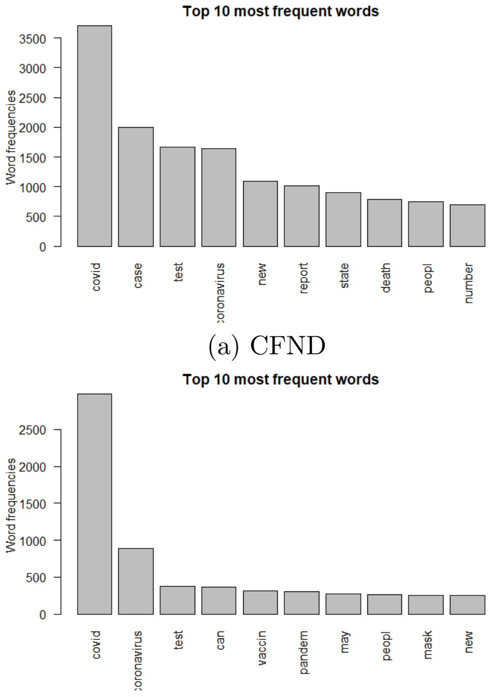
Figure 1. outlines the most frequent words that occur in the overall data set for both the CFND and the CoAID. The graph contains no distinction between fake or real words, however gives us an overall understanding of the vocabulary used between Twitter data and popular news journal title data.

Attribute	Real	Fake	Both
Unique Words	22916	19728	37503
Mean words	31.97	21.65	27.05
Mean char	218.37	143.26	182.57

Table 1. Numerical data set attributes of CFND [1]

From Table 1. it is evident that text contained in real news is typically longer than fake news text in terms of unique words, mean words per tweet, and mean characters per post [1].

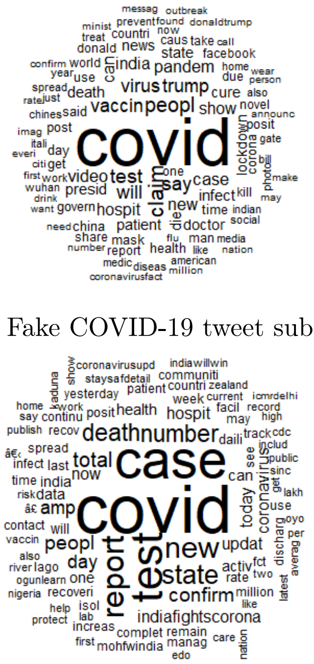
Figure 2 shows word clouds for the fake (2a) and real (2b) tweet posts; the significant overlap can be observed between both fake and real



(a) CFND

(b) CoAID

Figure 1: Top 10 most frequent words in (**a**) CFND and (**b**) CoAID



(a) Fake COVID-19 tweet subset

(b) Real COVID-19 tweet subset

Figure 2: Word clouds of most frequent words in subsets of the CFND

subsets. Moreover, within the real subset, there are more frequent uses of unique words whereas the fake subset uses a smaller vocabulary (fewer unique words).

3.2 Misinformation Detection Accuracy

The logistic regression model detected misinformation with 93.6% accuracy, further detailing its performance is Figure 3, a confusion matrix.

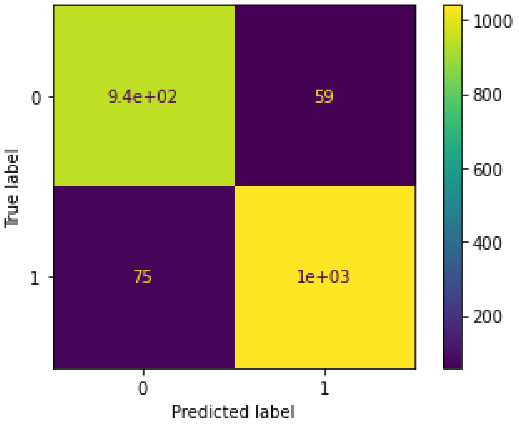


Figure 3: Confusion matrix of the linear regression model

The K-means clustering algorithm had an accuracy of 58%, however it did produce distinct clusters, shown in Figure 4 which gives us confidence that with further adjustments, we can increase the algorithms accuracy.

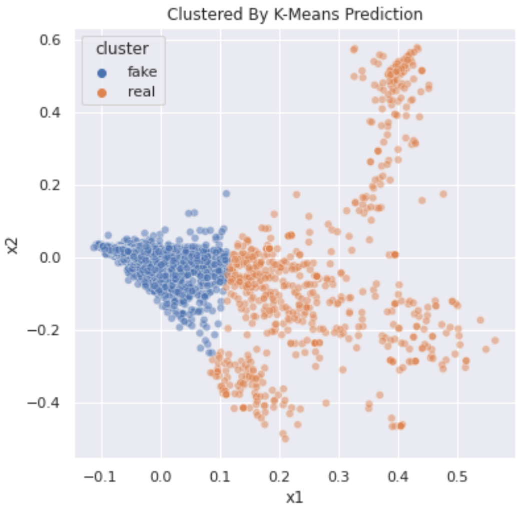


Figure 4: Clustering created by the K-means algorithm using 2 component (x1 & x2) PCA

The performance of the K-means clustering model was poor due to the unsupervised nature of the data fitting. Moreover, there is a significant overlap between both features in fake posts and real posts.

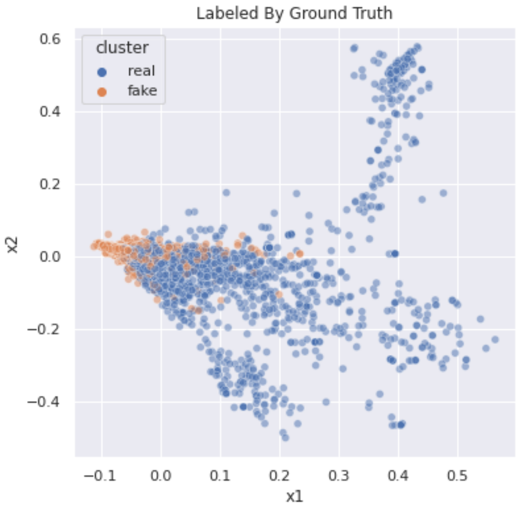


Figure 5: Clusters labeled by their ground truth

It is important to note the significant overlap found in features during the data exploration phase, specifically during the word cloud generation, and the K-means clustering model. This indicates a clear pattern to take into consideration.

4 Discussion

4.1 Characteristics of misinformation

We decided on using a supervised machine learning algorithm because we had a labeled data set available and supervised learning would allow us to get a more accurate model. Moreover, due to the small size of the CFND, a deep learning approach would not have been feasible due to their natural propensity towards large data. Regressional models, such as logistic regression and K-means clustering, are effective in determining the distinctive quantitative differences between the fake and real classes.

Our results show that there are indeed differing characteristics between fake and real COVID-19 related tweets. Tweets expressing fake information were noticeably shorter and expressed less unique words. In addition, there is different distribution in the number of frequent words used. Furthermore, the algorithm we to produced was able to accurately guess whether a tweet contained misinformation or not. This supports our hypothesis that fake and real news are fundamentally different.

There could be several sources of error for the logistic regression and K-means clustering models. For example, in twitter, the same information is constantly shared and posted in tweets. Because of this, it is possible that many of the tweets in the fake COVID-19 data set have very similar content due to being re posted frequently. This could cause the algorithm to only detect very specific messages and loose overall accuracy.

4.2 Guidelines for healthcare policies

There are several key characteristics found in misinformation regarding COVID-19: the infrequency of unique words, the total length of the text, the political context, and the range of ideas addressed. Namely, fake twitter posts tends to contain fewer unique words, specifically an overall smaller vocabulary. Moreover, the overall length of fake posts seem to be relatively shorter when compared to real posts. Misinformed media also follows a trend of being more politically focused. These types of posts often rely on exploiting the political opinions of individuals rather than being based in scientifically supported research. Lastly, it was found during in the K-means clustering analysis, refer to Figure 4, that fake twitter posts tends to be more concentrated in terms of its contents whereas real posts are more broad and sporadic.

We propose a short list of guidelines based the the aforementioned characteristics of misinformation for healthcare workers to consider when consulting patients. Patient consultation is an important modality for the dissemination of accurate healthcare related information. Some healthcare workers, such as pharmacists, are an under utilized resources for improving community level comprehension. As such, it is important to take advantage of these resources and increase overall public awareness of COVID-19 misinformation.

4.3 Future Investigations

We have some concerns regarding the high accuracy of our linear regression model. We plan to conduct extensive validation on it before using it to verify existing fake news detection systems as described in the introduction.

One possible venue for future studies could be exploring the use of sentiment as a model parameter. Observing a few of the key words that were frequent in both data sets, a disparity was noticed between the sentiment in fake news when compared to real posts. Namely, there existed a more negative and combative vocabulary, using words such as infect, kill, and die, in the fake news data set. As such, we predict that sentiment plays a critical role in the prediction of fake COVID-19 information.

Conclusions

In this study, we determined several key characteristics of COVID-19 misinformation which could be used to guide healthcare policies. We were able to use a logistic regression model, with an accuracy of 93.6%, to predict whether COVID-19 related facts were true or false. This model could be used to validate content moderation algorithms in websites such as Twitter or Instagram.

Acknowledgements

Thank you to Anish R. Verma and the UBDC mentoring team for providing guidance to us during this opportunity.

References

[1] Sourya Dipta Das, Ayan Basak, and Saikat Dutta. A heuristic-driven ensemble framework for covid-19 fake news detection. *arXiv preprint arXiv:2101.03545*, 2021.

[2] Limeng Cui and Dongwon Lee. Coaid: Covid-19 healthcare misinformation dataset, 2020.

[3] Peter Willett. The porter stemming algorithm: then and now. *Program*, 2006.