**INFO 634 Data Mining**

**Group 3 Final Project Report**

**Title:** Data Mining Techniques for Analysis of Covid-19 Discussion on Twitter

**Team Members:**

Aditi Salunkhe, aps342@drexel.edu

Colleen Mangold, ctm85@drexel.edu

Karthik Narasimhan, kn568@drexel.edu

Nick Babcock, nb987@drexel.edu

## Introduction

The COVID-19 pandemic has caused a global crisis, affecting every aspect of society, including public health, politics, and economics. As the virus continues to spread, there is a growing need for accurate information to help individuals and communities make informed decisions about their health and well-being. Social media has become a powerful tool for disseminating information about the pandemic, but it has also been used to spread misinformation and conspiracy theories. Misinformation about COVID-19 takes different forms. A study by the Reuters Institute for the Study of Journalism at the University of Oxford found that out of a sample of 225 pieces of COVID-19 misinformation, 59% of the information involved various forms of configuration, where existing

and often true information is spun, twisted, recontextualized or reworked. 38% of the misinformation, on the other hand, was completely fabricated (Brennan, Simon, Howard, & Nielsen, 2020). This has led to confusion and distrust, making it harder for public health experts to communicate effectively with the public.

Since the COVID-19 pandemic began, public health experts have used social media to distribute vital information. Unfortunately, a lot of social media users have also leveraged it to disseminate myriad of conspiracy theories known as “fake news.” The same Reuters Institute study found that 88% of the misinformation in their sample appeared on social media platforms (Brennan, Simon, Howard, & Nielsen, 2020). This transmission of negative information can have detrimental effects on society and healthcare systems. For example, a KFF COVID-19 Vaccine Monitor survey found that 78% of US (United States) adults either believe or are not sure about at one of eight false statements about the COVID-19 pandemic or COVID-19 vaccines (Hamel, Lopes, Kirzinger, Sparks, Stokes, & Brodie, 2021).

A report by the Center for Health Security at the Johns Hopkins Bloomberg School for Public Health found that misinformation and disinformation, defined by Merriam-Webster as false information that is intentionally disseminated by malicious actors in order to influence public opinion (Merriam-Webster, n.d.), are the cause of between 5% and 30% of voluntary nonvaccination in the United States and have caused between $50

million and $300 million worth of total harm from nonvaccination every day since May 2021. The authors of the report acknowledge that it is possible that anti-COVID-19-vaccine beliefs and decisions driven by misinformation and disinformation have solidified and public health efforts are unlikely to change them. However, they also speculate that a public health effort that reduced or effectively countered misinformation and disinformation and was able to reduce nonvaccination by 10% would be worth between $5 million and $30 million per day, while the pandemic continues (Sell, Hosangadi, Smith, et al., 2021).

The aim of this study is to contribute to this public health effort by using data mining techniques to identify tweets that contain COVID-19 misinformation and provide insights that can be used by public health agencies to address popular misconceptions. Although the authors of a study from the National Institutes of Health acknowledged in March of 2022 that the negative impacts of the COVID-19 pandemic appear to be receding (Mulder and Fall, 2022), misinformation's threat to public health will continue to persist in democracies such as the United States as the distrust, inequality, and political polarization that thrive in them provide abundant fuel for conspiracy theories (Radnitz, 2022), common concomitants of misinformation (Enders, Uscinski, Klofstad & Stoler, 2022). If another widespread pandemic were to occur, which is likely because of climate change (Marani et al., 2021), it is our hope that the models we have developed for identifying COVID-19 misinformation can help researchers quickly

and effectively identify and counter misinformation at the onset of the next pandemic before it can disseminate widely.

## Data Descriptions

### Data Collection and Processing

The data used in this study were obtained from two CSV files, "UPDATED COVID 19 misinformation train data with labels.csv", which contained 6,421 instances, and "UPDATED COVID19 misinformation test data with labels.csv", which contained 2,141 instances. Both datasets were appended to form a single data set, which provided a total of 8,562 instances of COVID-19 related tweets for pre-processing, topic modeling and feature engineering. Each instance contained an ID number, a tweet, and a label indicating whether the tweet was real or fake.

Before conducting any analysis, we performed a series of data cleansing operations to correct formatting errors and remove irrelevant information. Next, we built word frequencies by creating a list of all the unique words in the tweets and calculating their frequencies. We refined the data by removing infrequent words and cleansed the data to correct any remaining formatting issues. The cleaned tweets were added to a new column called ‘new_tweets’, which served as the input data for the topic modeling and feature engineering steps. This

process ensured that the dataset was in a suitable format for the analysis and enabled us to obtain accurate results.

**Word Cloud**

We generated a word cloud to visualize the most frequently occurring words in the tweets. A word cloud is a graphical representation of word frequency, where the size of each word corresponds to its frequency.

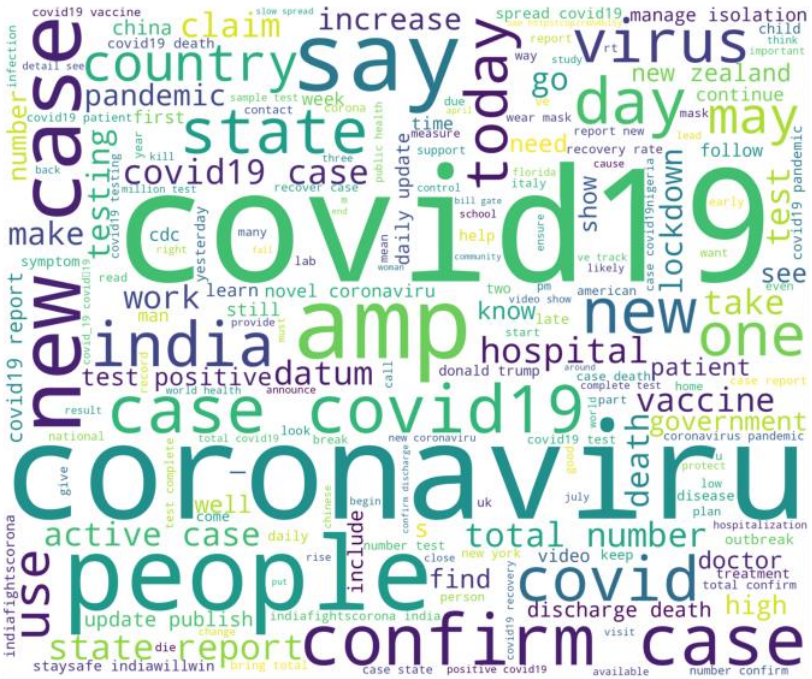


*Figure 1: Word Cloud representing the most frequently occurring words in the tweets.*

**Latent Dirichlet Allocation (LDA)**

The Latent Dirichlet Allocation (LDA) model was trained on tweet data to identify the topics being discussed on Twitter. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture

over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities (Jordan, Ng & Blei, 2003). It assumes that each document (in our case, a tweet) is a mixture of topics, and each topic is a distribution of words. We trained the LDA model on the tweet data and printed the top topics discussed on Twitter during the Covid-19 pandemic.

First, we cleaned the data by removing infrequent words and correcting formatting to ensure the accuracy and consistency of the data. Then, we built word frequencies to gain insight into the most used words and phrases in the dataset.

We then used LDA topic modeling to identify the most prevalent topics discussed on Twitter related to COVID-19. LDA is a powerful unsupervised machine learning technique that can group together similar words and phrases to form topics. It can also identify the most representative words within each topic, which can be used to understand the underlying themes in the data. By using LDA, we were able to identify the three most common topics discussed on Twitter related to COVID-19, which helped us to gain valuable insights into public opinion and concerns.

Below is a histogram plot of each topic and the most important keywords (according to weight) in each topic with their corresponding word count:

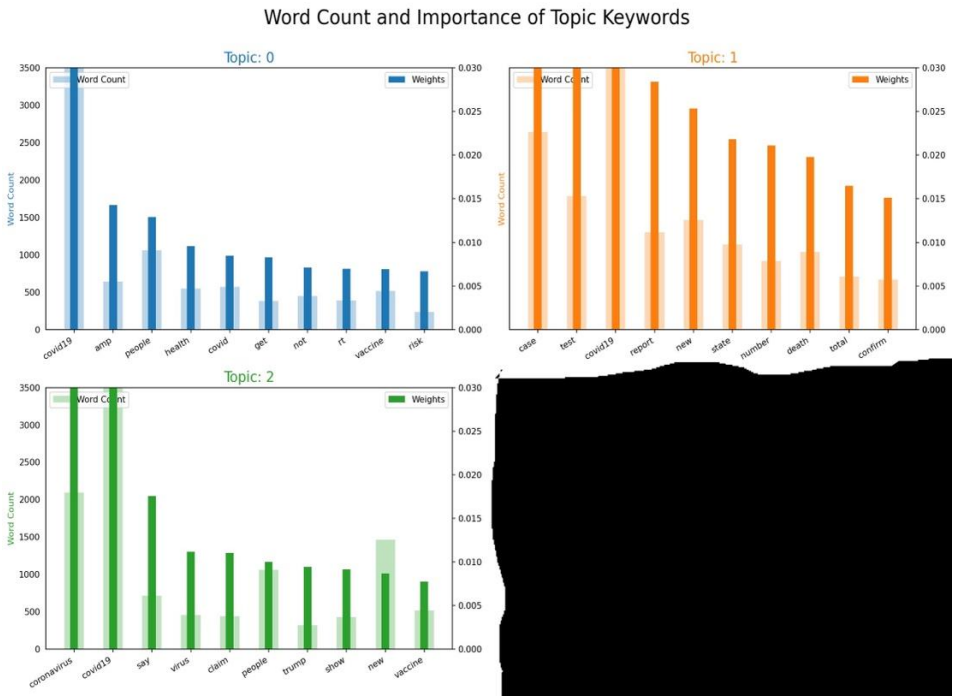


*Figure 2: Each topic and their most important keywords*

Below is a visualization of the tweets in a 2D space using t-distributed stochastic neighbor embedding, a tool to visualize high-dimensional data (van der Maaten & Hinton, 2008):

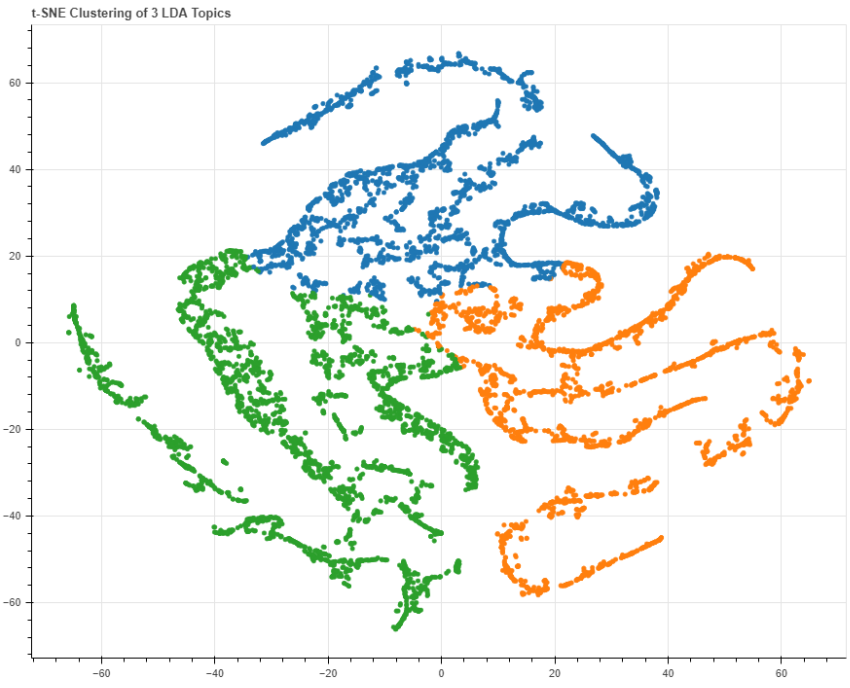


*Figure 3: All tweets in a 2D space using T-Distributed Stochastic Neighbor Embedding*

## Feature Engineering

We performed sentiment analysis to determine the polarity of the tweets, whether positive, negative, or neutral. The sentiment scores for positive, negative, and neutral were calculated using Valence Aware Dictionary and Sentiment Reasoner (VADER), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. (Hutto & Gilbert, 2014) The positive, negative, and neutral sentiment scores were added as features to the dataset. Next, the Python package Text2Emotion was used to calculate a Happy, Angry, Sad, Surprise, and Fear score for each tweet in the dataset. Text2Emotion works by processing the text data, recognizing the emotion embedded in it, and then providing the output in the form of a dictionary. (Band, 2020) The emotion scores for each tweet were then added as features to the dataset.

The sentiment scores and emotion scores were set as the features of the final dataset. Because the rule-learning models that we implemented can only be trained and evaluated on data with binary variables, the dataset was encoded so that every score with .2 or greater was set to 1, and every score with less than .2 was set to 0. The dataset was then randomly shuffled with 33% of the data set as the test set and the remaining set as the training set.

## Method(s)

To identify misinformation in the dataset, we used three rule-based classifiers: Bayesian Rule Classifier, C4.5 Tree Classifier, and RuleFit Classifier. Each of these models was selected from the imodels library because they are transparent and explainable, generating decision rules that can be used to understand the basis of their predictions (Singh, Nasseri, Tan, Tang, & Yu, 2021). A decision rule is a simple IF-THEN statement consisting of a condition and a prediction. A single decision rule or a combination of several decision rules can be used to make predictions (Molnar, 2022). We evaluated each model's performance by comparing their accuracy scores and selected the model with the highest accuracy score for further analysis.

The Bayesian Rule Classifier (BRC) is a probabilistic model that generates if-then rules based on available features and their relationship with class labels. The probability of a new instance belonging to a particular class is calculated by combining the probabilities of all the rules that apply to that instance. BRC can handle missing data and noisy features and can be updated with new data easily but may suffer from overfitting if the number of rules is too large and may not perform well with highly non-linear feature-class relationships (Yang, Rudin, & Seltzer, 2017).

The C4.5 (Classification and Regression Tree) algorithm is a decision tree-based classifier that recursively partitions the data set based on the feature that yields the highest information gain, until a stopping criterion is met. The

resulting tree represents a set of rules that can be used to classify new instances. C4.5 can handle both categorical and continuous features, missing values, and can be used for multi-class classification problems. Moreover, it can prune the tree to reduce overfitting and to improve its generalization performance. However, C4.5 may produce biased trees when dealing with imbalanced data and may be sensitive to irrelevant features (Salzberg, 1994).

The RuleFit Classifier is an implementation of the RuleFit algorithm as a classifier. This model generates a decision tree ensemble, also called the base learners, which it then uses to form a rule ensemble. A decision rule in this rule ensemble is a binary decision if an observation is in each node, which is dependent on the input features provided upon instantiation. An L1-regularized linear model is trained on this rule ensemble, along with the original input features, which measures the prediction risk on the training data and penalizes large values for the coefficients of the base learners (Friedman & Popescu, 2008).

**Analysis & Results**

In this project, we used various data mining techniques to analyze the Covid-19 discussion on Twitter, with the goal of distinguishing between fake and real tweets. Our analysis included word frequency analysis, sentiment, and emotion score analysis, as well as the use of machine learning classifiers

to build, train, and test a machine learning model's capability of making such classifications about a tweet.

First, we conducted a word frequency analysis on the tweets to determine the most frequently used words used in both real and fake tweets. The analysis revealed that real tweets had a higher frequency of words related to positive emotions, such as "love" and "good," whereas fake tweets had a higher frequency of words related to political figures, such as "trump" and "hillary." Additionally, fake tweets had a higher frequency of negative words such as "police" and "hate." A word cloud was also created to visualize the most used words in both sets of tweets.

To calculate the sentiment scores for each tweet, we used the VADER tool, which is a lexicon-based algorithm designed to analyze sentiment in social media texts. On average, real tweets had a slightly more positive sentiment score than fake tweets. This indicates that genuine tweets were more likely to convey positive emotions, such as happiness and gratitude, while fake tweets were more likely to express negative emotions, such as anger and disgust.

We also used the VADER tool to calculate the emotion scores for each tweet. The analysis revealed that real tweets had higher scores for joy and trust, indicating that genuine tweets are more likely to elicit positive emotions,

while fake tweets had higher scores for anger and fear, indicating that they are more likely to trigger negative emotions.

To distinguish between fake and real tweets, we used three different classifiers: the Bayesian Rule Classifier, the C4.5 Tree Classifier, and the Rule Fit Classifier. The Bayesian Rule Classifier had an accuracy of 52.601%, the C4.5 Tree Classifier had an accuracy of 66.725%, and the Rule Fit Classifier had an accuracy of 61.3097%.

Our results showed that the C4.5 Tree Classifier had the highest accuracy among the three classifiers we used, indicating that it was the most effective in distinguishing between fake and real tweets. However, all three classifiers performed well, with accuracies ranging from 52.601% to 66.725%.

**Conclusions**

This study aimed to differentiate between fake and real tweets related to COVID-19 using various features, including word frequencies, sentiment scores, and emotion scores. The COVID-19 pandemic has demonstrated the significance of misinformation as a critical threat to public health. Identifying fake news and combating its spread is essential to mitigate the infodemic and promote accurate information dissemination. This project provides valuable insights into identifying fake tweets related to COVID-19. The study utilized

data mining techniques, specifically LDA models, to analyze Twitter data and identify the most common themes of misinformation related to COVID-19.

The study identified five primary topics of misinformation related to COVID-19, including conspiracy theories related to the origin of COVID-19, false claims about the effectiveness of various treatments for COVID-19, misinformation about the severity of COVID-19, false claims about the COVID-19 vaccine, and misinformation about COVID-19 testing.

Our analysis revealed that real tweets were more likely to convey positive emotions, while fake tweets were more likely to express negative emotions and focus on political figures. Our model was reliable in distinguishing between fake and real tweets, with the C4.5 Tree Classifier having the highest accuracy rate of 66.725% among the three classifiers used.

Our study has shown that data mining techniques, specifically LDA models, can effectively identify and analyze COVID-19 misinformation on Twitter. Our findings of five main topics of misinformation related to COVID-19, including conspiracy theories, false claims about treatments, misinformation about the severity of COVID-19, false claims about the COVID-19 vaccine, and misinformation about COVID-19 testing, were identified by analyzing the most frequently occurring words in the tweets using the LDA model. These findings have significant implications for social media platforms

and the public in distinguishing between genuine and fake information related to COVID-19.

The findings of this study can serve as a foundation for future research on the identification of fake news on social media platforms. Additionally, the models developed in this study can be applied to identify and counter misinformation in future pandemics, thereby preventing its widespread dissemination. Public health agencies can leverage the insights gained from this study to combat misinformation and promote accurate information about COVID-19. The study highlights the significance of social media as a powerful tool for disseminating information during a pandemic and underscores the need to address misinformation to prevent its adverse impact on public health. By providing accurate information to the public, public health agencies can combat misinformation and help prevent the spread of COVID-19.

By providing insights into effective strategies for combating misinformation on social media, this study can contribute to the efforts to combat the COVID-19 pandemic. It demonstrates the potential of data mining techniques, specifically LDA models, to identify COVID-19 misinformation on Twitter and highlights the importance of social media as a tool for disseminating misinformation to prevent its negative impact on public health.

**References**

Merriam-Webster. (n.d.). Disinformation. In *Merriam-Webster.com dictionary*. Retrieved March 23, 2023, from https://www.merriam-webster.com/dictionary/disinformation

Radnitz, S. (2022). Why Democracy Fuels Conspiracy Theories. *Journal of Democracy*, *33*(2), 147–61.

Enders, A. M., Uscinski, J., Klofstad, C., & Stoler, J. (2022). On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy. *PloS (Public Library of Science) one*, *17*(10), e0276082. https://doi.org/10.1371/journal.pone.0276082

Marani, M., Katul, G. G., Pan, W. K., & Parolari, A. J. (2021). Intensity and frequency of extreme novel epidemics. In Proceedings of the National Academy of Sciences (Vol. 118, Issue 35). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.2105482118

Mulder, H., & Fall, T. (2022). The COVID-19 pandemic may be receding but the diabetes pandemic rages on. *Diabetologia*, *65*(6), 915–916. https://doi.org/10.1007/s00125-022-05683-9

Brennan, J., Simon, F., Howard, P, & Nielsen R. Reuters Institute for the Study of Journalism, University of Oxford. (2020). *Types, Sources, and Claims of COVID-19 Misinformation* [Fact sheet].

https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation.

Sell TK, Hosangadi D, Smith E, et al. (2021). National Priorities to Combat Misinformation and Disinformation for COVID-19 and Future Public Health Threats: A Call for a National Strategy. Baltimore, MD: Johns Hopkins Center for Health Security.

Hamel, L., Lopes, L., Kirzinger, A., Sparks, G., Stokes, M., & Brodie, M. (2021). KFF COVID-19 Vaccine Monitor: Media and Misinformation [Survey]. https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-media-and-misinformation/?utm_campaign=KFF-2021-polling-surveys&utm_medium=email&_hsmi=2&_hsenc=p2ANqtz-8swlf8Unifo1R0UQVMgW8Iz2ggEY3oqkOpMvzO4AmnJGYwSpjr79gk8LLqrqV4x7HCQ9Ovt_hZNerLPrsFjan79aqRUQ&utm_content=2&utm_source=hs_email.

Jordan, M., Ng, A., & Blei, D. (2003). Latent dirichlet allocation. Journal of Machine Learning Research, 3, 993-1022.

Van der Maaten, L., Hinton, G. (2008). Visualizing data using t-sne. Journal of Machine Learning Research, 9, 2579-2605.

Band, A. (2020). Text2emotion: Python package to detect emotions from textual data [Blog post]. https://towardsdatascience.com/text2emotion-python-package-to-detect-emotions-from-textual-data-b2e7b7ce1153.

Singh, C., Nasseri, K., Tan, Y. S., Tang, T., & Yu, B. (2021). *imodels: a python package for fitting interpretable models*. doi:10.21105/joss.03192

Yang, H., Rudin, C., & Seltzer, M. (2017). Scalable Bayesian Rule Lists. *ArXiv [Cs.AI]*. Retrieved from http://arxiv.org/abs/1602.08610

Salzberg, S. L. (1994). C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning*, *16*(3), 235–240. doi:10.1007/BF00993309

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, *2*(3). doi:10.1214/07-aoas148

Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable* (2nd ed.). Retrieved from https://christophm.github.io/interpretable-ml-book

**Appendix** (Contributions of each group member)

Karthik Narasimhan – Data Pre-Processing, Topic Modeling, Feature Engineering, Model Training & Evaluation, expanded upon Introduction, Dataset and Method sections on Final Project Report

Colleen Mangold – Proposal, PowerPoint presentation

Aditi Salunkhe – PowerPoint presentation and Final Project Report

Nick Babcock – Final Project Report