Employing Neural Language Models and A Bayesian Hierarchical Framework for Classification and Engagement Analysis of Misinformation on Social Media

Abbey List
Duke University
Undergraduate Honors Thesis in Computer Science and Statistical Science
abbey.list@duke.edu

Date: _____

Approved:

_____

Sam Wiseman (Advisor)

_____

Yue Jiang (Advisor)

_____

Bhuwan Dhingra (Committee Member)

_____

Colin Rundel (Committee Member)

_____

Alexander Volfovsky (Committee Member)

Thesis submitted in partial fulfillment of the requirements for Graduation with Distinction for Double Honors in the Department of Computer Science and the Department of Statistical Science

2022

ABSTRACT

While social media can be an effective tool for maintaining personal relationships and making global connections, it has become a powerful force in the damaging spread of misinformation, especially during universally difficult and taxing events such as the COVID-19 pandemic. In this study, we collected a sample of Tweets related to COVID-19 from Twitter accounts of influential political media commentators and news organizations, assigning labels of misinformation, misleading, or legitimate to each Tweet. We constructed a Bayesian hierarchical negative binomial regression model to analyze any associations between Tweet engagement and misleading status while controlling for factors such as political lean, lexical diversity, and Retweet status. We found evidence that engagement had a positive association with misleading status and text readability, as well as a negative association with Retweets. We also employed a DeBERTa neural language classification model to predict the presence of misinformative or misleading content in Tweets, and we experimented with external datasets, multitask fine-tuning, backtranslation, and weighted loss to achieve accuracy of 0.683 and a macro F1-score of 0.593. We then examined DeBERTa explainability through word attributions with integrated gradients and found that tokens with the highest influence on model predictions often possessed connotations or context that was understandably related to the predicted label. The results of this study indicate that misleading status, Retweet status, and linguistic features may hold associations with overall Tweet engagement, and the DeBERTa model represents a potentially useful tool that can examine Tweet text alone without an external knowledge base and determine whether misinformation is present.

Acknowledgements

I'd first like to thank my thesis advisors Dr. Sam Wiseman and Dr. Yue Jiang for their invaluable support and guidance throughout the course of this project. I cannot overstate my appreciation for their willingness to answer my questions and provide direction, always with patience and understanding, and I am extremely fortunate to have received from them such a positive introduction to undertaking research. I would also like to acknowledge my committee members Dr. Bhuwan Dhingra, Dr. Colin Rundel, and Dr. Alexander Volfovsky for contributing to the project and providing feedback – I am incredibly grateful for the time they took to discuss the thesis with me and share their insights to help me further improve this work. I am very appreciative towards Dr. Joan Combs-Durso, Dr. Mine Çetinkaya-Rundel, and Dr. Susan Rodger for their encouragement and advice, not only during my thesis project but also throughout my educational career at Duke. I would also like to thank my friends and family for supporting me through the process, listening to my ideas, and staying up late to help me proofread.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

In recent decades, the ubiquity of social media has made the dissemination of lies and propaganda easier than ever before, allowing untrustworthy sources to achieve sizable followings across various platforms. Particularly in periods of great collective stress, such as a polarizing election or a global pandemic, media consumers are vulnerable to the vast amounts of propaganda, misinformation, and other "fake news" to which they are exposed each day. However, the technological developments that allow illegitimate sources to produce misleading content with alarming speed and skill are the same advancements that have yielded opportunity for great innovation in the fields of natural language processing and statistical analysis.

The aims of the current study are two-fold, ultimately focused on understanding misleading propaganda in social media and its trends as they relate to significant current events. We recognize that propaganda can include information presented to encourage positive behaviors, such as displaying infection rates between vaccinated and unvaccinated populations to promote a vaccine. This project will center primarily around propaganda with negative effects or "fake news," which presents misleading or false information usually to provoke a specific reaction. Firstly, we intend to utilize Tweets related to COVID-19 from the past 1.5 years to train a language classification model to identify Tweets which contain untrustworthy propaganda. We believe the results of this model will not only yield insights into common characteristics of Tweets containing misinformation, but also provide a practically useful tool for warning the public of fake news on social media.

In addition to automating Tweet classification, we aim to analyze potential associations between misleading Tweets and engagement metrics while controlling for factors such as time period, lexical diversity, and political lean. Utilizing a Bayesian negative binomial regression with random effects and random slopes based on Twitter account, we will examine whether overall Tweet engagement (a combined metric derived from Retweets, Likes, Replies, and Quotes) may be associated with whether the Tweet is misleading, as well as whether any association may fluctuate based on political lean. Through these classification and analysis tasks, we hope to provide valuable information about the general indicators of negative propaganda in social media and how they may relate to overall engagement and social media traffic.

2 Related Work

An ingredient common to environments where misinformation can thrive is often the combination of novel, tumultuous circumstances with a lack of scientific consensus and emerging evidence-based claims that may contradict each other (Vraga & Bode, 2020). These conditions are often satisfied

in a multitude of modern debates, including climate change, voter fraud, and immigration reform (Treen et al., 2020; Berlinski et al., 2021; Abascal et al., 2021). While these issues are undoubtedly consequential and merit serious consideration, misleading content is perhaps most insidious and potentially lethal in the realm of public health information, and social media provides the perfect opportunity for these dangers to spread to millions of users in seconds. Historically, health topics that attract high rates of online misinformation have included vaccines, Ebola, Zika virus, and cancer (Wang et al., 2019), which share the common threads of uncertainty surrounding scientific guidance and potential severity of their effects. Consequences of dissemination of erroneous health-related information can be relatively benign, but they can also be tragically fatal: during the 2014 Ebola outbreak, falsehoods were spread over text about salt water as an effective cure, resulting in the deaths of several Nigerians and hospitalization of many more (Neporent). Fallout is not only limited to personal repercussions; inaccurate information and fear of Ebola drove rioters to attack the facilities and healthcare workers at a Liberian hospital in 2014 (Fantz).

According to a 2019 investigation which analyzed trends in the dissemination of fake news on social media, the rate at which users encounter false content has been increasing on Twitter since 2015 (Allcott et al., 2019), and a 2018 review claims that the spread of misinformation has become exacerbated by personally curated, tight-knit online circles that create echo chambers, as well as a general distrust in institutions like government and healthcare (Chou et al., 2018; Norman). While social media can be a convenient vector for the dissemination of misleading claims, it can also be leveraged to correct misperceptions: several experiments have found that on Facebook News Feed, false information correction mechanisms which utilized both automatic algorithms and other Facebook accounts to share legitimate news stories were successful in correcting the misunderstandings of users (Bode & Vraga, 2017; 2015). In an attempt to broadcast credible health science, numerous medical journals have opened social media accounts on Twitter and other platforms, gaining significant influence in these online spaces (Goel & Gupta, 2020); however, there is still danger in normalizing the practice of obtaining scientific information from social networking websites without simultaneously teaching skills to discern legitimate and illegitimate claims.

With the onset of COVID-19 in late 2019, recent studies of misinformation have examined the impact of the pandemic on trust in online communication and social media fatigue. In 2020, the World Health Organization (WHO) even established the term 'infodemic,' referring to the overabundance and mixing of reliable and questionable information that heavily strains the management of public health crises (WHO). A review of misleading posts on various social media outlets did not find evidence of differences in interaction patterns between legitimate and questionable sources, although the volume of deceptive content differed between websites (Cinelli et al., 2020). A user's likelihood of sharing misinformation has been found to hold an association with the user's purpose

for using social media: those interested in entertainment or in promoting a personal brand appear more likely to spread questionable information. Those intending to explore novel topics and stay current on global events appear less likely to engage in this behavior, although they also seem more prone to social media fatigue and burnout (Islam et al., 2020).

Automating the challenging task of classifying propaganda and fake news and identifying their characteristics in online sources has become increasingly popular in statistics and natural language processing. Misinformation detection can be formulated as a classification problem which seeks to categorize text into levels of legitimacy, a regression problem which quantifies the degree of legitimacy, or even a semi-supervised clustering problem that utilizes extralinguistic properties like user credibility to infer legitimacy (Su et al., 2020; Yang et al., 2019). Statistical models including support vector machines and naïve Bayes classifiers have displayed promising accuracy in detecting misinformative content by utilizing a wide range of text characteristics, including complexity cues, Term Frequency-Inverse Document Frequency (TF-IDF) representation, and other linguistic aspects like lexical, syntactic, or semantic features. Recent approaches also focus on more computationally-powerful neural networks like LSTM models or RNN-based models, which circumvent the requirement for time-consuming hand-crafted features and allow for modeling more complex connections between text and legitimacy at the expense of interpretability (Su et al., 2020).

A 2019 study utilized quoted attribution techniques and influence mining in a Bayesian machine learning system to estimate the probability that an article was illegitimate, achieving results that were 63.33% effective at identifying fake news articles (Traylor et al., 2019). Through an unsupervised clustering approach, it was found that propagandists on Twitter often interacted with similar accounts on similar timeframes, and their mean time between posts was 86% shorter than that of regular Twitter users (Orlov & Litvak, 2019). Furthermore, an analysis of Twitter posts related to a German federal election found that propagandists from U.S. campaigns were similarly interfering in German politics (Kellner et al., 2019). This provides evidence towards the detrimental universality of propaganda and has likely motivated events in the field of natural language processing such as open competitions dedicated to the detection of propaganda in news articles (Hua, 2019).

In addition to analyzing propaganda online, previous studies have focused on temporal aspects of the spread of fake news. A 2019 study was able to combine several deep learning approaches, including LSTM and Bi-RNN models, and utilize only the timestamps of Tweets to learn propagation patterns and identify misinformation (Kotteti et al., 2019). Analysis of social media posts over time has proven to be particularly informative in periods of disease outbreaks – a collection of Tweets related to Ebola during one of the first months of the outbreak in 2014 displayed that a wide range of information about disease risk factors, prevention, and incidence trends was disseminated

to millions of users, demonstrating the strong influence of social media during public crises (Odlum & Yoon, 2015). Furthermore, a 2013 analysis utilized a sampling-based algorithm to gain insights about the temporal diversity of Tweets across various disease outbreaks, displaying that characteristics of Tweets over time were highly variable between different time periods and even reflected the severity and duration of different outbreaks (Kanhabua & Nejdl, 2013).

As well, associations between misinformation online and content engagement have become the subject of numerous analyses of media text and social networks, as high engagement is crucial to the potential impact of fake news and propaganda. In 2020, an investigation of over 500,000 Tweets related to COVID-19 found that online bots share proportionally more misinformation than human users. The study used a binary indicator of whether a Tweet had high or low engagement based on Likes and Retweets, and while user account metadata appeared more important than legitimacy in determining engagement level, Tweets containing misinformation still seemed to yield higher engagement than those that were legitimate (Silva et al., 2020). Other studies have noted that this phenomenon may be part of a self-reinforcing cycle: while the content of fake news and propaganda could play a role in increasing user interaction, the ability of other users to observe high engagement metrics of a post also appears to be associated with reduced fact-checking behaviors and increased likelihood of post interaction, indicating that there may be a critical threshold of engagement beyond which the content of a post has less influence on its propagation than the mere appearance of its popularity (Valenzuela et al., 2019; Avram et al., 2020). A time-series analysis of the use of identity-based language in spreading misinformative or controversial messages in China investigated "science factionalism," a phenomenon where individuals categorize and identify themselves with distinct factions that support or oppose a scientific issue. A positive association was observed between science factionalism and the likes or comments on a message, and usage of negativity in media messages also appeared associated with identity-based scientific discourse. These findings are particularly salient towards investigating potential relationships between COVID-19 misinformation and engagement, given that the pandemic has complex relationships with not only scientific communication but also politics, economic background, religious beliefs, and other personal identity characteristics (Chen et al., 2021).

3 Data

3.1 Collection

We utilized Twitter's Developer API to pull Tweets from May 11th, 2020 (when Twitter instituted a new mechanism for labeling Tweets with potentially misleading COVID-19 discourse) to December 19th, 2021. Tweets were obtained from well-known public figures and organizations, where "well-

known" was defined as possessing over 1 million followers. We used the Media Bias Chart from Ad Fontes Media (adf) to select accounts of interest, choosing roughly equal numbers along the chart's continuum of politically right/politically-left, reliable/unreliable quadrants. Accounts were labeled with a political lean of either "Right" (21 accounts), "Left" (20 accounts), or "Center" (6 accounts) based on their position on the Media Bias Chart. Several examples of chosen accounts are listed in Table 1, with the full list of accounts appearing in Appendix A.1.

Left	Center	Right
Wonkette	PBS News	Fox News
Huffpost	BBC News	Ben Shapiro
Rachel Maddow	CBS News	The Blaze

Table 1: Examples of Chosen Accounts with Political Lean

All Tweets were pulled from the chosen accounts between the start and end dates, provided that they contained a word from the list [COVID, COVID-19, coronavirus, vaccine] or a hashtag from the list [#COVID, #COVID19, #coronavirus, #vaccine]. In addition to Tweet text, we also collected Retweets, Replies, Likes, Quotes, creation time, and Tweet ID for each Tweet. A total of 94,693 Tweets were pulled on December 23rd, 2021. For each Tweet, corresponding Likes, Retweets, Replies, and Quotes were summed to yield a single response metric of Tweet engagement.

While it might be argued that Tweets with earlier creation dates will naturally have higher engagement metrics, the newest collected Tweets (those posted in mid-December 2021) were examined again in February 2022, and their engagement metrics appeared almost completely unchanged from December 2021 (with some metrics even slightly decreasing in a few cases). From this, we believe the shortest time interval between posting and collection (about 4 days) is still sufficient to capture the full engagement potential of a Tweet.

3.2 Data Processing and Sampling

In the text of each Tweet, URLs were replaced with the token "URL", and any tagged user was replaced with "@USER". Newlines and emoticons were also removed. Once these substitutions were performed, duplicate Tweets were removed from the dataset, yielding 65,168 Tweets.

For the purpose of manual labeling, we obtained a stratified random sample of Tweets from the set of 65,168 Tweets. The full timeframe (May 11th, 2020 to December 19th, 2021) was split into 10 intervals of roughly 2 months each, and within each interval, 2 Tweets were randomly selected from each account (if available). After this procedure, the final dataset consisted of 595 Tweets.

For usage in our misinformation classification model, further processing of the 595 Tweets included conversion to lowercase, replacement of more than three consecutive "@USER" tokens or spaces with a single "@USER" or single space (respectively), removal of the '#' character from hashtags, separating numbers from words (e.g. "Trump2020" becomes "Trump 2020"), and replacing contractions with full phrases (e.g. "don't" becomes "do not").

3.3 Manual Labeling

While Twitter implements an algorithm intended to provide labels like "Misleading information" or "Disputed claim" to questionable Tweets, this information is not included through the Developer API. Because of this, we manually labeled Tweets as "Misinformation," "Misleading," or "Valid" in addition to observing the Tweet on the Twitter platform to take note of whether Twitter added any labels. While none of the 595 Tweets were labeled by Twitter, many had been deleted since collection, which was also recorded.

Throughout the analysis, we classify and refer to Tweets as misleading, misinformation, or valid. We define "misleading" Tweets as those which present statements that are either disputed or cause readers to believe ideas that are false or defamatory, regardless of whether the explicit statements are true or false. We classify Tweets as "misinformation" when they present demonstrably false or inaccurate information, especially with the apparent intent to deceive. Under this paradigm, misleading Tweets are a superset of Tweets that are classified as misinformation. However, Tweets will be labeled with only their most severe level of deception, i.e. misinformation Tweets are considered "Misinformation" rather than both "Misinformation" and "Misleading." Misleading or misinformative Tweets often include one or more of the following propaganda techniques, derived from Martino et al. (2020): loaded language, name calling, exaggeration or minimization, appeal to fear, oversimplification, and appeal to authority. "Valid" Tweets are those which are determined to be neither misleading nor misinformation. Examples of each type of Tweet are given in Table 2.

	Label
Finally confirmed! Vitamin D pills proven to eradicate the risk of death from COVID-19!	Misinformation
We were told that the vaccine would 'stop the spread'...another example of #FauciLies	Misleading
The FDA has given emergency authorization to certain coronavirus diagnostic tests.	Valid

Table 2: Examples of Labeled Tweets

Figure 1 displays the distributions of several variables of interest. From 1A, we see that the majority of Tweets are Valid, followed by Misleading, then Misinformation. Within Misinformation

or Misleading Tweets, politically-right Tweets comprise the largest proportion, while politically-left Tweets comprise the largest proportion of Valid Tweets. From 1B and 1C, it appears that Misinformation and Misleading Tweets may be positively associated with Engagement, and partisan Tweets may be associated with slightly higher Engagement than centrist Tweets.

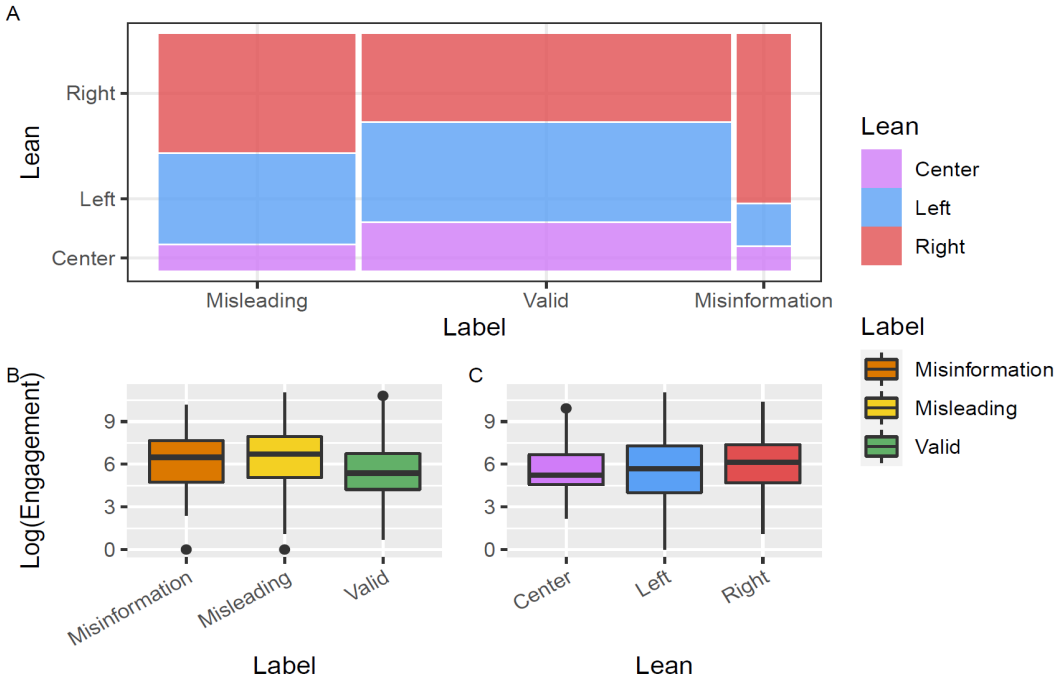


Figure 1: Visualizing Tweet Labels and Political Lean with Engagement

As a secondary variable of interest, Tweets were also labeled with whether they contained uncertainty – because virtually any statement can be logically correct as long as it contains a caveat like "might" or "maybe," we labeled Tweets which displayed expressions of hesitation or doubt, including words like "maybe," "might," "perhaps," "could," and "suggest." Phrasing statements as questions also fell into the uncertain category.

3.4 Other Variables

In addition to Tweet text, Likes, Retweets, Replies, Quotes, creation time, political lean, uncertainty, deletion status, and engagement, several other variables were calculated or derived for use in modelling. To account for several key periods of the COVID-19 pandemic, a categorical time variable was derived from the creation times of Tweets. There were 4 categories: before 200,000 COVID-19 deaths in the U.S. (Pre200D), after 200,000 COVID-19 deaths in the U.S. (200D), after the Johnson & Johnson vaccine was released and the U.S. surpassed 500,000 deaths (JJ500D), and after the Delta

variant became the dominant strain in the U.S. (Delta). We also utilized an indicator of whether a Tweet was a Retweet and/or contained a quote.

We created several variables based upon Tweet text. Lexical diversity was calculated using the Type-Token Ratio (TTR), giving the number of unique words (types) in a Tweet divided by the total number of words (tokens). A metric of entropy, defined as $-\sum_y y \times \log(y)$, was also calculated for each Tweet, where y takes on the relative frequency of each type found in the Tweet. Tweet readability was calculated using the Automated Readability Index (ARI) (Senter & Smith, 1967) – ARI is obtained as $\text{ARI} = 0.5\text{ASL} + 4.71\text{AWL} - 21.43$, defining ASL as average sentence length and AWL as average word length. We make an important distinction between readability and ARI (readability index) – ARI was roughly designed based on reading grade level, with higher ARI indicating a higher required grade level to read a text (note that ARI increases with sentence length and word length). Thus, lower text readability corresponds to higher readability index (i.e. higher education level required for comprehension) and vice versa.

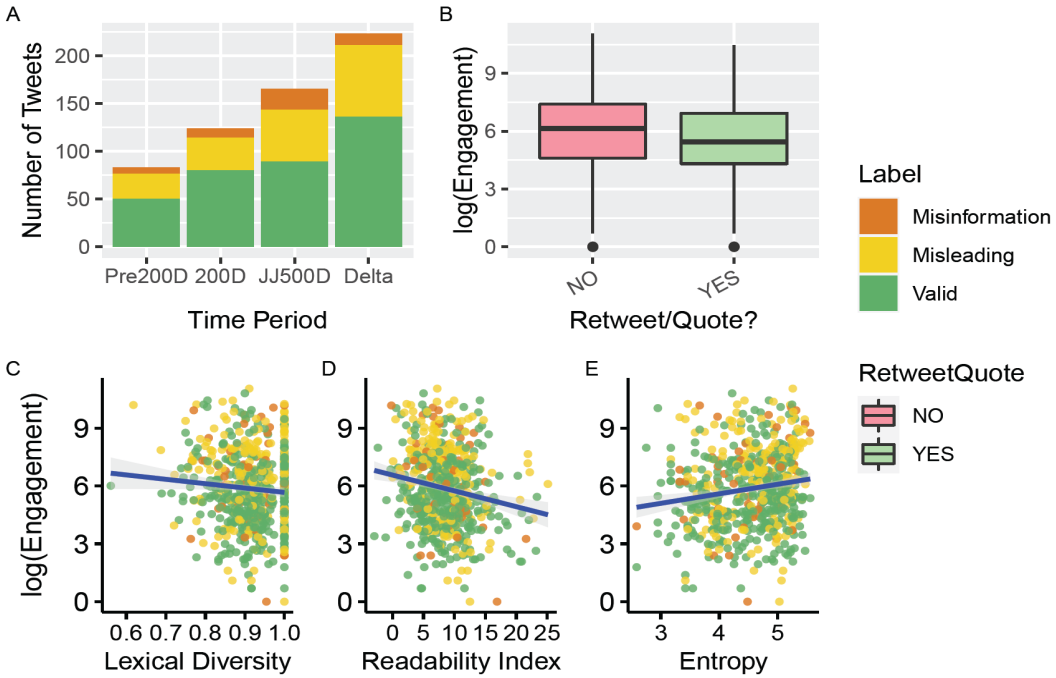


Figure 2: Visualizing Time Period, Retweet Status, and Text Features with Engagement

Several visualizations of these covariates are displayed in Figure 2. It appears that the number of Tweets overall increased across these 5-month periods of the pandemic, although the proportions of Valid, Misleading, and Misinformation Tweets remain similar. Tweets that are not Retweets and do not contain quotes seem potentially associated with higher Engagement. Finally, Engagement

may hold weak negative associations with lexical diversity and readability index, contrasting with a possible weak positive association with entropy.

3.5 Other Data

Besides our main dataset of 595 manually labeled Tweets, we also utilized a larger corpus of Tweets related to COVID-19 for the purpose of providing more training data for our misinformation classification model. The CMU-MisCov19 dataset (Memon & Carley) is a corpus of 4573 Tweets related to COVID-19 which are annotated with one of 17 different labels, including categories like Conspiracy, Fake Cure, False Fact or Prevention, Sarcasm/Satire, Politics, and Panic Buying. The Tweet IDs and labels are publicly available, and we were able to rehydrate the Tweet text for 3466 of the 4573 original examples. The CMU-MisCov19 Tweets underwent the same data preprocessing as our 595 manually labeled Tweets before being utilized during training of our classification model.

4 Methodology

4.1 Analyzing Engagement Metrics

Towards the aim of assessing how engagement metrics may be associated with various characteristics of Tweets, we will construct a Bayesian hierarchical negative binomial regression model with a log link function in order to understand the potential associations between misleading or misinformative Tweets and overall engagement while adjusting for potential confounding variables like political lean, time period, expressed uncertainty, and text features. We include a random intercept for Twitter account in addition to random slopes for time period based on account.

We believe the negative binomial distribution is an appropriate assumption for our response variable, as overall engagement is the sum of four discrete metrics (Likes, Retweets, Replies, and Quotes). While a Poisson regression model is another potentially suitable option, the response variable displays overdispersion with a mean of about 2,225 and variance over 33 million, which is problematic for the assumption of equality between mean and variance in a Poisson distribution. A negative binomial regression model has another free parameter to account for overdispersion, making it more suitable for our data. A multiple linear regression model may have been useful in modelling engagement – although the raw response is very right-skewed, taking the logarithm yields a more symmetric distribution that may be represented well by a normal distribution. However, linear regression has the potential to predict negative engagement, which is impossible in practice. Although our model is intended to be used mainly for interpretation rather than prediction, we believe that modelling engagement directly yields a more natural representation, which is better achieved by count distributions like the negative binomial distribution.

Engagement might also be modelled as a categorical response, with several levels corresponding to different amounts of engagement. The approach of Silva et al. (2020) even utilizes a binary response of only high versus low engagement. We believe a binary measure of engagement may be somewhat oversimplistic, especially for Tweets near the boundary between these two levels. While finer stratification may yield improvements (e.g., Very Low, Below Average, Average, Above Average, etc.), modelling counts directly will remove the need to create arbitrary thresholds between engagement levels and most accurately represent the engagement of each Tweet.

Regarding the hierarchical structure of our modelling approach, we aim to address the possible correlation in engagement for Tweets from the same account. We anticipate that different accounts may have different baseline levels of engagement due to follower counts and general exposure, influencing our decision to include random intercepts for account to address this source of random variability. See Appendix A.2 for visualization of the apparent clustering in Engagement based on Twitter account. Also, due to external factors such as fluctuating post frequency or changes in general media presence (e.g. controversy surrounding Joe Rogan in early 2022 based on his COVID-19 vaccine comments (Bond) and past utterances of racial slurs (D'Innocenzio)), we believe that the associations between time periods and engagement may be subject to random variability not accounted for by our confounders; thus, we incorporate random slopes for time periods based on accounts.

We now present the mathematical model formulation, adapted from Chapters 17 and 18 of *Bayes Rules!: An Introduction to Applied Bayesian Modeling* (Johnson et al., 2021). Let Y_{ij} represent the total engagement for Tweet i within account j, defined as $Y_{ij} = \text{Retweets}_{ij} + \text{Replies}_{ij} + \text{Likes}_{ij} + \text{Quotes}_{ij}$. We also define a covariance matrix for the joint multivariate normal distribution of random effects as follows:

$$\Sigma = \begin{pmatrix} \sigma_0^2 & \rho_{12}\sigma_0\sigma_1 & \rho_{13}\sigma_0\sigma_2 & \rho_{14}\sigma_0\sigma_3 \\ \rho_{12}\sigma_1\sigma_0 & \sigma_1^2 & \rho_{23}\sigma_1\sigma_2 & \rho_{24}\sigma_1\sigma_3 \\ \rho_{13}\sigma_2\sigma_0 & \rho_{23}\sigma_2\sigma_1 & \sigma_2^2 & \rho_{34}\sigma_2\sigma_3 \\ \rho_{14}\sigma_3\sigma_0 & \rho_{24}\sigma_3\sigma_1 & \rho_{34}\sigma_3\sigma_2 & \sigma_3^2 \end{pmatrix}$$

We will utilize a decomposition of Σ which consists of R (correlations between group-specific intercepts and slopes), τ (combined degree to which intercepts and slopes vary by group), and π (relative proportion of between-group variability stemming from varying intercepts vs. varying slopes), defined as follows:

$$R = \begin{pmatrix} 1 & \rho_{12} & \rho_{13} & \rho_{14} \\ \rho_{12} & 1 & \rho_{23} & \rho_{24} \\ \rho_{13} & \rho_{23} & 1 & \rho_{34} \\ \rho_{14} & \rho_{24} & \rho_{34} & 1 \end{pmatrix}$$

$$\tau = \sqrt{\sigma_0^2 + \sigma_1^2 + \sigma_2^2 + \sigma_3^2}$$

$$\pi = \begin{pmatrix} \pi_0 \\ \pi_1 \\ \pi_2 \\ \pi_3 \end{pmatrix} = \begin{pmatrix} \frac{\sigma_0^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2+\sigma_3^2} \\ \frac{\sigma_1^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2+\sigma_3^2} \\ \frac{\sigma_2^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2+\sigma_3^2} \\ \frac{\sigma_3^2}{\sigma_0^2+\sigma_1^2+\sigma_2^2+\sigma_3^2} \end{pmatrix}$$

so that $\Sigma = \text{diag}(\sigma_0, \sigma_1, \sigma_2, \sigma_3) \times R \times \text{diag}(\sigma_0, \sigma_1, \sigma_2, \sigma_3)$ and $\begin{pmatrix} \sigma_0 \\ \sigma_1 \\ \sigma_2 \\ \sigma_3 \end{pmatrix} = \tau\sqrt{\pi}$.

Below, we provide the complete model formulation.

$$Y_{ij}|\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j}, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14} \sim \text{NegBin}(\mu_{ij}, r)$$

$$\text{with } \log(\mu_{ij}) = \beta_{0j} + \beta_{1j}\mathbb{I}(\text{Time}_{ij} = \text{200D}) + \beta_{2j}\mathbb{I}(\text{Time}_{ij} = \text{JJ500D}) + \beta_{3j}\mathbb{I}(\text{Time}_{ij} = \text{Delta}) +$$

$$\beta_4\mathbb{I}(\text{Misinfo-or-Misleading}_{ij} = \text{Yes}) + \beta_5\mathbb{I}(\text{Deleted}_{ij} = \text{Yes}) + \beta_6\mathbb{I}(\text{Uncertain}_{ij} = \text{Yes}) +$$

$$\beta_7\mathbb{I}(\text{Retweet-or-Quote}_{ij} = \text{Yes}) + \beta_8\text{LexicalDiversity}_{ij} + \beta_9\text{ARI}_{ij} + \beta_{10}\text{Entropy}_{ij} +$$

$$\beta_{11}\mathbb{I}(\text{PoliticalLean}_{ij} = \text{Left}) + \beta_{12}\mathbb{I}(\text{PoliticalLean}_{ij} = \text{Right}) +$$

$$\beta_{13}\mathbb{I}(\text{PoliticalLean}_{ij} = \text{Left and Misinfo-or-Misleading}_{ij} = \text{Yes}) +$$

$$\beta_{14}\mathbb{I}(\text{PoliticalLean}_{ij} = \text{Right and Misinfo-or-Misleading}_{ij} = \text{Yes})$$

$$\begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix} |\beta_0, \beta_1, \beta_2, \beta_3, \sigma_0, \sigma_1, \sigma_2, \sigma_3 \sim N\left(\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}, \Sigma\right)$$

$\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14} \sim$ weakly informative priors (see Appendix A.3)

$$r \sim \text{Exp}(1)$$

$$R \sim \text{LKJ}(1)$$

$$\tau \sim \text{Gamma}(1, 1)$$

$$\pi \sim \text{Dirichlet}(4, 1)$$

For the assumptions of the negative binomial regression model, see Appendix A.4 for discussion of independence and the response distribution, as well as analysis of multicollinearity, linearity of covariates with (logged) engagement, and model fit through generalized variance inflation factors (GVIF), Cook's distance, leverage, and residual plots. We also examine model diagnostics such as traceplots, ACF plots, and posterior predictive distributions in Appendix A.5 to assess mixing and goodness-of-fit.

We also perform a model sensitivity analysis in order to investigate the influence of Tweets from Joe Rogan, as we encountered some difficulty in classifying the political affiliation of his account. Joe Rogan is a media commentator who hosts a podcast centered around politics, philosophy, and current events, and because he holds some opinions that are considered far-right along with opinions that are considered far-left, his political leaning is somewhat unclear. See Appendix A.6 for the sensitivity analysis involving experimentation with his political classification as well as a numeric estimation of his political beliefs using the *tweetscores* R package.

4.2 Classifying Tweets with Misinformation

Effective rhetorical tactics are commonly used to spread misinformation and misleading sentiments, including name-calling, oversimplification, and appeals to fear (Martino et al., 2020). From this, we aim to test the hypothesis that Tweets likely containing misinformative content can be recognized and classified by language models based on the text alone, without direct access to any particular knowledge base.

4.2.1 Task

Our task can be formally described as 3-label classification which determines whether a Tweet contains misinformation, contains misleading content, or neither.

Task: Identifying Misinformation or Misleading Sentiment in Tweets

Does the Tweet contain misinformation (MIS), statements that are misleading (MIL), or neither (VAL)?

- **MIS**: statements with demonstrably false or inaccurate information, especially with the apparent intent to deceive.
- **MIL**: statements which are either disputed or cause readers to believe ideas that are false or defamatory.

- **VAL**: statements that contain neither misinformation nor misleading content.

Examples of MIS (Misinformation), MIL (Misleading), and VAL (Valid) Tweets can be found in Table 2.

4.2.2 Approach

To classify Tweets into Misinformation, Misleading, or Valid, we utilize the DeBERTa neural language model, extending the BERT (Bidirectional Encoder Representations from Transformers) model to add "a disentangled attention mechanism and an enhanced mask decoder" (He et al., 2021).

Broadly, transformer networks convert a text sequence into numeric vectors which represent the words in the text as well as their positioning in the original sequence. These word embeddings and position encodings are passed through a series of encoders which calculate new representations of the input. In encoder-decoder transformers, the last encoder output is passed through several decoders, which convert the representation into the final output, such as a label for each token in a sequence. However, for classification with BERT/DeBERTa, the transformer consists only of encoders, converting text input into contextual representations to produce a single class prediction. Transformer networks also use an attention mechanism, a process by which the network can "focus" on the most informative features of the text. This procedure utilizes hidden states of the encoders and decoders to create a context vector, or weighted average of encoder states, which is then used to create modified decoder states that magnify significant features of the input and minimize less significant features (Vaswani et al., 2019; Wood).

Devlin, Chang, Lee, and Toutanova (2019) present BERT as a transformer network that applies attention and a multi-layer bidirectional encoder to learn both left and right contexts of each word in a text. It stacks a number L of encoders that pass input through multi-head attention and a feed-forward neural network. Inputs contain tokens [CLS] and [SEP] at the beginning and end of each sequence, respectively (e.g. a Tweet of "coronavirus is a hoax" becomes {[CLS], coronavirus, is, a, hoax, [SEP]}). Outputs are vectors with a prediction for each input example (Devlin et al., 2019).

For BERT, predicted labels are defined as follows, using $BERT_{out}$ (BERT's output vector representation of classification token [CLS]), a normalization layer η, and a softmax layer σ:

$$\phi = \sigma(\boldsymbol{W}^s \cdot \eta(BERT_{out}) + b^s)$$

In the previous equation, ϕ is the predicted label, $\boldsymbol{W}^s \in \mathbf{R}^{N_c \times B}$ and $b^s \in \mathbf{R}^{N_c}$ are parameters to the softmax layer, N_c is the size of the model output (corresponding to the number of possible labels analyzed for the model), and B is the dimension of the BERT output.

According to He, Liu, Gao, and Chen (2021), DeBERTa leverages BERT with a disentangled attention mechanism: rather than utilizing a vector that is the sum of the word embedding and position embedding, two separate vectors encoding words and positions, respectively, are used to represent words. Attention is then calculated with disentangled matrices based upon the word embeddings and positions. Like BERT, DeBERTa is pre-trained using mask language modeling, but it incorporates absolute positioning of words in addition to relative positions (He et al., 2021).

We also utilize the cross entropy loss function below, where y_i is the true label of Tweet i:

$$L = -\sum_i y_i \cdot \log(\phi_i)$$

4.2.3 EXTENSIONS

In addition to training and testing the DeBERTa model on our dataset of 595 Tweets, we incorporate several extensions in an attempt to maximize performance.

First, we utilize the CMU-MisCov19 dataset (Memon & Carley), which is a corpus of 4573 Tweets related to COVID-19 misinformation. There are 17 different labels used to annotate these Tweets. Conspiracy, Fake Cure, False Fact or Prevention, Sarcasm/Satire, Politics, and Panic Buying are several of these categories. 3466 of the 4573 original examples could be retrieved from Twitter. We implement sequential fine-tuning, first training on the CMU-MisCov19 data before training on our dataset. Since our original dataset of 595 examples is fairly small, we believe that the CMU-MisCov19 corpus will provide a wider variety of misleading Tweets, allowing the model to learn characteristics of misinformation more accurately and generalize to a broad range of social media text.

Continuing with the CMU-MisCov19 corpus, we also implement a multitasking training procedure, providing the model with batches from the CMU-MisCov19 data and our dataset in an alternating order. While we anticipate improvement in performance when first training on the CMU-MisCov19 data and then on our smaller dataset, a training procedure which alternates batches between CMU-MisCov19 and our data might prevent the model from overfitting on either dataset and improve its ability to generalize to new Tweets.

We also incorporate backtranslation as a data augmentation procedure: each Tweet in our training set is translated to French, and the French Tweet is translated back to English, ideally producing a new Tweet with the same sentiments but slightly altered language. As our dataset is fairly small at only 595 Tweets, generating synthetic data through backtranslation will provide the model with more training examples without requiring more manual labeling. Regarding the choice of an intermediate language, we noted that French possesses some similarities with English, and English-French trans-

lation tools are fairly advanced. However, French is distinct enough that translation back to English would likely yield several differences from the original English text, making French suitable for the goal of creating new augmented inputs. The backtranslation procedure will leverage our original small dataset to create "new" Tweets that allow the model to learn from a wider variety of text.

We implemented weighted loss to adjust the importance assigned to each class: labels of misleading and misinformation were less common than valid labels, but we believe false negatives (i.e. failing to classify a misinformative Tweet as such) are potentially more harmful than false positives. This motivated our decision to assign the highest weight to the misinformation label and the lowest weight to the valid label. Exact weight values were calculated based on the inverse frequency of each label in the training set. Finally, we tuned dropout, learning rates, and weight decay for optimal performance.

4.2.4 Datasets and Evaluation

For model training, we will utilize both our dataset of 595 Tweets and the CMU-MisCov19 dataset. Data cleaning, preprocessing, and examples are described in Section 3, and the data distributions are shown in Table 3. As we are more interested in the wide range of Tweets provided by the CMU-MisCov19 corpus rather than the final model performance for CMU-MisCov19, we provide only training, validation, and testing totals for the CMU-MisCov19 dataset, rather than breakdowns by label.

Class	Training	Validation	Testing
Misinformation	42	5	4
Misleading	150	12	27
Valid	289	37	29
(CMU-MisCov19)	2807	312	347

Table 3: DeBERTa Model Data Distributions

For the backtranslation procedure, we translated only the training set from our original dataset (rather than the training set for CMU-MisCov19) into French and back to English using the *DeepL* Python library. If the backtranslated English Tweet was identical to its original English form, it was not added to the training set. After adding the (non-duplicate) backtranslated Tweets, there were 82, 291, and 559 Tweets that were Misinformation, Misleading, and Valid, respectively, in our training set. We utilize accuracy and macro-F1 score to assess performance.

4.2.5 Model and Training Details

We utilized the pre-trained 'deberta-base' DeBERTa model from the Hugging Face library[1], and we used the DeBERTa tokenizer with padding and a maximum length of 256. We pre-processed and tokenized the Tweets as described in Section 3, and we provided the corresponding token IDs, attention masks, and labels to the model. Class labels were transformed with an ordinal encoder. During training, we used an AdamW optimizer, and gradients were clipped to 1. We trained the model for a maximum of 10 epochs with early termination if the validation F1 score began to decrease. Learning rates, dropout rates, and weight decay rates, as well as exact weights utilized for the weighted loss, can be found in Appendix A.7.

We will first report performance on a majority class baseline, and we will then report performance for the model trained with the following:

- Only our original dataset
- Sequential fine-tuning on the CMU-MisCov19 dataset, then our dataset
- Multitasking with the CMU-MisCov19 dataset and our dataset
- Multitasking and incorporating backtranslated Tweets
- Multitasking, backtranslated Tweets, and weighted loss
- Sequential fine-tuning, backtranslated Tweets, and weighted loss

5 Results

5.1 Negative Binomial Regression

The results of the negative binomial regression model are displayed in Table 4. Full model output, including estimates for all random intercepts and random slopes, can be found in Appendix A.8.

We do not observe strong evidence of differences in Tweet engagement across various time periods of the pandemic, but for a Tweet which is misleading or contains misinformation, we may expect its engagement to multiply by about 1.888 as compared to a valid Tweet, holding all other factors constant. While the 95% credible interval of (-0.020, 1.351) includes 0, we note that the vast majority of probability density is concentrated in the positive range.

[1] https://huggingface.co/

Covariate	Estimate	95% Credible Interval
Intercept	5.068	(2.755, 7.400)
Time = 200D	0.094	(-0.316, 0.509)
Time = JJ500D	-0.143	(-0.567, 0.272)
Time = Delta	-0.253	(-0.683, 0.167)
Misinfo-or-Misleading = Yes	**0.636**	**(-0.020, 1.351)**
Deleted = Yes	0.132	(-0.062, 0.327)
Uncertain = Yes	-0.035	(-0.280, 0.220)
Retweet-or-Quote = Yes	**-0.345**	**(-0.563, -0.129)**
Lexical Diversity	1.065	(-0.429, 2.529)
Readability Index (ARI)	**-0.040**	**(-0.067, -0.012)**
Entropy	0.038	(-0.187, 0.257)
Political Lean = Left	0.644	(-0.899, 2.231)
Political Lean = Right	0.854	(-0.621, 2.338)
Political Lean = Left and Misinfo-or-Misleading = Yes	**-0.894**	**(-1.695, -0.151)**
Political Lean = Right and Misinfo-or-Misleading = Yes	-0.157	(-0.917, 0.571)

Table 4: Negative Binomial Regression Model Results (Unexponentiated)

Political Lean	Estimate	95% Credible Interval
Center (β_M)	0.636	(-0.020, 1.351)
Left ($\beta_M + \beta_{LM}$)	-0.253	(-0.607, 0.107)
Right ($\beta_M + \beta_{RM}$)	**0.483**	**(0.175, 0.797)**

Table 5: Associations Between Misleading/Misinfo and Engagement By Political Lean

Regarding other features such as whether a Tweet has been deleted or whether it contains uncertain language, we see that these 95% credible intervals contain 0 and do not give strong evidence of associations with engagement, but for Tweets that are Retweets or contain quotes from others, engagement is expected to differ from original Tweets with no quotes by a factor of 0.709, holding all else equal, and the 95% credible interval of (-0.563, -0.129) does not include 0. Text characteristics of lexical diversity and entropy did not show evidence of associations with Tweet engagement; however, an increase of 1 unit in readability index appears to be associated with a multiplicative change of 0.961 in engagement with all other variables held constant, with a 95% credible interval (-0.067, -0.012) entirely below 0.

Neither right nor left political leanings appear to have strong evidence of differences from centrist Tweets in engagement metrics. There is also little evidence that the association between engagement and misleading or misinformation Tweets differs for right-leaning vs. centrist Tweets. However, while coefficient estimates for misleading/misinformative status and left political lean are positive, we see that the interaction between these attributes has a negative coefficient, with a 95% credible interval of (-1.695, -0.151) that does not include 0. This provides evidence that the association between engagement and whether a Tweet is misleading or contains misinformation for left-leaning Tweets is more negative the same association for centrist Tweets. Using the interac-

tion effect, we decompose the association between engagement and the misleading/misinformation indicator for all political leanings in Table 5, denoting β_M, β_{LM}, and β_{RM} as the coefficients for the misleading/misinformation indicator, the interaction between left political lean and misleading/misinformation, and the interaction between right political lean and misleading/misinformation, respectively.

We see that while the negative interaction effect between misleading or misinformative Tweets and left political lean indicated a lower association between engagement and misleading status for left Tweets than for centrist Tweets, there is little evidence that this association is different from 0 within left-leaning Tweets. However, within right-leaning Tweets, there is evidence that engagement for a misleading or misinformative Tweet is expected to multiply by 1.620 as compared to legitimate (right-leaning) Tweets, holding all else equal.

5.2 Misinformation Classification

The results for the DeBERTa classification model, as well as the baseline, are reported in Table 6.

From the results, we see that all DeBERTa models showed improvement over the baseline majority class, and the combination of multitask fine-tuning, backtranslation, and weighted loss yielded the highest performance with an accuracy of 0.683 and macro-F1 score of 0.593. Interestingly, multitasking with or without backtranslated Tweets did not perform as well as sequential fine-tuning first on CMU-MisCov19 and then on our dataset, but multitasking with backtranslated Tweets and weighted loss was able to exceed sequential fine-tuning performance.

Model	Accuracy	Macro F1 Score
Baseline Majority Class	0.483	0.217
Original 595 Tweets	0.617	0.535
CMU-MisCov19, Then Original 595 Tweets	0.667	0.574
Multitasking	0.635	0.418
Multitasking, Backtranslation	0.667	0.463
Multitasking, Backtranslation, Weighted Loss	0.683	0.593
CMU-MisCov19, Then Original 595 Tweets + Backtranslation + Weighted Loss	0.677	0.553

Table 6: Misinformation Classification Model Results

6 Discussion

6.1 Engagement Analysis Findings

From our analysis of factors that are associated with Tweet engagement metrics, we observed evidence that readability index, whether a Tweet contains misinformation or misleading content, and

whether the Tweet is quoting another entity (either with quotations or via a Retweet) are associated with overall Tweet engagement. As well, for Tweets with a left political lean, the direction of the association between misleading/misinformative content and engagement appears to change from the direction of this association for centrist Tweets.

The most salient result is that of misleading or misinformative centrist Tweets – these Tweets are expected to have engagement 1.888 times that of valid centrist Tweets, holding all else constant. Because misinformation often utilizes rhetorical tactics that attract readers or grab attention (e.g. emotional appeal, bandwagon, loaded language, appeal to authority, etc.), this result may reflect that these strategies succeed in garnering more user interaction. While most factual information is often shared with the intent of providing the truth without a particular agenda in mind, misleading information is generally the opposite: the truth should be spun or obscured in pursuit of a certain political, social, or business incentive. Given this common aim of misrepresentative Tweets, it may be expected (and concerning) that these Tweets will receive higher engagement and continue to spread an ideology as far as possible.

A Retweet or Tweet that contains a quote appears to be associated with decreasing engagement, as does high readability index. For Retweets or direct quotes, users might prefer to go directly to the original Tweet or find more information about a quote from a news article rather than interacting with a second-hand version, resulting in lower engagement for these Tweets. Noting that higher readability index corresponds to lower readability (as it indicates a higher education level required to comprehend the text), the negative association between ARI and engagement corresponds to a positive association between text readability and engagement. Users may be more likely to engage with Tweets that they can view and understand quickly, as opposed to Tweets with more specialized and complicated syntax, perhaps explaining the apparent increase in engagement as Tweet readability increases and readability index decreases.

It is particularly intriguing that the interaction between political lean and misinformation status indicated that, holding all else constant, right-leaning and center Tweets are expected to see a positive change in engagement if they include misleading or misinformative messages, while a more negative trend is expected for left-leaning Tweets. While all three political sectors produce Tweets that can be considered misleading or misinformation, different reasoning may explain why it appears advantageous for some groups and not others. The centrist group had the lowest proportion of misleading/misinformation Tweets, implying they may be less common within this political sphere. Thus, they may garner more attention when they are released due to their peculiarity, or because some users may be more inclined to trust relatively non-partisan sources. Furthermore, when considering sources of misinformation in media, right-leaning commentators might tend to be more

well-known and brash, with higher shock value (e.g. Alex Jones insisting "I have the government documents where they said they're going to encourage homosexuality with chemicals so that people don't have children" (Hananoki), Tucker Carlson asserting that "[COVID-19] does feminize people. No one ever says that but it's true" (Baragona), etc.). Combined with the finding from a 2021 analysis that right-leaning Twitter users had slightly higher vulnerability to misinformation (Nikolov et al., 2021), this might explain why including misinformative or misleading sentiments appears to be an advantageous tactic for the right rather than the left.

We did not find evidence that Tweets in later time periods (after 200,000 COVID-19 deaths in the U.S., after Johnson & Johnson released a vaccine and the U.S. surpassed 500,000 COVID-19 deaths, or after the Delta variant became the dominant strain in the U.S.) had different engagement metrics than Tweets from the earliest time period (before 200,000 COVID-19 deaths in the U.S.). This is somewhat surprising, as we might expect earlier Tweets to have higher engagement due to the novelty of the pandemic. Conversely, one might argue that later Tweets could have higher engagement as panic grew from the rising death toll, stringent restrictions on activities, and rapidly emerging variants. Both these arguments may be useful in speculating why there is no notable difference over time – near the beginning of the pandemic, curiosity and uncertainty may have driven engagement, whereas concern over the worsening outlook, unrest, and pandemic fatigue became the motivators that maintained this engagement as time continued, and thus there were no time periods with much lower or higher engagement than the earliest time period. Alternatively, our earliest time period is in May 2020, about 6 months after the pandemic began – if we examined Tweets posted since the first discovery of COVID-19, perhaps more differences over time would have become apparent.

Deleted Tweets and Tweets that contained language of uncertainty did not show strong evidence of associations with the response as compared to currently available Tweets or Tweets without hesitant language, respectively. We might expect deleted Tweets to have less engagement since they can no longer attract more interaction after being removed. However, as discussed in Data, it appeared that a period of only about 4 days was enough for a Tweet to reach its full engagement potential, and since many deleted Tweets were removed after this threshold, it is not surprising that deletion did not display a relationship with engagement. Furthermore, while uncertainty can be characteristic of misleading messages, it is unlikely that users consider the hesitancy in a Tweet's tone before deciding to interact with it, which may explain the lack of association between uncertainty and engagement.

We also did not find evidence that the other text features, lexical diversity and entropy, held any associations with overall engagement. As Twitter is a casual social media application for most users, it might be that lexical diversity and entropy (a measure of how much information is contained

per token) are relatively unimportant to those scrolling through their Twitter feeds. If users enjoyed these features, books or news articles would be more efficient sources; therefore, it is likely that other factors, such as eye-catching keywords, images, or personal connection, may play larger roles in users' decisions to engage with Tweets.

Partisan Tweets did not display evidence of differences in engagement as compared to centrist Tweets. This is somewhat unexpected, since Democrats/left-leaning individuals make up 69% of the top 10% of Tweeters on the website, while Republicans/right-leaning individuals make up only 26% (pew). It might be that while a larger portion of Tweets and total engagement on the website stem from left-leaning accounts due to their majority, there are still sufficient audiences for both political affiliations so that engagement **per Tweet** is roughly balanced, at least in comparison to Tweets from accounts near the political middle.

In Appendix A.6, we performed a sensitivity analysis that involved altering the political categorization of Joe Rogan to address the highly polarized but heavily fluctuating nature of his content and examine how this influenced model results. When Rogan's political leaning is listed as Right or Left, we found that although the credible intervals for the interaction between left lean and misinformative/misleading Tweets shifted to include 0, other estimates and credible intervals were largely equivalent to when his politics are listed as Center. This indicates that the political affiliation for his 15 Tweets did not have a substantial impact on model results, despite his relatively high number of misleading/misinformative Tweets. We also used the *tweetscores* package to estimate Joe Rogan's political ideology using the political "elites" he followed with a Bayesian latent spatial model to represent the social network between users and political accounts. Methods of Metropolis sampling and maximum likelihood estimation yielded a substantial left lean (-0.87) for Rogan, while a correspondence analysis with an expanded set of political elites yielded a substantial right lean of equivalent magnitude (+0.87). While not incredibly useful for classifying Joe Rogan under our schema, this ideology estimate supported the idea that Rogan's political views are quite difficult to categorize, and our analysis showed that our model was not very sensitive to the political affiliation of his Tweets.

6.2 DeBERTa Model Findings

Our DeBERTa classification results showed that all models were able to exceed the performance of the majority class baseline, indicating that the model is able to successfully learn certain linguistic features of Tweets with misinformation or misleading content. It seems that sequentially fine-tuning first on CMU-MisCov19 data and then on our original dataset yielded higher performance than multitask training – since we report test performance only on our dataset, it is possible that training on our dataset all at once (after the CMU-MisCov19 dataset) better prepared the model for test exam-

ples from our data, while multitasking provided greater ability to generalize but lower performance on our test set. Including backtranslated Tweets in the training set appears to increase performance overall, as the model is exposed to a wider range of linguistic variability, and this method did not require any additional manual labeling.

Weighted loss also improved model predictions – with unweighted loss, we noted that the model almost never predicted any misinformation labels, likely due to their scarcity in the training data. However, weighted loss with higher penalties for misleading and misinformation labels provided an incentive for these classes to be predicted correctly, yielding higher performance especially regarding macro-F1 score. We see that the model using multitask fine-tuning, backtranslated Tweets, and weighted loss achieved the highest performance, with an accuracy of 0.683 and a macro F1-score of 0.593.

Next, we analyze the explainability of the highest-performing DeBERTa model using the Python library *transformers-interpret*, which utilizes integrated gradients to compute token attributions for the input. The package first leverages the model tokenizer to parse the input text, then accumulates gradients along each dimension on the straightline path between a given baseline (often the zero vector) and the input vector, providing a single attribution score for each token. The attribution score may be interpreted as a token's relative influence on the assignment of a particular label (Sundararajan et al., 2017). We note that the efficacy of the integrated gradients method is somewhat debated: firstly, it does not necessarily account for interactions between tokens, and it has also been shown to display somewhat counterintuitive behavior (Hase & Bansal, 2020; Adebayo et al., 2018). However, we believe that the methods are still reasonably useful, if not for a rigorous quantitative investigation, then for a qualitative analysis of tokens with high importance within the model.

First, we examine attribution scores of three examples which the model predicted correctly, and then we give scores for three examples which were predicted incorrectly.

Correctly Labeled Tweets

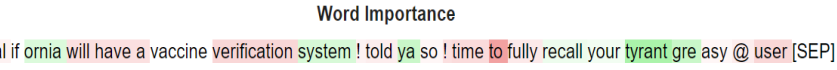
Legend: Negative Neutral Positive

1A. True Label: Misinformation (Correct): "RT @USER: in Toronto! But this could happen anywhere. COVID has now become the pretext for unlimited government control over our lives"

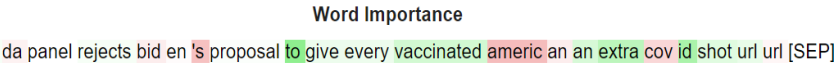
Word Importance

[CLS] rt @ user : in tor onto ! but this could happen anywhere . cov id has now become the pretext for unlimited government control over our lives [SEP]

1B. True Label: Misleading (Correct): "California will have a vaccine verification system! Told ya so! Time to fully recall your tyrant greasy @USER"



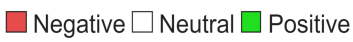
1C. True Label: Valid (Correct): "FDA panel rejects Biden's proposal to give every vaccinated American an extra COVID shot URL URL"



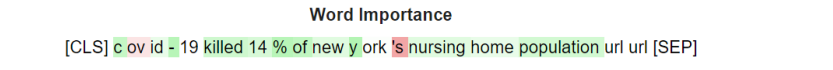
From the correctly labeled examples, we can observe the relative magnitudes of token attribution scores and roughly determine which tokens contributed to (or conflicted with) the assigned label. For the Tweet correctly labeled Misinformation, it appears that the words "pretext," "unlimited," and "government" contribute positively to the Misinformation label – "pretext" is defined as a purpose or motive presented to conceal true intentions, which is strongly connected to misinformation. "Unlimited" is a non-specific descriptor that is difficult to verify, and propaganda messages are commonly related to issues in government, providing a practical explanation for the high positive influence of these tokens on the Misinformation label.

For the Misleading Tweet, "tyrant" as well as the first tokenized syllable of "greasy" contributed positively to the label, perhaps explained by the loaded and inflammatory connotation of these words. "Verification" has a slight negative contribution to the misleading label, perhaps because verification and fact-checking are unlikely to be encouraged in misleading (or misinformative) Tweets. Finally, in the correctly-labeled Valid Tweet, "vaccinated" and "extra" possess a positive attribution towards the Valid label – claims about "vaccinated" individuals or providing an "extra" item might be easier to verify than general claims about vaccines or "unlimited" government control (see Misinformation Tweet above), which may increase their prevalence in legitimate Tweets rather than misleading or misinformative ones.

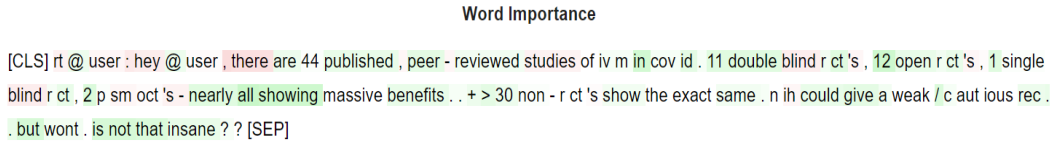
Incorrectly Labeled Tweets



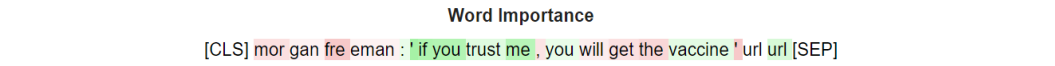
2A. True Label: Misinformation (Incorrectly Labeled "Valid"): "COVID-19 killed 14% of New York's nursing home population URL URL"



2B. True Label: Misleading (Incorrectly Labeled "Valid"): "RT @USER: Hey @USER, There are 44 published, peer-reviewed studies of IVM in COVID. 11 Double blind RCT's, 12 open RCT's, 1 single blind RCT, 2 PSM OCT's- nearly all showing MASSIVE benefits..+ >30 non-RCT's show the exact same. NIH could give a weak/cautious rec.. but wont. Isn't that insane??"



2C. True Label: Valid (Incorrectly Labeled "Misleading"): "Morgan Freeman: 'If you trust me, you will get the vaccine' URL URL"



From the incorrectly labeled examples, we can observe roughly which tokens were most influential in the model's incorrect label assignment. For the Misinformation Tweet that was incorrectly labeled "Valid," it appears that "killed," "14% of," and "population" contributed positively to the "Valid" label – this is somewhat understandable, since definitive statements like "killed" and numeric figures about populations are easier to verify than more general statements and are perhaps less likely to appear in misleading Tweets. During labeling, the figure provided in this Tweet was determined to be incorrect, but because the DeBERTa model does not have access to a knowledge base, it is not surprising that it found nothing particularly indicative of deception or dishonesty in this text.

For the Misleading Tweet incorrectly labeled "Valid," it appears that numeric values like "11", "12", and "2," as well as the phrase "all showing," contributed positively to the "Valid" label. Similar to the Misinformation Tweet incorrectly labeled "Valid," numeric figures and definitive phrases like "all showing" may be less likely to appear in Tweets that intend to deceive readers, and because the model cannot check the provided numbers directly, it assigned the "Valid" label to this Tweet. Finally, the Valid Tweet incorrectly labeled "Misleading" displays that "if you trust me" and "vaccine" contribute positively to the Misleading label. A phrase like "If you trust me, you will..." contains emotional tactics such as guilt or ultimatums that are commonly used in misleading or manipulative content – while this Tweet was manually labeled "Valid" because it is a clear quote from another individual (i.e. the emotional sentiment is not necessarily from the account which posted the Tweet, but only from the quoted individual), it is understandable that this phrase contributes to a misleading label. Also, many misleading or misinformative Tweets in this dataset were related to statements

about vaccines, making it possible for the token “vaccine” to gain a misleading connotation within the model.

These depictions of word attributions for example inputs display that the model is providing largely understandable labels for Tweets in our dataset, as words or tokens with high-magnitude attribution scores often possess meanings which correspond to the assigned label (regardless of its correctness). The model seems to miss phrases that may be helpful in longer Tweets – for example, “Isn’t that insane??” in 2B is loaded language that should have contributed more positively to a Misleading label, rather than an incorrect Valid label. A knowledge base of facts and numeric figures about COVID-19 may have also been beneficial in allowing the model to recognize Tweets that did not possess overtly manipulative language but presented incorrect information.

We can also visualize attention heads in the model to examine which tokens are receiving the most attention. In Figure 3, we can observe attention from the [CLS] token in the final model layer for Tweet 2A: “COVID-19 killed 14% of New York’s nursing home population URL URL.” There are 12 attention heads, each represented by a different color. The attention from the [CLS] token is concentrated on the first character of ’COVID-19,’ the phrase ’killed 14%’, and the [SEP] token, similar to the findings of the token attributions – this supports the observation that these tokens contributed highly to the model prediction.

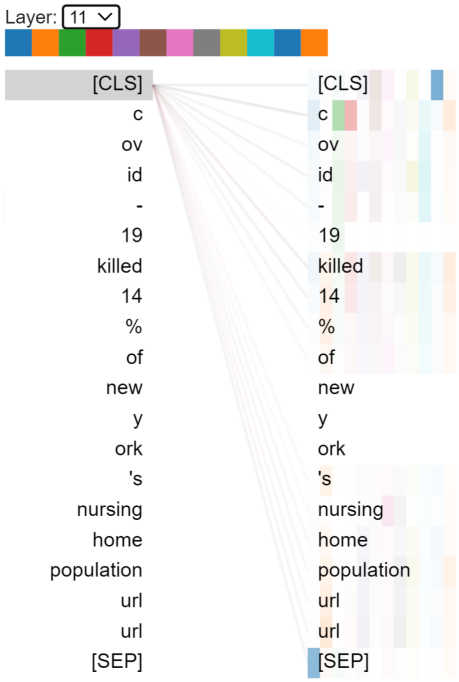


Figure 3: Attention from [CLS] Token

Integrated gradients and attention visualizations are useful methods for investigating the relative importance of each token in a certain model prediction, and further analysis might involve examining attention across multiple layers, text perturbation, or adversarial examples to draw conclusions about model sensitivity to changing inputs.

6.3 Limitations and Future Work

One limitation of this study was manual labeling and the relatively small dataset used for the negative binomial regression and DeBERTa models – a larger sample size may have yielded better classification performance and inference. Also, while specific criteria were used for labeling Misinformation,

Misleading, and Valid Tweets, there is still some subjectivity inherent in deciding a Tweet's group based upon a single judgment. In the future, it may be advantageous to crowdsource multiple labelers (e.g. through Amazon Mechanical Turk) or even leverage the Twitter API to retrieve Twitter's misinformation labels directly. Although the website's own labels were not available through the API at the time of this study (and no Tweets were seen to have any such labels when they were checked manually on a web browser), this may provide an avenue to automate the collection of misinformative Tweets and gain a much larger sample. Also, perhaps relabeling existing COVID-19 misinformation datasets (such as CMU-MisCov19) with our own labeling schema may have also resulted in a more diverse range of Tweets, although the labeling process would be intensive.

For the analysis of Tweet engagement, the representation of time as categorical and the aggregate metric of engagement remove the ability to see certain details in change over time or the different components of engagement, respectively – it may be worthwhile to explore time as a continuous measurement across the overall timeframe to examine whether the same trends apply to continuous and categorical time. Fitting separate models for Likes, Retweets, Replies, and Quotes may also illuminate whether certain covariates have differential associations depending on the exact mode of engagement. Also, the DeBERTa classification model only utilized Tweet text as input – while this may be ideal for a model that is intended to be used for immediately classifying Tweets as they are posted, it is possible that other extralinguistic features like deletion status, political lean, or creation time may have associations with whether a Tweet is misinformation, misleading, or valid. Future experiments might involve concatenating vector representations of these features with word/position embedding vectors of Tweet text to investigate whether the DeBERTa model can leverage these characteristics.

From our sensitivity analysis based on Joe Rogan's Tweets, we believe it may be useful to experiment with political ideology as a numeric continuum rather than a 3-class schema of Left, Center, and Right. We believed a categorical political lean was advantageous for interpretability, and nearly all of our accounts were fairly clear in their political affiliation, suggesting that political categories would be sufficient. However, for Joe Rogan (or private individuals with less involvement in politics), political leaning is much more variable and difficult to classify, and a numeric estimate of ideology may be preferable. While the *tweetscores* package is helpful for estimating political standing based on followers, several of our accounts did not follow any political elites and thus could not receive an estimate from this package, and current tools that measure political leaning based on Tweets themselves (e.g. software from the Polarization Lab at Duke University) require users to log in to their account to obtain an estimate. Especially for private individual accounts and politically-ambiguous actors, we believe that the expansion of political estimation software will be crucial towards providing more accurate analysis of a broad range of social media sources.

A sociological limitation of this research is that it is likely ill-suited for generalization towards different languages or countries. We utilized only English Tweets from public figures or news organizations that are prominent in the United States. However, other countries may have very different rhetoric surrounding COVID-19 and vastly different political systems that cannot be easily separated into distinct parties. Text features like readability, lexical diversity, and uncertainty may present differently in other languages as well, which could result in different findings from those for English Tweets, and the notable time periods of the pandemic in the U.S. may not coincide with those in other nations. In the future, it may be helpful to include Tweets from multiple languages and broaden covariates relating to time and politics to avoid excessive cultural specificity and better encapsulate global phenomenons.

7 Conclusion

By obtaining a sample of Tweets related to COVID-19 from news organizations and influential political media commentators, we obtained several insights about factors associated with overall engagement and methods that can assist with the classification of misinformation in social media posts.

Firstly, we observed evidence of a positive association between engagement and whether a Tweet is misinformative/misleading (for centrist and right-leaning Tweets), while we found that engagement displayed negative associations with readability index and whether the Tweet was a Retweet (or contained quotes). Although we did not see strong evidence that political lean was associated with engagement, it appeared that the interaction between left political lean and misleading/misinformative content had a negative coefficient that yielded a much lower, if not negative, association between misleading content and engagement compared to the positive associations seen for centrist and right-leaning Tweets. From our sensitivity analysis on the political affiliation of Joe Rogan, we found that his categorization was not highly influential for model results, although quantitative estimates of his political ideology from the *tweetscores* R package supported our observation that he is politically ambiguous. By utilizing a DeBERTa model to classify Tweets into misinformation, misleading, and legitimate, we found that the model is able to learn linguistic features of deceptive Tweets without access to any external knowledge base. Implementation of other procedures such as fine-tuning on a related dataset, multitask training, backtranslation, and weighted loss are able to improve model performance as well, with the combination of multitask fine-tuning, backtranslation, and weighted loss obtaining the highest performance overall. From our analysis of DeBERTa explainability through word attributions with layer integrated gradients and attention visualizations, we observed that tokens with more direct, verifiable meaning and numeric figures

were often influential in Tweets predicted as “Valid,” while inflammatory or emotional phrases had higher weight in predictions of “Misleading” or “Misinformation,” indicating that the model focuses on largely explainable portions of the input text when making predictions.

We hope that our findings in the classification of misinformation and its associations with post engagement can contribute towards future work in detecting misleading content on social media and analyzing its relationships with political and cultural factors.

A Appendix

A.1 Full List of Accounts and Political Leanings

Left	Center	Right
1. Wonkette	1. ABC News	1. Fox News
2. Jezebel	2. BBC News	2. Tomi Lahren
3. The Young Turks	3. CBS News	3. Greg Gutfeld
4. The Root	4. PBS News	4. Dana Perino
5. The Week	5. Wall Street Journal	5. Jesse Watters
6. Vox	6. Joe Rogan*	6. Jeanine Pirro
7. HuffPost		7. Tucker Carlson
8. CNN		8. Sean Hannity
9. Ana Navarro-Cárdenas		9. Laura Ingraham
10. Jim Acosta		10. New York Post
11. Wolf Blitzer		11. The Blaze
12. Christopher Cuomo		12. The Daily Wire
13. Sanjay Gupta		13. Ben Shapiro
14. Don Lemon		14. PragerU
15. MSNBC News		15. Dinesh D'Souza
16. George Stephanopoulos		16. Steven Crowder
17. Rachel Maddow		17. Ted Cruz
18. Nicolle Wallace		18. The Epoch Times
19. Joy Reid		19. Dave Rubin
20. The New York Times		20. Breitbart News
		21. The Federalist

* See sensitivity analysis in Appendix A.6 for discussion and handling of Joe Rogan's political affiliation.

Table 7: Full List of Twitter Accounts Used to Collect Tweets

A.2 Engagement Metric Clustering by Account

Here, we present several visualizations that appear to display engagement clustering by account, influencing the hierarchical structure of our modelling approach.

In Figure 4, we can observe the logged engagement of all Tweets, grouped by Twitter account.

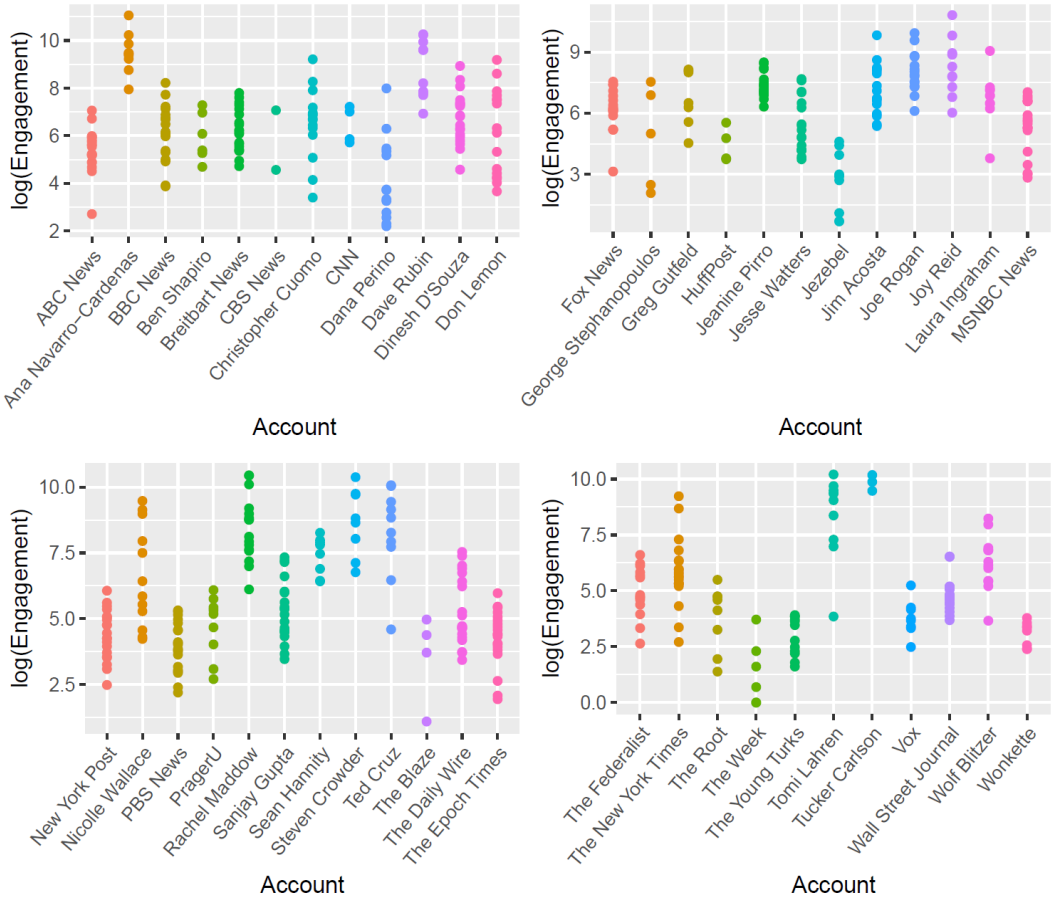


Figure 4: log(Engagement) of Tweets Grouped by Account

From the plots, it appears that engagement may be somewhat clustered by account. Accounts that have experienced high engagement such as Ana Navarro-Cárdenas, Joy Reid, and Steven Crowder seem more likely to have high engagement for many of their Tweets, while accounts that have experienced low engagement such as Jezebel, The Epoch Times, and The Young Turks seem more likely to sustain fairly low engagement. From these visualizations, we believe it is reasonable to assume possible clustering in engagement by account, motivating the use of a random intercept on account.

Next, we can observe changes in engagement across time period for each account. In Figure 5, we display a subset of accounts and their Tweet engagement in each time period. We also provide the line of best fit for each account.

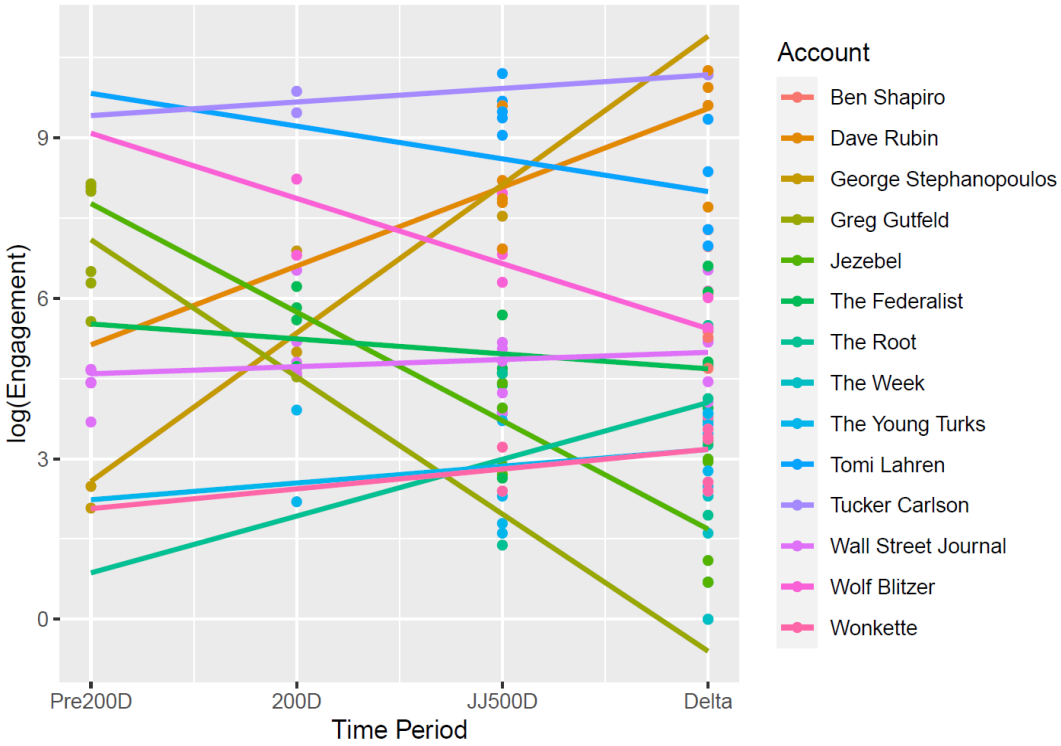


Figure 5: log(Engagement) Across Time Period by Account

The most striking characteristic of the plot above is the vast differences in best-fit lines for each account. Some accounts, like Jezebel and Greg Gutfeld, appear to display a sharp decrease in engagement as time continues, while others, like Dave Rubin and George Stephanopolous, present a sharp increase. Some accounts have fairly shallow slopes, such as Tucker Carlson and Wall Street Journal. This indicates that the association between time period and engagement may differ depending on account – because Tweets were pulled over only about 20 months, engagement is subject to various external forces that are somewhat difficult to predict (e.g. social media hiatus, sudden involvement in controversy, etc.), motivating our decision to allow slopes for time periods to vary by account in our hierarchical model.

A.3 Prior Distributions for Bayesian Negative Binomial Regression

Utilizing the *rstanarm* package in R to automatically adjust and rescale priors, the prior distributions used for the intercept and each coefficient are as follows:

$$\beta_0 \sim N(3, 2)$$

$$\beta_1 \sim N(1, 4.920)$$

$$\beta_2 \sim N(1, 4.464)$$

$$\beta_3 \sim N(1, 4.128)$$

$$\beta_4 \sim N(1, 4.073)$$

$$\beta_5 \sim N(1, 4.025)$$

$$\beta_6 \sim N(1, 5.222)$$

$$\beta_7 \sim N(1, 4.013)$$

$$\beta_8 \sim N(1, 27.662)$$

$$\beta_9 \sim N(1, 0.472)$$

$$\beta_{10} \sim N(1, 3.225)$$

$$\beta_{11} \sim N(1, 4.097)$$

$$\beta_{12} \sim N(1, 4.019)$$

$$\beta_{13} \sim N(1, 5.797)$$

$$\beta_{14} \sim N(1, 4.797)$$

A.4 Negative Binomial Regression Assumptions

A.4.1 Independence

As described in Appendix A.2, Tweets from the same account may be likely to have correlated engagement metrics – large accounts may have fairly high numbers of viewers for their Tweets, resulting in higher engagement, while small accounts may experience the opposite. To address this, we have incorporated random intercepts for accounts and random slopes for time period based on account. Beyond this independence violation, we do not anticipate any further correlation between engagement metrics of different Tweets; we believe independence is otherwise satisfied.

A.4.2 Distribution of the Response

Our response variable is a single metric of engagement, calculated by summing Likes, Retweets, Replies, and Quotes for each Tweet. This value will be a non-negative integer quantity for all Tweets. The negative binomial distribution also has an extra parameter to account for overdispersion. Because our response is a count variable that is somewhat overdispersed, we believe a negative binomial distribution is a reasonable assumption.

A.4.3 Residuals

In Figure 6, we observe several residual tests for the negative binomial model. The left plot displays observed versus expected quantiles, while the right plot gives rank-transformed model predictions versus model residuals.

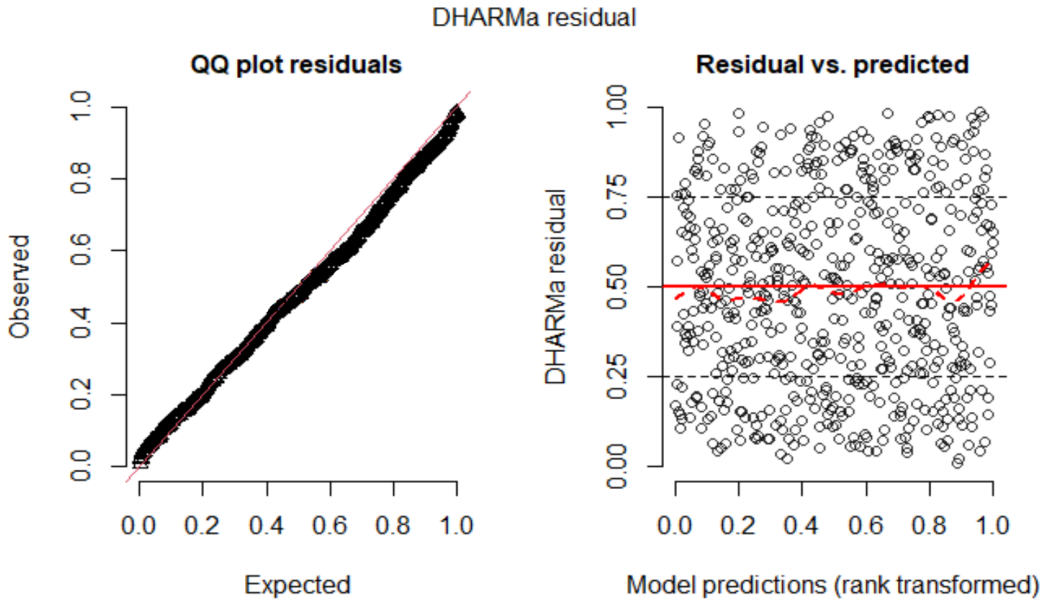


Figure 6: Residual Analysis of Negative Binomial Model

We also give the average residuals for categorical predictors in Table 8.

Category	Average Residual
Time = Pre200D	0.001
Time = 200D	0.031
Time = JJ500D	0.046
Time = Delta	0.033
Misinfo-or-Misleading = Yes	0.011
Misinfo-or-Misleading = No	0.046
Deleted = Yes	0.041
Deleted = No	0.020
Uncertain = Yes	0.080
Uncertain = No	0.022
Retweet-or-Quote = Yes	0.007
Retweet-or-Quote = No	0.053
Political Lean = Right	0.031
Political Lean = Left	0.048
Political Lean = Center	-0.003

Table 8: Average Residuals for Categorical Variables

The QQ plot of residuals and observed vs. expected quantiles do not show any large deviations, and the plot of residuals vs. rank-transformed predicted values displays no concerning irregularities. We also performed a test of correct distribution (KS test), as well as tests for outliers and dispersion – none were statistically significant. Finally, all average residuals for categorical variables are close to 0, indicating no issues with model fit.

A.4.4 Predictors versus Response

In Figure 7, we display our three numeric predictors versus engagement.

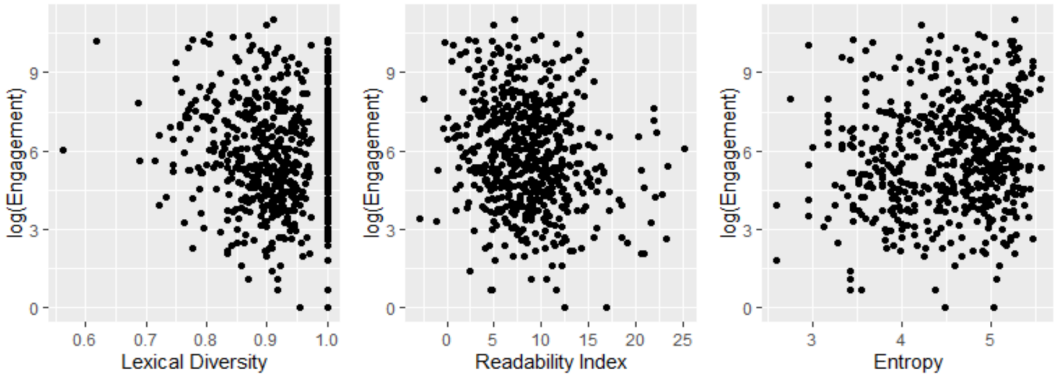


Figure 7: Numeric Predictors vs. Engagement

From the plots, we do not see any irregular or nonlinear patterns, indicating that the assumption of linearity between log(Engagement) and numeric predictors is satisfied.

A.4.5 Leverage, Cook's Distance, Standardized Residuals

In Figure 8, we display leverage, Cook's distance, and standardized residuals. A leverage threshold of 0.047 is used based on the formula $\frac{2\times(p+1)}{n}$, where p is the number of estimated coefficients and n is the number of observations.

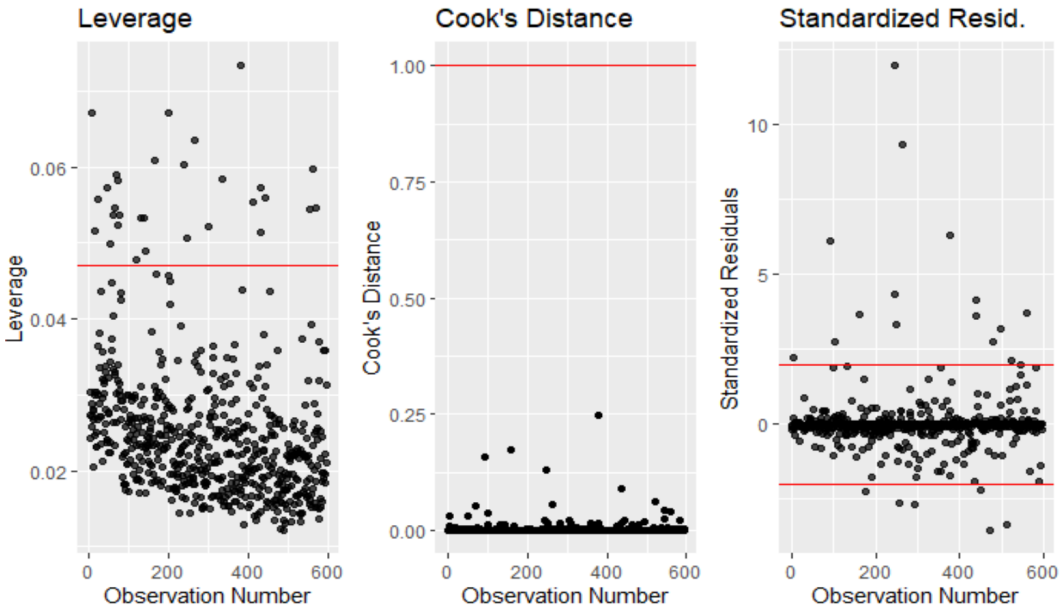


Figure 8: Leverage, Cook's Distance, and Standardized Residuals

There are 30 observations with high leverage, 21 observations with high standardized residuals, and no observations with high Cook's distance. Because these observations comprise a relatively small portion of our dataset and no irregularities were observed in other residual plots, we do not believe that any influential points greatly altered the model fit.

A.4.6 Generalized Variance Inflation Factors

We display GVIF in Table 9, utilizing $\text{GVIF}^{\frac{1}{2\times df}}$ to make the quantities comparable between covariates with multiple estimated coefficients and covariates with only a single estimated coefficient.

Covariate	$\text{GVIF}^{\frac{1}{2\times df}}$
Time Period	1.053
Misinfo-or-Misleading	3.216
Deleted	1.050
Uncertain	1.010
Retweet-or-Quote	1.079
Lexical Diversity	2.953
Readability Index (ARI)	1.148
Entropy	2.361
Political Lean	1.653
Misinfo-or-Misleading/Political Lean	1.816

Table 9: Generalized Variance Inflation Factors

Because all variance inflation factors are well below 10, we do not anticipate any problems with multicollinearity.

A.5 Model Diagnostics

We examine posterior predictive distributions, traceplots, and ACF plots from our Bayesian negative binomial regression model below.

A.5.1 Posterior Predictive Distributions

We first examine posterior predictive distributions of engagement in Figure 9. The bottom plot gives the same distribution as the top plot but on a more restricted domain.

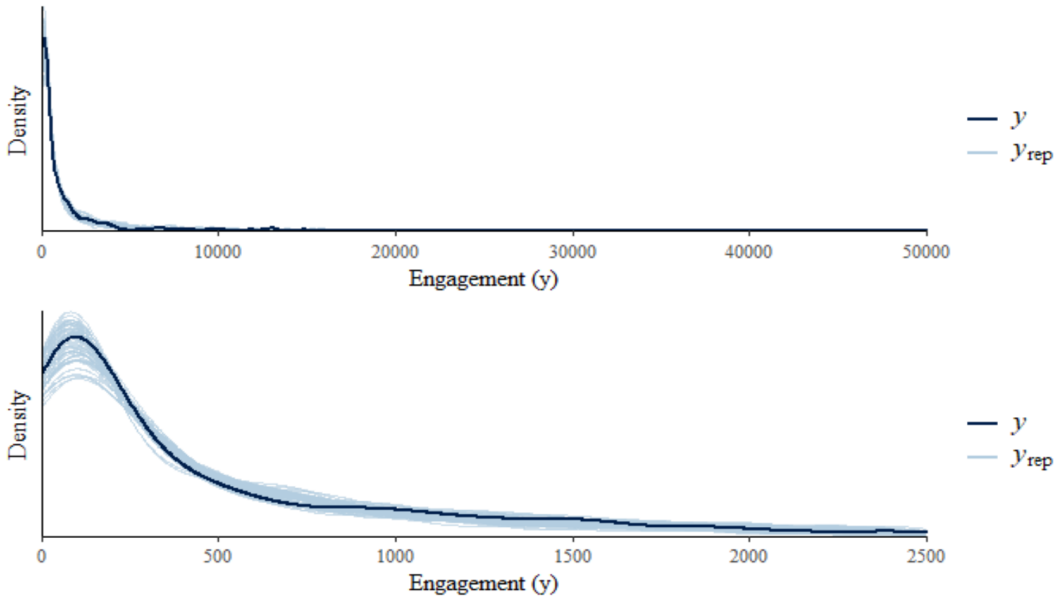


Figure 9: Posterior Predictive Distributions

From the plots, it appears that the posterior predictive distributions of the response correspond well to the true distribution, indicating no issues with model fit.

A.5.2 TRACEPLOTS

Figure 10 displays traceplots for all model coefficients.

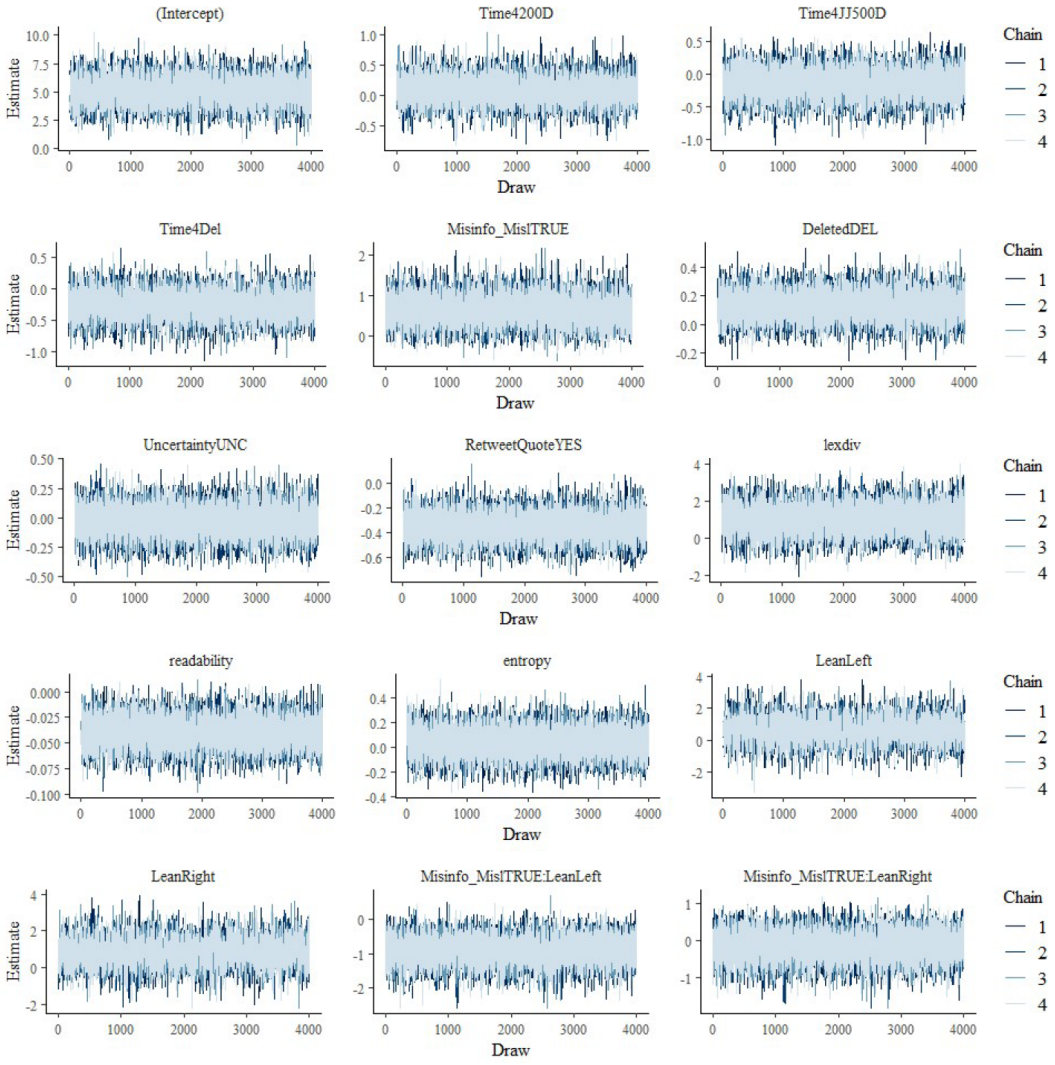
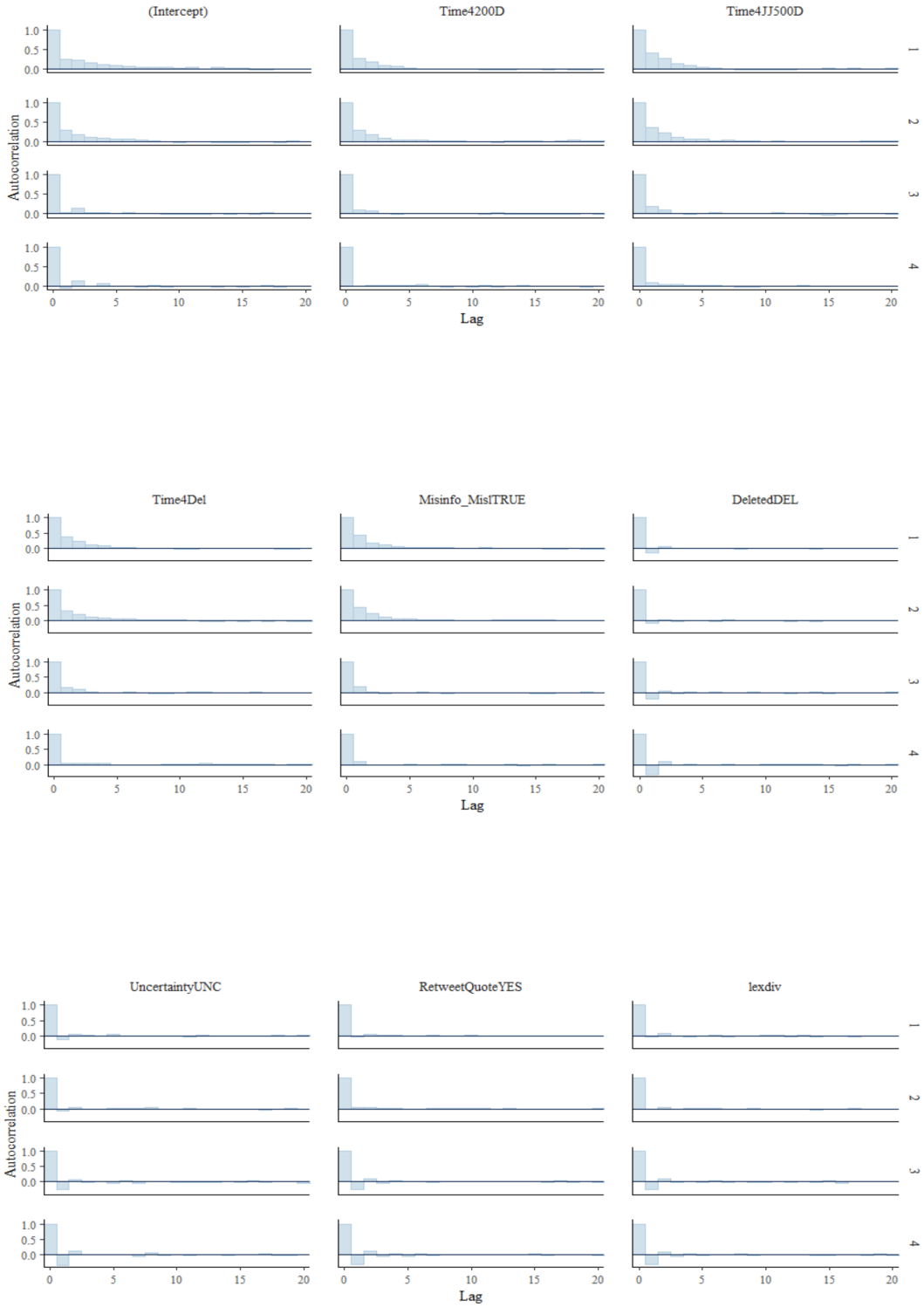


Figure 10: Traceplots of Coefficients

From these plots, it appears that the sampler explored a relatively wide range of values for each coefficient without displaying any irregularities or getting stuck in a certain area of the parameter space, indicating sufficient mixing.

A.5.3 ACF PLOTS

We display ACF plots for all model coefficients in Figure 11 below.



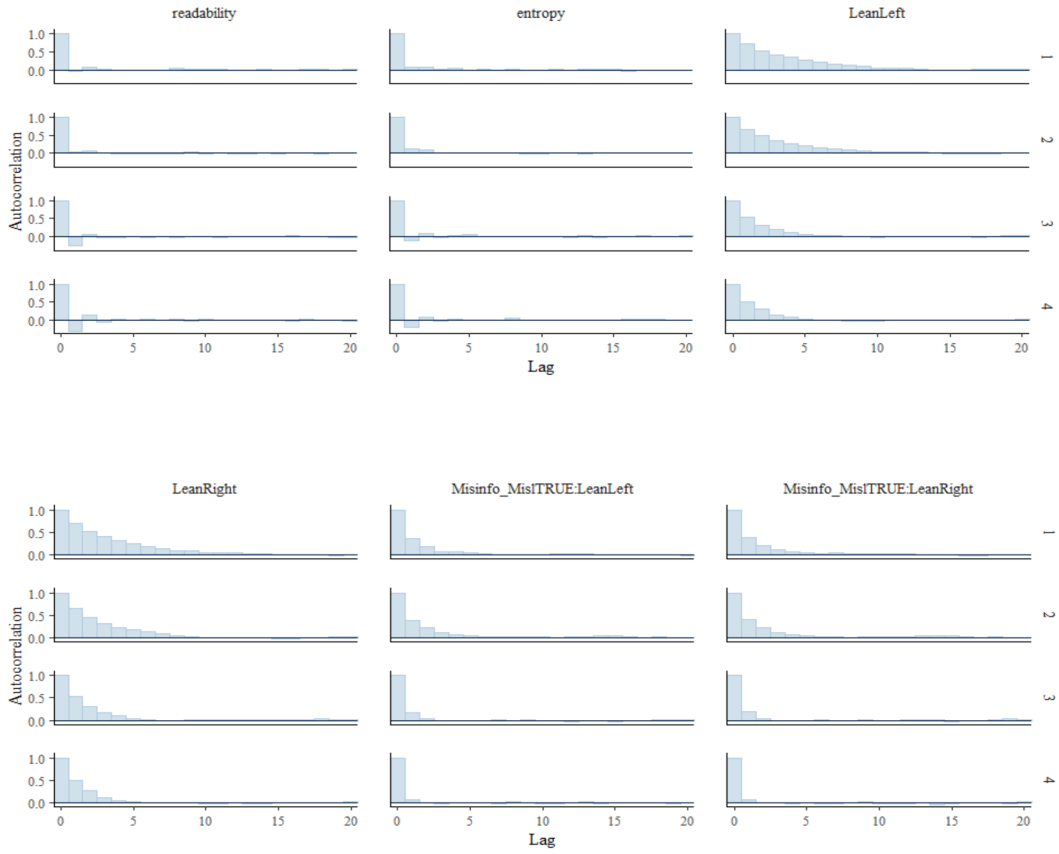


Figure 11: ACF Plots for Coefficients

From the ACF plots, it appears that there is almost zero autocorrelation after 4-5 draws for most covariates. The coefficients for left and right political leans display some autocorrelation past this point, but because the value is still fairly small, we are not concerned about autocorrelation in model draws.

A.6 Joe Rogan Sensitivity Analysis

In this section, we present the results of a sensitivity analysis which alters the political classification of Joe Rogan to analyze whether this heavily influences the model.

Joe Rogan is a media commentator who frequently shares his opinions on politics, philosophy, science, and comedy through his podcast and various social media accounts. He is known for holding political beliefs that range from far-right to far-left, depending on the specific issue, and he frequently has guests on his podcasts from all sides of the political spectrum.

In analyzing Rogan's Tweets, we found that many were politically polarizing in some form, either leaning substantially right or left. This posed a problem for the classification of his political affiliation – he has claimed to be an independent (Cleary), indicating that perhaps the Center category in our schema may be the best fit, although almost none of his individual Tweets are likely to be considered centrist. Because he holds views that would cause some to label him right-wing and others to label him left-wing, our main model attributes the Center label to him, and we experiment with Left and Right labels in Tables 10 and 11 to observe how this may change model results. While the political affiliation of one account may seem unlikely to alter model estimates, we noted that 13 out of 15 Tweets contributed by Joe Rogan were either misinformation or misleading. We believe this distribution makes it possible for his listed political leaning to influence at least the interaction between misleading/misinformative Tweets and political lean, motivating this analysis to investigate.

Covariate	Estimate	95% Credible Interval
Intercept	4.650	(2.310, 7.036)
Time = 200D	0.081	(-0.321, 0.486)
Time = JJ500D	-0.155	(-0.575, 0.253)
Time = Delta	-0.262	(-0.687, 0.138)
Misinfo-or-Misleading = Yes	0.389	(-0.254, 1.101)
Deleted = Yes	0.132	(-0.063, 0.324)
Uncertain = Yes	-0.023	(-0.269, 0.234)
Retweet-or-Quote = Yes	-0.346	(-0.561, -0.129)
Lexical Diversity	1.083	(-0.396, 2.541)
Readability Index (ARI)	-0.039	(-0.067, -0.011)
Entropy	0.050	(-0.171, 0.281)
Political Lean = Left	1.108	(-0.518, 2.732)
Political Lean = Right	1.163	(-0.386, 2.717)
Political Lean = Left and Misinfo-or-Misleading = Yes	-0.575	(-1.357, 0.162)
Political Lean = Right and Misinfo-or-Misleading = Yes	0.089	(-0.676, 0.805)

Table 10: Negative Binomial Regression Model: Joe Rogan Labeled Left

It appears that model results are largely equivalent regardless of the political leaning for Joe Rogan. When he is listed as either Left or Right, the 95% credible interval for the interaction

between left political lean and misinformative/misleading status includes 0, while it does not include 0 when he is labeled Center.

Covariate	Estimate	95% Credible Interval
Intercept	4.642	(2.307, 7.059)
Time = 200D	0.083	(-0.322, 0.502)
Time = JJ500D	-0.154	(-0.579, 0.265)
Time = Delta	-0.260	(-0.678, 0.148)
Misinfo-or-Misleading = Yes	0.378	(-0.277, 1.090)
Deleted = Yes	0.140	(-0.060, 0.337)
Uncertain = Yes	-0.032	(-0.283, 0.222)
Retweet-or-Quote = Yes	-0.348	(-0.563, -0.136)
Lexical Diversity	1.092	(-0.403, 2.533)
Readability Index (ARI)	-0.040	(-0.067, -0.012)
Entropy	0.054	(-0.174, 0.275)
Political Lean = Left	0.982	(-0.606, 2.556)
Political Lean = Right	1.208	(-0.273, 2.760)
Political Lean = Left and Misinfo-or-Misleading = Yes	-0.641	(-1.428, 0.100)
Political Lean = Right and Misinfo-or-Misleading = Yes	-0.142	(-0.621, 0.862)

Table 11: Negative Binomial Regression Model: Joe Rogan Labeled Right

The estimates for left and right political leaning have higher positive magnitude when Joe Rogan is labeled Left or Right as compared to when he is labeled Center, although the 95% credible intervals still include 0. Despite the large percentage of misleading or misinformative content in Joe Rogan's Tweets, it seems that his political affiliation alone does not drastically alter model results.

To gain a quantitative estimate of Joe Rogan's political affiliation, we utilized the R package *tweetscores*, which estimates the political ideology of Twitter accounts using a Bayesian spatial model with latent variables to represent links between average users and a set of political "elites" in the social network. Based upon the political elites that a user follows, the package implements two main methods, Metropolis sampling and maximum likelihood estimation, to obtain a numeric estimate for a user's political ideology from the joint posterior distribution of the political leanings of accounts they follow, their level of political interest, and the popularity of accounts they follow (Barberá, 2017). Estimates range from around -2 to +2, where Barack Obama and Fox News score around -1.4 and +1.4, respectively. Joe Rogan follows 37 political elites, and he scores around -0.87 using either of the methods described above, indicating a left lean. However, the package implemented a third procedure which utilizes correspondence analysis and expands the definition of political elites to include accounts that are not political themselves but have highly partisan follower bases. With this method, Joe Rogan follows 120 elites, and he scored around +0.87, indicating a right lean. Although the *tweetscores* package does not necessarily help us classify Joe Rogan's political affiliation, it supports our observation that his politics are difficult to pinpoint and fluctuate heavily between parties. See the Discussion for further comments.

A.7 DeBERTa Hyperparameters

We fine-tuned several hyperparameters of the DeBERTa model: we experimented with learning rates in [0.1, 0.01, 0.001, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}, 1×10^{-7}] (in addition to a default value of 2×10^{-5}), dropout rates in [0.0, 0.1, 0.15, 0.2, 0.25, 0.3, 0.4, 0.5], and weight decay rates in [0.1, 0.01, 0.001, 1×10^{-4}, 1×10^{-5}, 1×10^{-6}, 1×10^{-7}].

The selected parameters were a learning rate of 2×10^{-5}, a dropout rate of 0.0, and weight decay of 0.01. Additionally, in the AdamW optimizer, we utilized $\epsilon = 1 \times 10^{-8}$ to improve numerical stability.

We also utilized inverse frequency of each label to implement weighted loss. We assigned label x with a weight equal to $1 - \frac{\text{number of training Tweets with label } x}{\text{total number of training Tweets}}$. Weights were [0.91202, 0.68777, 0.40021] for Misinformation, Misleading, and Valid labels, respectively.

A.8 Full Negative Binomial Regression Model Output

The full output of the Bayesian hierarchical negative binomial regression model, including all estimates for random intercepts and random slopes, is provided in Tables 12 and 13.

A.8.1 Fixed Effects

Covariate	Estimate	95% Credible Interval
Intercept	5.068	(2.755, 7.400)
Time = 200D	0.094	(-0.316, 0.509)
Time = JJ500D	-0.143	(-0.567, 0.272)
Time = Delta	-0.253	(-0.683, 0.167)
Misinfo-or-Misleading = Yes	0.636	(-0.020, 1.351)
Deleted = Yes	0.132	(-0.062, 0.327)
Uncertain = Yes	-0.035	(-0.280, 0.220)
Retweet-or-Quote = Yes	-0.345	(-0.563, -0.129)
Lexical Diversity	1.065	(-0.429, 2.529)
Readability Index (ARI)	-0.040	(-0.067, -0.012)
Entropy	0.038	(-0.187, 0.257)
Political Lean = Left	0.644	(-0.899, 2.231)
Political Lean = Right	0.854	(-0.621, 2.338)
Political Lean = Left and Misinfo-or-Misleading = Yes	-0.894	(-1.695, -0.151)
Political Lean = Right and Misinfo-or-Misleading = Yes	-0.157	(-0.917, 0.571)

Table 12: Negative Binomial Regression Model: Fixed Effects (Unexponentiated)

A.8.2 RANDOM EFFECTS

Account	Intercept	Time = 200D	Time = JJ500D	Time = Delta
ABC News	0.121	-0.399	-0.616	-0.062
Ana Navarro-Cárdenas	3.010	0.266	0.342	0.642
BBC News	0.614	-0.317	0.102	0.607
Ben Shapiro	0.030	0.001	0.008	0.020
Breitbart News	-0.130	0.023	0.189	-0.174
CBS News	0.582	-0.018	-0.007	0.186
Christopher Cuomo	1.010	0.476	-0.272	-0.577
CNN	0.388	0.001	-0.020	0.102
Dana Perino	-1.022	0.260	-0.123	-0.651
Dave Rubin	1.885	-0.302	-0.322	0.805
Dinish D'Souza	0.353	-0.340	-0.073	0.505
Don Lemon	1.022	0.264	-0.414	-0.333
Fox News	-0.418	-0.279	0.412	0.413
George Stephanopoulos	-0.337	0.397	0.582	-0.028
Greg Gutfeld	0.736	-0.280	-0.188	0.243
HuffPost	-1.208	0.026	-0.001	-0.373
Jeanine Pirro	0.688	0.119	0.094	0.110
Jesse Watters	-1.241	0.723	0.616	-0.776
Jezebel	-2.509	0.269	0.265	-0.963
Jim Acosta	1.745	0.342	-0.193	0.002
Joe Rogan	2.054	-0.094	-0.079	0.665
Joy Reid	2.301	0.310	0.387	0.405
Laura Ingraham	0.368	0.307	0.376	-0.143
MSNBC News	0.094	-0.350	-1.028	-0.304
New York Post	-1.819	-0.454	0.135	0.165
Nicolle Wallace	1.149	-0.772	-0.053	1.234
PBS News	-1.224	-0.247	-0.117	-0.152
PragerU	-1.222	-0.130	-0.156	-0.237
Rachel Maddow	2.870	0.204	-0.332	0.438
Sanjay Gupta	-0.205	0.196	-0.421	-0.517
Sean Hannity	0.782	-0.223	-0.242	0.421
Steven Crowder	1.799	-0.288	-0.296	0.767
Ted Cruz	1.739	0.051	0.098	0.463
The Blaze	-1.503	0.030	0.022	-0.459
The Daily Wire	-0.853	-0.521	0.464	0.630
The Epoch Times	-1.736	-0.335	0.251	0.127
The Federalist	-1.176	-0.291	-0.381	-0.104
The New York Times	0.136	0.918	0.922	-0.562
The Root	-1.252	-0.149	-0.204	-0.240
The Week	-2.533	0.046	0.027	-0.789
The Young Turks	-2.464	-0.415	-0.457	-0.349
Tomi Lahren	1.598	0.431	0.535	0.107
Tucker Carlson	2.035	0.096	-0.007	0.451
Vox	-1.929	0.084	0.055	-0.641
Wall Street Journal	-0.461	0.048	0.051	0.294
Wolf Blitzer	0.346	0.660	0.294	-0.582
Wonkette	-2.246	-0.201	-0.289	-0.478

Table 13: Negative Binomial Regression Model: Random Effects (Unexponentiated)

BIBLIOGRAPHY

Media Bias Chart. *Ad Fontes Media*. URL https://adfontesmedia.com/interactive-media-bias-chart/. 5 Nov 2021.

Differences in how Democrats and Republicans behave on Twitter. *Pew Research Center*. URL https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/. 15 Oct 2020.

Maria Abascal, Tiffany J. Huang, and Van C. Tran. Intervening in anti-immigrant sentiments: The causal effects of factual information on attitudes toward immigration. *The ANNALS of the American Academy of Political and Social Science*, 697(1):174–191, 2021. URL https://doi.org/10.1177/00027162211053987.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 9525–9536, 2018. URL https://doi.org/10.48550/arxiv.1810.03292.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2), 2019. URL https://doi.org/10.1177/2053168019848554.

Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review*, 2020. URL http://dx.doi.org/10.37016/mr-2020-033.

Justin Baragona. Tucker Carlson bizarrely claims getting COVID 'does feminize people'. *MSN*. URL https://www.msn.com/en-us/news/world/tucker-carlson-bizarrely-claims-getting-covid-does-feminize-people/ar-AARCYHS. 08 Dec 2021.

Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. *Political Analysis*, 23(1):76–91, 2017. URL https://doi.org/10.1093/pan/mpu011.

Nicolas Berlinski, Margaret Doyle, Andrew M. Guess, Gabrielle Levy, Benjamin Lyons, Jacob M. Montgomery, Brendan Nyhan, and Jason Reifler. The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science*, pp. 1–16, 2021. URL https://doi.org/10.1017/XPS.2021.18.

Leticia Bode and Emily K. Vraga. In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, 65(4): 619–638, 2015. URL https://doi.org/10.1111/jcom.12166.

Leticia Bode and Emily K. Vraga. See something, say something: Correction of global health misinformation on social media. *Health Communication*, 33(9):1131–1140, 2017. URL https://doi.org/10.1080/10410236.2017.1331312.

Shannon Bond. What the Joe Rogan podcast controversy says about the online misinformation ecosystem. *NPR*. URL https://www.npr.org/2022/01/21/1074442185/joe-rogan-doctor-covid-podcast-spotify-misinformation. 21 Jan 2022.

Kaiping Chen, Yepeng Jin, and Anqi Shao. Science factionalism: How group identity language affects public engagement with misinformation and debunking narratives on a popular Q&A platform in China. *arXiv e-prints*, 2021. URL https://doi.org/10.48550/arXiv.2112.07968.

Wen-Ying Sylvia Chou, April Oh, and William M. P. Klein. Addressing health-related misinformation on social media. *Journal of the American Medical Association*, 320(23):2417–2418, 2018. URL https://doi.org/10.1001/jama.2018.16865.

Matteo Cinelli, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo, and Antonio Scala. The COVID-19 social media infodemic. *Scientific Reports*, 10, 2020. URL https://doi.org/10.1038/s41598-020-73510-5.

John Cleary. Is Joe Rogan a Republican? What are his politics? *Heavy*. URL https://heavy.com/news/joe-rogan/is-joe-rogan-a-republican/. 16 Mar 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019. URL https://www.semanticscholar.org/paper/BERT%3A-Pre-training-of-Deep-Bidirectional-for-Devlin-Chang/df2b0e26d0599ce3e70df8a9da02e51594e0e992.

Anne D'Innocenzio. Joe Rogan apologizes for racial slur after video surfaces. *ABC News*. URL https://abcnews.go.com/Entertainment/wireStory/joe-rogan-apologizes-racial-slurs-video-surfaces-82695536. 6 Feb 2022.

Ashley Fantz. Ebola facility in Liberia attacked; patients flee. *CNN*. URL https://www.cnn.com/2014/08/17/world/africa/ebola-liberia-attack/index.html. 18 Aug 2014.

Ashish Goel and Latika Gupta. Social media in the times of COVID-19. *Journal of Clinical Rheumatology: Practical Reports on Rheumatic Musculoskeletal Diseases*, 26(6):220–223, 2020. URL https://doi.org/10.1097/RHU.0000000000001508.

Eric Hananoki. A comprehensive guide to Alex Jones: Conspiracy theorist and Trump "valuable asset". *Media Matters*. URL https://www.mediamatters.org/donald-trump/comprehensive-guide-alex-jones-conspiracy-theorist-and-trump-valuable-asset. 01 Dec 2016.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, 2020. URL https://arxiv.org/pdf/2005.01831.pdf.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Wei Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. *2021 International Conference on Learning Representations*, 2021. URL https://www.microsoft.com/en-us/research/publication/deberta-decoding-enhanced-bert-with-disentangled-attention-2/.

Yiqing Hua. Understanding BERT performance in propaganda analysis. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2019. URL https://doi.org/10.18653/v1/D19-5019.

A. K. M. Najmul Islam, Samuli Laato, Shamim Talukder, and Erkki Sutinen. Misinformation sharing and social media fatigue during COVID-19: An affordance and cognitive load perspective. *Technological Forecasting and Social Change*, 159, 2020. URL https://doi.org/10.1016/j.techfore.2020.120201.

Alicia A. Johnson, Miles Q. Ott, and Mine Dogucu. *Bayes Rules!: An Introduction to Applied Bayesian Modeling*. Chapman and Hall/CRC, 2021. ISBN 1032191597.

Nattiya Kanhabua and Wolfgang Nejdl. Understanding the diversity of Tweets in the time of outbreaks. *Proceedings of the 22nd International Conference on World Wide Web*, pp. 1335–1342, 2013. URL https://doi.org/10.1145/2487788.2488172.

Ansgar Kellner, Lisa Rangosch, Christian Wressnegger, and Konrad Rieck. Political elections under (social) fire? Analysis and detection of propaganda on Twitter. *Institute of System Security: Computer Science Report*, 2019. URL https://arxiv.org/pdf/1912.04143.pdf.

Chandra Mouli Madhav Kotteti, Xishuang Dong, and Lijun Qian. Rumor detection on time-series of Tweets via deep learning. *IEEE Military Communications Conference (MILCOM)*, pp. 1–7, 2019. URL https://doi.org/10.1109/MILCOM47813.2019.9020895.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of propaganda techniques in news articles. *Proceedings of the 14th International Workshop on Semantic Evaluation*, pp. 1377–1414, 2020. URL http://dx.doi.org/10.18653/v1/2020.semeval-1.186.

Shahan Ali Memon and Kathleen M. Carley. CMU-MisCov19: A novel Twitter dataset for characterizing COVID-19 misinformation. *Carnegie Mellon University*. URL https://zenodo.org/record/4024154#.Yiwn2HrMLD6. 20 Jan 2022.

Liz Neporent. Nigerian Ebola hoax results in two deaths. *ABC News*. URL https://abcnews.go.com/Health/nigerian-ebola-hoax-results-deaths/story?id=25842191. 30 Sep 2014.

Dimitar Nikolov, Alessandro Flammini, and Filippo Menczer. Right and left, partisanship predicts (asymmetric) vulnerability to misinformation. *Harvard Kennedy School (HKS) Misinformation Review*, 2021. URL https://doi.org/10.37016/mr-2020-55.

Jim Norman. Americans' confidence in institutions stays low. *Gallup*. URL https://news.gallup.com/poll/192581/americans-confidence-institutions-stays-low.aspx. 13 Jun 2016.

Michelle Odlum and Sunmoo Yoon. What can we learn about the Ebola outbreak from Tweets? *American Journal of Infection Control*, 43(6):563–571, 2015. URL https://doi.org/10.1016/j.ajic.2015.02.023.

Michael Orlov and Marina Litvak. Using behavior and text analysis to detect propagandists and misinformers on Twitter. *Communcations in Computer and Information Science: Information Management and Big Data*, 898, 2019. URL https://doi.org/10.1007/978-3-030-11680-4_8.

R. J. Senter and E. A. Smith. Automated Readability Index. *Cincinnati Univ OH*, 1967. URL https://apps.dtic.mil/sti/citations/AD0667273.

Mirela Silva, Fabrício Ceschin, Prakash Shrestha, Christopher Brant, Juliana Fernandes, Catia S. Silva, André Grégio, Daniela Oliveira, and Luiz Giovanini. Predicting misinformation and engagement in COVID-19 Twitter discourse in the first months of the outbreak. *arXiv e-prints*, 2020. URL https://www.researchgate.net/publication/346614769_Predicting_Misinformation_and_Engagement_in_COVID-19_Twitter_Discourse_in_the_First_Months_of_the_Outbreak.

Qi Su, Mingyu Wan, Xiaoqian Liu, and Chu-Ren Huang. Motivations, methods and metrics of misinformation detection: An NLP perspective. *Natural Language Processing Research*, 1(1-2): 1–13, 2020. URL https://doi.org/10.2991/nlpr.d.200522.001.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *Proceedings of International Conference on Machine Learning*, pp. 3319–3328, 2017. URL https://doi.org/10.48550/arXiv.1703.01365.

Terry Traylor, Jeremy Straub, Gurmeet, and Nicholas Snell. Classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator. *2019 IEEE 13th International Conference on Semantic Computing*, pp. 445–449, 2019. URL https://doi.org/10.1109/ICOSC.2019.8665593.

Kathie M. d'I. Treen, Hywel T. P. Williams, and Saffron J. O'Neill. Online misinformation about climate change. *WIREs Climate Change*, 11(5), 2020. URL https://doi.org/10.1002/wcc.665.

Sebastián Valenzuela, Daniel Halpern, James E. Katz, and Juan Pablo Miranda. The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation. *Digital Journalism*, 7(6):802–823, 2019. URL https://doi.org/10.1080/21670811.2019.1623701.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *31st Conference on Neural Information Processing Systems*, pp. 6000–6010, 2019. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Emily K. Vraga and Leticia Bode. Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication*, 37(1): 136–144, 2020. URL https://doi.org/10.1080/10584609.2020.1716500.

Yuxi Wang, Martin McKee, Aleksandra Torbica, and David Stuckler. Systematic literature review on the spread of health-related misinformation on social media. *Social Science and Medicine*, 240, 2019. URL https://doi.org/10.1016/j.socscimed.2019.112552.

WHO. Novel Coronavirus (2019-nCoV) Situation Report - 13. *World Health Organization*. URL https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200202-sitrep-13-ncov-v3.pdf?sfvrsn=195f4010_6. 2 Feb 2020.

Thomas Wood. Transformer neural networks. *DeepAI*. URL https://deepai.org/machine-learning-glossary-and-terms/transformer-neural-network. 11 Mar 2022.

Shuo Yang, Kai Shu, Suhang Wang, Renjie Gu, Fan Wu, and Huan Liu. Unsupervised fake news detection on social media: A generative approach. *Proceedings of 33rd AAAI Conference on Artificial Intelligence*, 2019. URL http://dx.doi.org/10.1609/aaai.v33i01.33015644.