Understanding Misinformation: A Case Study of COVID-19 Social Media Posts

Group 2: Teshinee Kukamjad, Jolina Hor, Sam Reade *All authors have equal contribution*

Introduction: On March 11, 2020, the World Health Organization (WHO) officially declared COVID-19 a global pandemic. Just a little over three years later, on May 5th, this global health emergency ended. Unfortunately, throughout this period, the pandemic not only posed a significant public health challenge to the world but also gave rise to widespread misinformation on social media platforms. According to Pew Research Center (2021), nearly half of Americans reported obtaining either ‘some’ (30%) or ‘a lot’ (18%) of news and information about COVID-19 vaccines from online sources. While this statistic may not appear as significant, such information can lead to rumors, stigma, discrimination, and false theories, ultimately creating confusion that restricts individual and public health responses during times of crisis.

Background: In response to this growing issue, numerous researchers have developed models aimed at detecting and mitigating the spread of false information. Murugesan et. al (2022), for example, utilizes machine learning techniques like Adaboost & Decision tree algorithm to identify fake news in the medical domain. Similarly, Farhoudinia et. al (2024) looked into using sentiment analysis on a COVID-19 Twitter dataset to detect fake news. While these research papers have offered valuable insights, there is still a research gap in combining all detection methods together to identify the most significant features contributing to misinformation classification.

Research Question: Which social media features are most important for distinguishing real vs misinformation posts?

Data: The *Fighting an Infodemic: COVID-19 Fake News Dataset* (Patwa et al., 2021) consists of 10,700 English-language social media posts related to the COVID-19 pandemic. The content was gathered from Twitter, Facebook, and Instagram using web scraping and the Twitter API. Real news posts were collected from trusted organizations, including the World Health Organization (WHO), the Centers for Disease Control and Prevention (CDC), and Covid India Seva. In contrast, fake news posts were gathered from misinformation posts on social media. The authors included 5,600 real news posts and 5,100 fake news posts to ensure a balanced dataset. The dataset contains three columns: a numeric index, the text of the post, and a binary label indicating whether the post is classified as “real” or “fake.” The research team manually verified each post to ensure accurate labeling.

Data Cleaning: Social media text is often noisy and contains elements like links, mentions, and numbers. We applied several preprocessing steps to clean the text and extract meaningful language patterns. First, we converted all text to lowercase to treat words like "COVID" and "covid" as the same. We then replaced URLs with the token *link*, mentions (e.g., @user) with *mention*, and numbers (e.g., 24k, 3.5M) with *number*. As these specific values are often unique to the post, we replace those with simple labels so we can keep the important structure of the post without adding unnecessary information to the model. We used TweetTokenizer to split each post into tokens. We filtered out standard English stopwords from the NLTK library and the top 20 most frequent words in the dataset, as they might not help the model differentiate between posts containing real information and misinformation.

Method: We want to identify the most influential factors for detecting misinformation in tweets using feature engineering and machine learning. For each individual tweet we will extract a set of key features, including tweet length, hashtag count, link presence, emotion, assigned using the DisTilBERT model, sentiment polarity, which will have the levels positive, negative, or neutral, and important words identified through TF-IDF weighting. These features are then used to train a Random Forest classification model. The dataset will be split into training and testing sets to evaluate the model’s performance, making sure that it generalizes well to unseen data. We selected the Random Forest algorithm because it is robust to overfitting, can handle a mix of categorical and numerical features, and provides interpretable feature importance metrics. Random Forest was especially attractive for this project because of its interpretability. This makes it ideal for our goal of not only classifying tweets as real or fake, but also understanding which tweet characteristics are most predictive of misinformation. The model's output will help us pinpoint patterns in tweet content and structure that correlate strongly with authenticity, providing both predictive power and actionable insights.

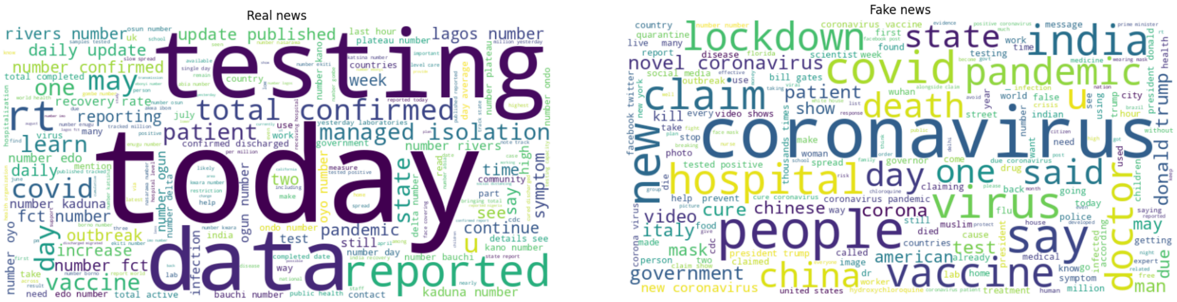

Figure 1: Word clouds of posts containing real information (left) and misinformation (right)

Expected Results: We expect to find that emotional and structural features such as sentiment polarity and emotion labels differ significantly between real and fake tweets. The results will be visualized using several techniques. We will use feature importance plots to highlight which variables most strongly influence the classification, confusion matrices to assess the accuracy of the model across real and fake tweet predictions, and emotion distribution charts to compare the prevalence of specific emotional tones across tweet types. These visualizations will provide interpretability and validation of our model’s performance. In addition to highlighting the role of emotional and structural features in distinguishing real from fake tweets, we expect several other informative results from this project. We anticipate that fake tweets will exhibit higher rates of extreme emotional content like anger, fear, or disgust. While real tweets may display more neutral tones. We also expect structural features such as the presence of links or hashtags to be different between tweet types, with fake tweets potentially using more hashtags or fewer credible links to drive engagement. The Random Forest model’s feature importance scores will help uncover which features consistently contribute most to accurate classification, potentially revealing new insights into patterns of misinformation dissemination.

Discussion: This project brings up several interesting avenues for future research. One idea involves expanding the analysis to examine tweets over multiple years, instead of limiting the scope to a narrow time frame. This would allow us to investigate whether the emotional and structural patterns that characterize misinformation are stable over time or evolve with changing social and political contexts. Longitudinal analysis could also help identify whether certain events or periods such as elections or times of crisis lead to spikes in specific types of fake content, or shifts in how misinformation is conveyed. Another valuable extension would be to analyze how emotional content in fake posts correlates with user engagement metrics, such as likes, retweets, and replies. Understanding whether emotionally charged misinformation tends to generate higher engagement can provide insight into the ways that drive its spread and popularity on social media. These extensions would deepen our understanding of misinformation and help design more effective tools for detection and mitigation.

Works Cited

Farhoudinia, B., S. Ozturkcan, and N. Kasap. "Emotions Unveiled: Detecting COVID-19 Fake News on Social Media." *Humanities and Social Sciences Communications*, vol. 11, 2024, article no. 640. https://doi.org/10.1057/s41599-024-03083-5.

Mitchell, Amy, and Jacob Liedke. “About Four-in-Ten Americans Say Social Media Is an Important Way of Following COVID-19 Vaccine News.” *Pew Research Center*, 24 Aug. 2021, https://www.pewresearch.org/short-reads/2021/08/24/about-four-in-ten-americans say-social-media-is-an-important-way-of-following-covid-19-vaccine-news/.

Murugesan, S., and K. Pachamuthu. "Fake News Detection in the Medical Field Using Machine Learning Techniques." International Journal of Safety and Security Engineering, vol. 12, no. 6, 2022, pp. 723–727. https://doi.org/10.18280/ijsse.120608.

Patwa, Parth, et al. “Fighting an Infodemic: COVID-19 Fake News Dataset.” *Combating Online Hostile Posts in Regional Languages during Emergency Situations*, edited by Tanmoy Chakraborty, Kai Shu, Huan R. Bernard, Huan Liu, and Md. Shad Akhtar, Springer, 2021, pp. 21–29, https://doi.org/10.1007/978-3-030-73696-5_3.