(Mis)leading the COVID-19 Vaccination Discourse on Twitter: An Exploratory Study of Infodemic Around the Pandemic

Shakshi Sharma®, Rajesh Sharma, and Anwitaman Datta®

Abstract—In this work, we collect a moderate-sized representative corpus of tweets (over 200000) pertaining to COVID-19 vaccination spanning for a period of seven months (September 2020–March 2021). Following a transfer learning approach, we utilize a pretrained transformer-based XLNet model to classify tweets as misleading or nonmisleading and manually validate the results with random subsets of samples. We leverage this to study and contrast the characteristics of tweets in the corpus that are misleading in nature against non-misleading ones. This exploratory analysis enables us to design features such as sentiments, hashtags, nouns, and pronouns which can, in turn, be exploited for classifying tweets as (non-)misleading using various machine learning (ML) models in an explainable manner. Specifically, several ML models are employed for prediction, with up to 90% accuracy, with the importance of each feature is explained using SHAP Explainable AI (XAI) tool. While the thrust of this work is principally exploratory in nature to obtain insight on the online discourse on COVID-19 vaccination, we conclude the article by outlining how these insights provide the foundations for a more actionable approach to mitigate misinformation. We have made the curated data as well as the accompanying code available so that the research community at large can reproduce, compare against, or build upon this work.

Index Terms—COVID-19, explainable AI (XAI), misinformation, social media.

“We live in an era of unprecedented scientific breakthroughs and expertise. But we’re also stymied by the forces of misinformation that undermine the true knowledge that is out there.” — Dr. Laolu Fayanju [1]

I. INTRODUCTION

The fears, uncertainties, and doubts (FUDs) around COVID-19 and the available vaccines have been amplified by the ongoing discourse on social media leading to an infodemic (a portmanteau of information and epidemic, referring to the spread of possibly accurate and inaccurate information about a disease, itself spreading like an epidemic). The infodemic has been fueled by prolific misinformation spreaders, e.g., the “disinformation dozen” [2], who may have special interests and agendas in doing so. Machingaidze and Wiysonge [3] studied vaccine hesitancy in several low- and middle-income countries (LMICs) and compared that with the USA and Russia, which were some of the countries at the forefront of COVID-19 vaccine research. The average vaccine acceptance rate in LMICs was reported to be 80.3%, compared to 64.6% in the USA and 30.4% in Russia. Such vaccination hesitancy has made the impact of a harsh pandemic much worse than necessary, e.g., a study estimated a factor of up to eight more fatalities [4] attributed to this.

Given the pivotal role of social media both as a medium of propagation of misleading information, as well as an easy source to probe the prevailing sentiments at scale, we chose to use openly accessible COVID-19 vaccine-related tweets to drive our research. Unlike traditional surveys, for example, [1] which are able to collect and analyze detailed demographic breakdowns such as age, race, gender, education, income, geography, etc., our approach is not granular. However, the advantages of sensing information from open social media are its sheer scale and the possibility to carry out continuous and real-time monitoring once the data pipeline is set up.1

First, we wanted to establish an understanding of what are the main topics of discussion and contentions around COVID-19 vaccination to identify the drivers of vaccine hesitancy. This exploratory analysis can provide actionable insights to policy-makers on how to approach and address vaccine hesitancy.

Second, we wanted to pinpoint misleading tweets promptly and in an automated yet explainable manner. Being able to recognize misleading tweets at scale is useful to identify when and which specific misinformation is gaining traction. This could be useful in many ways. Public health policy-makers can use the information to prioritize their combat against a particular sub-genre of misinformation in a timely fashion. It can aid the governance of the social media platform, e.g., by censoring or tagging misleading tweets, as might be applicable as per terms of usage and regulatory requirements. Outside the scope of this work, we could also help determine suitable counter-points, and timely automated rebuttals of misleading information (this is a follow-up study we are currently carrying out). The explainability aspect of the classification process proximately helps in validating the quality of the classification. In the longer run (outside the scope of

1This work presents only a retrospective analysis.
the current article, it could help decipher temporal shifts in the behavior of misinformation propagators (which could help update and keep the ML models efficacious) and also generalize the approach beyond the current topic.

Accordingly, the key highlights and contributions of the current work are as follows.

1) **Data** (following FAIR principle, i.e., findable, accessible, interoperable, and reusable): From an initial set of over 200,000 tweets which we originally collected, we curated a representative denoised collection of 114,635 tweets related to COVID-19 vaccine, along with two mutually exclusive randomly sampled subsets of 1500 and 1000 tweets manually labeled in terms of whether they are misleading or nonmisleading. We have provided the dataset2 comprising separately the manually labeled data used for training and testing, as well as the results of algorithmic prediction over a larger corpus of data, while adhering to Twitter’s content redistribution policy.3 The accompanying source code has also been open-sourced.4

2) **Exploratory Analysis:** We analyze the dataset across three dimensions: 1) *Language exploration* utilizing syntactic structure and the principal themes involved across misleading and nonmisleading tweets; 2) *Opinion study* leveraging the sentiments and emotions of both types of tweets; and 3) *Effect of visibility* analyzing meta-data of the tweets. These provide us insights to distinguish nonmisleading from misleading tweets.

3) **Classification and Explainability:** Aforementioned analysis aided the identification of potential features which could be explicitly leveraged to classify tweets to determine whether they are misleading or not. This explicit approach was compared against a black-box model (XLNet). We also looked into the mutual consistency of these approaches. The explainability aspect of our approach reinforces the credibility of the classifiers. The efficacy as well as marginal contributions of a subset of features were explored empirically for further validation.

II. RELATED WORK

In this work, we broadly discuss the literature from two perspectives, namely, online discourse on COVID-19 vaccinations and misinformation.

A. **Online Discourse on COVID-19 Vaccinations**

The COVID-19 vaccination debate [5] has been carried out on social media sites such as Reddit. The vast majority of Reddit thread comments collected are about conspiracy theories related to COVID-19 vaccines. On Twitter, utilizing temporal dimension [6], 12 million tweets in two months were released. Moreover, primary themes and user-level engagements around the pandemic were discussed. After the initiation of the Italian vaccination campaign, researchers [7] have been monitoring the online conversations of Italian Twitter users. They have discovered that there is a larger occurrence of high-credibility information despite the sharing of low-credibility information. The research [8] included a dashboard with statistics on the COVID-19 vaccination by collecting and analyzing English tweets. Another research [9] gathered almost 123 million multilingual tweets. Malagoli et al. [6] collected the COVID-19 vaccinations’ tweets for two months. Analyzed the tweets from two perspectives: 1) user engagement and 2) content properties in a temporal manner. Besides, COVID-19 fact-check dataset has also been collected and analyzed [10].

B. **Misinformation**

Misinformation is an umbrella term used for fake news, rumor, misleading, etc. The research has been performed on rumors [11], [12], fake news [13], [14], even on misinformation [15]. However, much less focus is on the misleading misinformation. For instance, Ferrara [16] focus on the influence of bots in disseminating COVID-19 conspiracy claims. The degree of misinformation surrounding the COVID-19 epidemic has also been explored [17]. Besides text, Ma and Stahl [18] used a multimodal discourse analysis technique to evaluate textual and visual information inside a public antivaccine Facebook page.

The role of Twitter in misinformation during COVID-19 [19] is a concern. Specifically, [6] is one of the closest to our study, where COVID-19 vaccination tweets are studied. Previous studies lacked an in-depth analysis of the COVID-19 vaccination data from a variety of perspectives. In the COVID-19 vaccination, for example, there are no significant themes presented in relation to various sentiments and emotions. Moreover, we bring the explainability element to our COVID-19 classification model, which, to our knowledge, has not been investigated on such a dataset. To summarize, we employed the transfer learning technique after collecting the tweets, which were then categorized as misleading or nonmisleading. Next, we used natural language processing (NLP)-based approaches to analyze, for instance, emotions with respect to misleading and nonmisleading tweets. In addition, we use machine learning (ML) techniques to determine if these NLP-based extracted features are adequate to discriminate between misleading and nonmisleading tweets in an explainable manner.

III. DATASET

The data collection and annotation process are covered in this section. The curated dataset (link provided in Section I) comprises two separate files, one containing manual labels used for training and testing, and another larger set with algorithmically predicted labels.

A. **Data Collection**

The first news of the world’s COVID-19 vaccine registration5 in August 2020, as well as Trump’s order to carry out vaccination even before it had been thoroughly tested and approved,6 signaled a vaccine rush. Naturally, this amplified

---

2https://researchdata.ntu.edu.sg/dataset.xhtml?persistentId=doi:10.21979/N9/QMLJQ
3https://developer.twitter.com/en/developer-terms/agreement-and-policy
4https://github.com/shakshi12/Misleading-Covid-vaccination-Tweets
5https://www.theguardian.com/society/2020/apr/18/coronavirus-vaccine-trials-could-be-completed-by-mid-august
6https://www.bbc.com/news/world-us-canada-53899908
the discussions around Covid-19 vaccines both offline and online. We examine the type of discussions about COVID-19 vaccination which spread over Twitter since Twitter provides an easy to mine source of information representing all the important narratives. We collected tweets related to COVID-19 vaccination from September 2020 to March 2021. Most countries had begun vaccine rollouts\(^7\) as of March 2021. A significant mass of anti-vaxxers remain, nevertheless, gradually, people are less hesitant to get vaccinated\(^8\) and most of the current narratives and discussion points were already formed and matured in the aforementioned period. As such, we focus on the period leading up to large-scale vaccine rollout, which is critical for studying and understanding the nature of misinformation and its spread.

Given the Twitter application programming interface (API) restrictions, collecting an exhaustive dataset was impractical. As such, we aimed for a representative sample instead. We queried the Twitter Streaming API with a variety of relevant keywords: vaccine, covidvaccine, chinesevirus, VaccinesWork, No Masks, antivaxxer, antivaccine, antivax, COVID19, covax. Since there are no precise keywords for the misleading category, we also considered anti-vaccine tweets to cover a broader range of aspects even though not all misleading tweets are anti-vaccine and vice-versa. This resulted in over 200,000 tweets. After filtering the tweets,\(^9\) the final dataset used in this study has 114,635 tweets in English.

### B. Data Annotation Process

To perform the analysis, we further process and label the tweets as misleading and nonmisleading.

**Misleading**, a type of misinformation, is defined as fabricating an issue or an individual that is not true\([20],[21]\). Specifically, after applying standard NLP techniques to clean the tweets, we designate a tweet to be misleading when the content of a tweet deviates from the evidence shared by news media or reputable sources such as the World Health Organization (WHO), and even if it uses facts in parts, it might add connotations to it that encourages vaccine hesitancy. Otherwise, we consider the tweet as nonmisleading. For instance, the tweet, "I know its so bloody weird that Gates funded every single vaccine moderna Pfizer Oxford etc. Something my gut says dont have the vaccine." spins a narrative to increase people’s distrust. Thus, in this work we consider it as a misleading tweet. The partial truth in the misleading tweets could be the result of the widespread COVID-19 vaccination half-truths and myths.\(^{10}\) These narratives are designed to prompt vaccine hesitancy\([22]\) and eventually become the root cause of further misleading information.

Since data annotation is a costly and labor-intensive method, thus, first manual labeling of a sample subset of tweets is performed by human annotators, which is used for model parameter tuning and validation, when we explored the transfer learning\([23]\) approach to annotate each tweet in the complete dataset as misleading or nonmisleading. To that end, we first took a random sample of 1500 tweets from the dataset. This sampling was agnostic of the topics these tweets might be about. We then manually annotated them as either misleading or nonmisleading tweets. The sample tweets were labeled by three annotators. Then, using Fleiss’ kappa score\([24]\) \((k)\) to test the quality of the dataset annotation, an interannotator agreement amongst three annotators was performed, producing \(k = 0.84\), validating a high quality of agreement. The resulting labels had the ratio 47.7:52.3 among misleading and nonmisleading tweets.

We experimented with the sampling of tweets by selecting various sample sizes and utilizing fivefold cross-validation. Initially, we fine-tuned the pretrained XLNet transformer model\(^{11}\) with a sample size of 100. The results, as can be seen in Table I, are less promising and show the model’s underfitting. Next, we increased the sample size to 500 and observed an improvement over the previous performance. We continued with increasing the sample size. We observe that the performance dropped when the sample size was 2000. This is because while the training set’s accuracy, precision, recall, and other metrics all reached 0.92, the test set’s results were noticeably poorer. Consequently, we used 1500 samples to fine-tune the pretrained XLNet model. To demonstrate the stability of the results, we repeated these experiments five times but with different random seeds (1, 124, 2012, 2022, 1000).

After choosing 1500 as the sample size for the tweets, we then finally fine-tune the XLNet language model using the manually labeled tweets. XLNet is an extension of the Transformer-XL model and learns the bidirectional contexts. BERT and RoBERTa were also tested. XLNet outperformed all other models in our case. The 1500 manually annotated sampled tweets were split into an 80:20 train and test set utilizing a fivefold cross validation technique. Specifically, 300 out of 1500 tweets were used as a test set, and the rest were used as a training set. Table II, row 1 shows the evaluation metrics calculated on the test set. The results shown are the average value on the five test sets obtained using the fivefold cross validation technique. Subsequently, we used the fine-tuned XLNet model to annotate the entire dataset. To validate the efficiency of the annotated tweets, we manually verified 1000 random tweets. This sample was also balanced in terms of the predicted labels. Table II, row 2 shows the evaluation metrics of this validation. The accuracy for the sample was 0.86, indicating that our model had been well-trained, and we can rely on these labels to get further insights.

In Sections IV–VII, we will provide the analysis of the whole dataset, which is divided into a ratio of 70:30 nonmisleading and misleading tweets, respectively.

### IV. LANGUAGE EXPLORATION

In this section, we uncover ten distinct Syntactic styles of misleading and nonmisleading tweets.

\(^{7}\)https://www.cnbc.com/2021/03/25/covid-live-updates.html

\(^{8}\)https://www.ajmc.com/view/a-timeline-of-covid-19-vaccine-developments-in-2021

\(^{9}\)We removed tweets that are not relevant to COVID-19 or are not in English.

\(^{10}\)https://edition.cnn.com/2020/12/18/health/myths-covid-vaccine-debunked/index.html

\(^{11}\)https://huggingface.co/transformers/model_doc/xlnet.html
we use Kolmogorov Smirnov Test. The difference in both the distributions is indeed significant, we used the NLTK library to extract the part-of-speech to distinguish the structural patterns of both types of tweets. We found that the distributions are more significant than others in our dataset (Section III), we study ten syntactic aspects (attributes) which were found to be more significant than others in our dataset.

## A. Uncover Syntax

After assigning the labels for each tweet (as discussed in Section III), we study ten syntactic aspects (attributes) which were found to be more significant than others in our dataset to distinguish the structural patterns of both types of tweets. We used the NLTK library to extract the part-of-speech tags. First, we visualize the Syntactic distributions of both misleading and nonmisleading tweets. Next, to validate that the difference in both the distributions is indeed significant, we use Kolmogorov Smirnov Test.

First, we look at the **Nouns**, the main building blocks of any sentence. We observe from Fig. 1(a) that visually there is a slight variation in both distributions. To determine whether this difference in the distributions is statistically significant, we calculated the \( p \)-value of Nouns (see Table III, row 1), which is much lower than the significance level (5% as default value), which implies that the two distributions are in fact dissimilar. Second, **Pronouns** are the substitute for Nouns. Fig. 1(b) shows that Pronouns are more used in nonmisleading tweets than misleading tweets. Third, **Type-Token Ratio (TTR)** measures the lexical diversity (quality) of the text. Specifically, it is the ratio between the total number of unique words (types) in the text and the total number of words in the text. The higher the value of this ratio, the higher the lexical diversity of the text

\[
TTR = \frac{\text{# of Types}}{\text{# of Tokens}} \times 100. \quad (1)
\]

We notice in Fig. 1(c) that the distributions are right-skewed, indicating that the text is of good quality in terms of lexical diversity. However, we observe some differences between the distributions. When TTR is near to 100, the density of the misleading distribution is lower in comparison to its mean. Whereas, nonmisleading distribution has a similar density with respect to its mean. This implies that misleading tweets are less lexically diverse in contrast to nonmisleading tweets. In addition, the \( p \)-value present in Table III, row 3 is lower, indicating the distributions are different.

### We investigate other Syntactic aspects namely, **Stop words** [Fig. 1(d)], **Verbs** [Fig. 1(e)], **Conjunctions** [Fig. 1(f)], **Adverbs** [Fig. 1(g)], **Determiners** [Fig. 1(h)], **Adjectives** [Fig. 1(i)], and **WH-words** [Fig. 1(j)]. Their \( p \)-values present in Table III. Although the values for Determiners, Adjectives, and WH-words are near but less than the significance level, we consider them distinguishable attributes in finding misleading and nonmisleading tweets.

### B. What Are the Principal Topics of Discussion?

Next, using topic modeling using the latent Dirichlet allocation (LDA) approach [25], we inspect the top five most talked-about topics among misleading and nonmisleading tweets (see Table IV). We employed grid-search to identify the optimal number of topics, determining that to be five for both misleading and nonmisleading tweets for our data collection. Following are the key themes of misleading tweets.

1. **Politics:** The frequent targets of these tweets are politics; for example, the tweet - “I don’t even trust this Govt to take a vaccine they are desperate to sell us This makes me feel sad But now if cellchitis is also a side effect of it And blood clots And Morrison brushes that aside and or lies to us about it What?”. Here, the key point of discussion is not to believe the government, and thus misleads the readers by bringing political angle into the vaccination debate and creating this type of vaccine prejudice.

2. **Myths and Side Effects:** The false stories and myths about vaccines have the greatest effect on people’s minds; for instance, this tweet - “IF you are allergic to eggs and chicken, you are not going to receive a dose of the 1.02 million doses of Oxford-Astrazeneca vaccine that is expected in Kenya on Tuesday 2 March, 2021.”
Fig. 1. Syntactic analysis of the misleading and nonmisleading tweets. The x-axis represents the counts of the specific syntactic attribute (for example, noun) and y-axis represents the density of the distribution. TTR is an abbreviation for TTR. Orange and blue colors represent the misleading and nonmisleading labels, respectively. (a) Nouns. (b) Pronouns. (c) TTR. (d) Stop words. (e) Verbs. (f) Conjunctions. (g) Adverbs. (h) Determiners. (i) Adjectives. (j) WH-words.

TABLE IV

| Top Five Most Discussed Topics in Misleading and Non-Misleading COVID-19 Vaccination Tweets |
|-----------------------------------------------|-----------------------------------------------|
| Misleading                                   | Non-Misleading                                |
|Politics                                      | Operation Warp Speed                          |
|Myths & Side-Effects                          | Shots                                         |
|Vaccine Efficacy                              | Vaccine Efficacy                              |
|Role of Trump                                 | Real Side-Effects                             |
|Vaccine Choices                               | Data & Facts                                  |

attempts to steer readers away from the true story,$^{14}$ namely the delivery of said volume of vaccines under the COVAX initiative.

3) **Vaccine Efficacy:** There is a lot of confusion related to the COVID-19 vaccine efficacy in the discussions, for instance, “How can we trust the vaccine when the efficacy is not reported precisely??”

4) **Trump’s Role:** Trump is well known for being heavily involved in many false reports.$^{15}$ This tweet confirms Trump’s involvement in the tweet, “Why did the whitehouse turn down Pfizer offer of the vaccine as the 1st to receive it? Seems Mr. Trump joins hands with Operation Warp Speed and they are deliberately slowing the speed of vaccine rollout.”

5) **Vaccines Choice:** There is skepticism about choosing the vaccines. For instance, this tweet tells one of the myths about altering DNA for the Johnson & Johnson vaccine - “Do not get the Johnson & Johnson version. The MOA is different than that of Pfizer’s or Moderna’s vaccine. Picture a tennis ball inside a basketball. J&amp;J’s enters the nucleus (tennis ball) that can permanently alter DNA.”

The key themes of nonmisleading tweets are as follows.

6) **Operation Warp Speed:** The most discussed topic is Operation Warp Speed, initiated by the U.S. government to facilitate the development and distribution of COVID vaccines. An example of a tweet with this theme - “[username] Pfizer vaccine was self funded. Nothing to do with warp speed, the government or the lying ex president trump. TRUMP IS A LIAR.”

7) **Shots and Real Side-Effects:** These two themes are about informing an individual’s actual experience after getting the vaccine shot. We provide an example of a tweet that is related to both the themes - “I had the PFIZER shot(s), 30 days apart, and both times it didn’t hurt. My arm was sore later on for a day or two. 24 hours after the second shot I got a fever and chills and promptly went to sleep. Woke up 10 hours later feeling great. No other side effects. GET THIS VACCINE!”

8) **Vaccine Efficacy:** Both misleading and nonmisleading have a shared theme. This indicates that they are both discussing the vaccine’s efficacy. misleading tweets, on the other hand, aim to cause uncertainty regarding vaccine efficacy, while nonmisleading tweets emphasize the positive aspects of vaccine efficacy, such as - “#JabMe: A single shot of either the Pfizer or Oxford vaccine provides about 80 percent protection against being treated in a hospital, according to the latest data from the U.K. vaccination program.”

9) **Data and Facts:** Nonmisleading tweets are more concerned with presenting accurate information, such as - “BBC: Around 5 million Europeans have already received the AstraZeneca vaccine. Of this figure, about 30 cases had reported ‘thromboembolic events’ - or developing blood clots. European medicines regulator said there was no indication the jab was causing the blood clots.”

The top five themes indicate the different subjects explored in both types of tweets. Precisely, Misleading tweets mostly mislead the reader using political dimension or raise fear among the people for vaccination. In comparison, Non-Misleading tweets discuss the real side-effects of the vaccination and try to bring the facts with evidence.

---

$^{14}$https://www.unicef.org/kenya/press-releases/over-1-million-covid-19-vaccine-doses-arrive-nairobi-via-covax-facility

$^{15}$https://www.cnbc.com/2021/01/13/trump-tweets-legacy-of-lies-misinformation-distrust.html
V. OPINION STUDY

Previously, we explored the Syntactic dimension. We now look into the second dimension, the role of opinion in relation to misleading and non-misleading tweets.

A. Sentiment Matters

We first explore the impact of sentiments on misleading and non-misleading tweets, considering three broad categories of sentiments: 1) positive; 2) negative; and 3) neutral. The sentiments are calculated using valence aware dictionary and sEntiment reasoner (VADER) API [26], [27]. To detect a tweet as positive, negative, or neutral, we utilized the compound score. The compound score is the sum of the valence scores of each word, yielding a range of $[-1, 1]$. A score of $-1$ indicates a strong Negative sentiment, whereas a score of $+1$ suggests a strong Positive sentiment. We utilized a 0.05 threshold value. The selection of the 0.05 threshold has been adopted from the previous literature [28], [29]. In addition, to validate the choice of threshold, we randomly sampled 150 tweets with a balanced combination of positive, negative, and neutral sentiment labels obtained from the VADER API (50 random tweets from each label). Following that, we manually annotated the sentiments for 150 tweets and got an accuracy of 0.96, confirming the efficacy of the selected threshold. Thus, a tweet is considered to have a positive sentiment if the compound score is greater than or equal to 0.05, neutral if the score is between 0.05 and -0.05; negative otherwise.

Fig. 2 shows that negative sentiments are more prevalent in misleading tweets followed by positive and neutral sentiments. An example of a misleading tweet with positive sentiment is “A little Angel in my dreams today told that our bodies will be developing antibodies on its own within a few days without vaccination.” While a reading of it indicates vaccine prejudice and skepticism, the sentiment analysis tool latches upon the positive sounding phrases in there.

We then identify the topics which are being discussed with respect to sentiments. Apart from vaccine efficacy which confirms the above mentioned tweet’s theme, the related topics operation warp speed and trials are also identified within the positive sentiments of misleading tweets (See Table V). We surmise that misleading tweets with positive sentiments inject the negativity by sugar-coating the tweets with positive words to easily trick people into either believing in their positive hypothetical situations or providing a new dimension to the topic.

Positive sentiments, on the other hand, dominate in nonmisleading tweets, though a substantial portion of them again have negative or neutral sentiments. An instance of a nonmisleading tweet with negative sentiment is “Dr Kathrin Jansen, Pfizer’s head of vaccine development: We were never part of the Warp Speed …We have never taken any money from the U.S. government, or from anyone. Trump is a liar.” In this instance, the negative sentiment of the nonmisleading tweet is due to it counteracting the misleading information. Many facts and news related to the pandemic have naturally negative sentiments. Similar to this tweet argument, topics that are discovered in the negative sentiments of nonmisleading tweets are operation warp speed and vaccine efficacy, in addition to, trials and data and facts (See Table V). These nonmisleading tweets with negative sentiments indicate that they are either attempting to clarify claims against COVID vaccination’s development companies or myths against the vaccination process with their choice of negative words.

Overall, negative sentiments are more common in misleading tweets, whereas, nonmisleading tweets have more positive sentiments. We go through five emotions in detail in Section V-B.

B. Intense Emotions

The sentiments serve as the foundation for analyzing the tweets. As a result, we dig deeper into the impact of emotions on tweets.

Fig. 3 displays the five different emotions - anger, fear, happiness, sadness, and surprise retrieved by employing the text2emotion API [30]. In this API, for each tweet, we get a

![Fig. 2. Sentiment analysis with respect to misleading and non-misleading tweets. The x- and y-axis denote the labels of the tweets and percentages, respectively. Negative, neutral, and positive are different categories of the sentiments.](image)

![Fig. 3. Emotion analysis of the misleading and non-misleading tweets. Each axis denotes the emotion. The five emotions depicted are happiness, fear, anger, surprise, and sadness. Non-misleading tweets are represented in green color, and misleading in red color (best seen in color).](image)

| Topic Modeling with Respect to Each Sentiments, M and NM Represent Misleading and Non-Misleading, OWS and VaEf Denote Operation Warp Speed and Vaccine Efficacy, Respectively |
|---------------------------------|---------------------------------|---------------------------------|
| Positive | Negative | Neutral |
| Trials, OWS, VaEf | VaEf, OWS, Trials, Myths | VaEf, OWS, Trials |
| Trump, Real side-effects, VaEf | Data & Facts, OWS, VaEf, Trials | Data & Facts, VaEf, Real side-effects |

TABLE V
dictionary where keys denote emotion categories, and values indicate its score. Higher the score, the higher the emotion in that tweet. We classify the tweets as per the emotion that has the highest score. In the figure, the value with respect to each emotion axis corresponds to the percentage of a particular emotion in a (Non-)Misleading class.

In misleading tweets, the most common emotion is fear, followed by surprise, sadness, happiness, and, finally, anger. Whereas, nonmisleading tweets have a tie for the first place with fear and surprise, followed by happiness, sadness, and anger.

Fear and surprise are the two most popular emotions across both types of tweets. It is understandable given that 45% of unvaccinated people are afraid to get the vaccine because they are worried about the adverse side-effects. To confirm this, we studied the topics around fear and surprise emotions. The topics which are similar to the above statement are trials and vaccine efficacy in both misleading and nonmisleading categories (See Table VI). However, the emotion fear is higher in misleading tweets in contrast to nonmisleading tweets, which is attributable to the observation that most misleading tweets reference fake and fabricated vaccine side-effects and false narratives of operation warp speed (Table VI). In contrast, the emotion surprise is higher in nonmisleading tweets than misleading tweets. These often discuss the governments’ fast response toward vaccination, fitting into the data and facts topic.

Furthermore, emotions, anger, and sadness are higher in misleading tweets. One of the possible reasons could be that these tweets often involve a political dimension and accuse the government of not making the right decisions. The topics covering under both emotions are politics, vaccine availability, and operation warp speed in the misleading category. Nonmisleading tweets that have emotion happiness discuss their experience about receiving the shot and facing no bogus side-effects spreading across the Internet. A related topic is real side-effects in the nonmisleading category. To summarize, the majority of the misleading tweets have Fear emotions more than nonmisleading tweets.

### VI. INFLUENCE OF VISIBILITY

So far, our analysis was confined to the content of the tweets themselves. The focus of the third dimension is looking at information and meta-data in the tweets that influence their visibility, such as words used, hashtags, likes to study whether there are distinctive characteristics across misleading and nonmisleading tweets.

#### A. Merry Words of Twitter

Certain words are used more frequently in the tweets than others. In Fig. 4(a) and (b), we use Word Clouds to summarize this for both types of tweets visually. The relative frequency of the words is reflected in the size of the words. The figures show that the most frequent words in both Word Clouds are completely different, indicating that the choice of the words in both types of tweets significantly varies. Shot, report, jab, ill, sore, and fact are all recurring words in the nonmisleading Word Cloud, whereas, wait, death, risk, effect, trump, and die are all frequent words in the misleading one.

To study this difference quantitatively, we computed the Kendall Tau correlation coefficient on the union of the top 50 misleading and nonmisleading words. Only seven words were found to be common in both classes. A score of $-0.81$ was observed, showing disagreement between the word groups (the Kendall Tau range is $[-1, 1]$, with $-1/1$ indicating strong dis/agreement, respectively). This clearly suggests that frequently recurring words are unrelated, implying that the word choices in both classes differ.

#### B. Much Ado About Hashtags

Extensive usage of hashtags is a popular way to enhance the exposure of the tweets. We investigate them from two perspectives.
1) **Unique Hashtags:** We explore such hashtags that are relatively unique to misleading versus nonmisleading tweets. Table VII lists some of the popular ones. Note that the hashtags mentioned in the table are chosen depending on how many times they appear in the tweets. In misleading tweets, the #untestedvaccine clearly indicates that the tweet refers to one of the vaccine myths. In contrast, the #vaccinatedandproud represents that tweet is in support of the vaccination process. Thus, the choice of the hashtags can provide a clue about the misleading tweets.

2) **Co-Hashtags:** We also consider the combination of hashtags that frequently occurred together in a tweet, i.e., co-hashtags. In non-misleading tweets, we find 280 co-hashtags, while 86 co-hashtags are found in misleading tweets. After filtering those co-hashtags that occurred more than once, we found that co-hashtags repeatedly occurred only in nonmisleading tweets. There is no pattern (consistency) regarding co-hashtags in misleading tweets, making their hashtags more random.

### C. As You Like It

The number of retweets, replies, and likes count are all essential visibility attributes. Fig. 5 depicts the median (mean) values of the counts of retweets, replies, and likes for both classes. In our case, the median and mean values are the same. On average, replies to the tweets remain the same regardless of the type of information it contains. However, there is a variation in the retweets count and likes count. Relatively, the misleading tweets get fewer retweets and likes than nonmisleading tweets.

### D. What’s in a Name?

The names of the COVID-19 vaccinations are frequently mentioned in tweets. In this regard, we attempt to assess the influence of vaccine names in both misleading and nonmisleading tweets. In our collected data, we discovered that the majority of the discussions revolve around the five vaccinations: 1) Pfizer; 2) Moderna; 3) AstraZeneca; 4) Covaxin; and 5) Johnson & Johnson. These names are used either in an individual or combined manner.

Fig. 6 shows that the proportion of misleading tweets is lower than the proportion of nonmisleading tweets until the number of vaccine names is fewer than or equal to three. When the count reaches four or five, the number of Misleading tweets begins to rise, i.e., when the number of vaccine names in a tweet grows to more than three, the likelihood of a Misleading tweet also grows.

### VII. Classification of Misleading Tweets

By providing the tweets as an input to the pretrained XLNet model, the focus was to obtain the labels for the tweets. These labels are treated as “ground-truth” for the prediction task. In this section, instead of using the tweets themselves, we use the descriptive features described in Sections IV–VI as an input to the various ML models to classify misleading and nonmisleading tweets. The purpose is to check if these features can distinguish misleading tweets. This helps us explicitly understand the differentiating characteristics across nonmisleading tweets. Specifically, the features constitute - Stop words, Pronouns, Nouns, Adjectives, Average length, WH-word, Adverbs, Conjunctions, Verbs, Determiners, TTR, Sentiments, Emotions, and Hashtags. For classification, we treat distinct sentiment categories as one feature by utilizing their raw scores. In emotions, there are dictionaries where keys denote emotion categories, and values indicate their scores. Higher the score, the higher the emotion in that tweet. Unlike sentiments, these emotion scores do not have a particular range for all categories. They cannot be readily used as a feature value in the emotions category. Thus, we applied one-hot encoding to the emotions feature.

On the train set, we use fivefold cross-validation, which covers 80% of the data. We utilized cross-validation as
implemented in Sklearn. The whole dataset (114,635) is divided into train and test sets. The \( k \)-fold cross-validation has been applied to the train set to find the optimal parameters of the model, and then the test set is used to evaluate the final performance of the trained model. Following the same strategy, the 80% train set (91,708) was used for fivefold cross-validation, and the final evaluation was performed on the 20% test set (22,927). To train and assess the model, each fold splits the train set into fold-train and fold-test sets. Finally, we evaluate the trained model’s performance on an unseen test set that accounts for 20% of the total data. Each train and test set is balanced. The evaluation metrics for the test set are shown in Table VIII in descending order of accuracy. In our example, the ensemble-based model, random forest (RF), performs best with an accuracy of 0.90, followed by extra trees (XTS), decision tree (DT), and so on. This shows that writing styles may effectively distinguish between misleading and non-misleading tweets. Other models’ findings, such as precision, recall, F1 score, and AUC ROC score, show that our results are consistent throughout, showing that the trained models are generalizable. The best-performing model corresponds to a learning rate of 0.01, using 100 trees with a random state of 1.

### A. Feature Importance

The SHAP Explainable AI (XAI) tool is then used to evaluate the contributions of each feature in the classification task. By computing the average marginal contributions of each feature, this tool helps in finding the relevant features in the prediction. Fig. 7 shows the importance of the features (SHAP ranking) in descending order. Sentiments, the most important contribution, have a negative influence on prediction, suggesting that a lower Sentiments value predicts a Misleading class and vice versa. This makes sense because higher Sentiments values indicate positive sentiments and lower values indicate negative sentiments, which is in line with the section Opinion Study’s conclusion that misleading tweets include more negative sentiments. In addition, Nouns, Emotions, and Conjunctions features have a negative influence on misleading tweets. This means that compared to nonmisleading tweets, misleading tweets include fewer Nouns and Conjunctions. This might be because the objective of the misleading tweets is to entice readers by utilizing fancy words or catchy phrases rather than delivering accurate information using Nouns and Conjunctions. The remaining features have a positive influence; for example, compared to nonmisleading tweets, misleading tweets have a larger number of Pronouns.

#### 1) Feature Ablation Study: After evaluating the importance of features, we try to see if there is a decline in accuracy if we exclude some features. We experiment with a few settings, eliminating certain variables based on their importance as suggested by the SHAP ranking in Fig. 7, and then rerunning all of the models in the same environment. Due to space limits, we only provide the results of the best-performing model, RF. The best accuracy attained with all features is 0.90. We start by removing Emotions and the rest of the features listed below Emotions as per the SHAP plot. Table IX, row 1 summarizes the findings. When we remove these features from our dataset, the values of the evaluation metrics decline, indicating that they are truly relevant. Next, we remove features that are less important than Emotions, such as Type token ratio and

---

**TABLE VIII**

| Models/Metrics | ACC  | PR   | RC   | F1   | AUC  |
|----------------|------|------|------|------|------|
| **RF**         | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 |
| **XTS**        | 0.90 | 0.90 | 0.89 | 0.89 | 0.90 |
| **DT**         | 0.88 | 0.88 | 0.88 | 0.84 | 0.88 |
| **XT**         | 0.88 | 0.88 | 0.85 | 0.87 | 0.86 |
| **BG**         | 0.87 | 0.87 | 0.87 | 0.87 | 0.88 |
| **NuSVC**      | 0.75 | 0.76 | 0.76 | 0.76 | 0.76 |
| **XGB**        | 0.75 | 0.75 | 0.75 | 0.77 | 0.77 |
| **KNN**        | 0.73 | 0.73 | 0.73 | 0.73 | 0.72 |
| **LGBM**       | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |
| **ADB**        | 0.70 | 0.70 | 0.70 | 0.70 | 0.70 |

---

**TABLE IX**

| Features/Metrics | ACC  | PR   | RC   | F1   | AUC  |
|------------------|------|------|------|------|------|
| w/o Emo & BF     | 0.86 | 0.86 | 0.86 | 0.86 | 0.86 |
| w/o TTR & BF     | 0.87 | 0.87 | 0.88 | 0.88 | 0.88 |
| w/o Adj & BF     | 0.88 | 0.88 | 0.88 | 0.88 | 0.89 |
| w/o Hashtags     | 0.89 | 0.89 | 0.89 | 0.89 | 0.89 |
| w/ ALL features  | 0.90 | 0.90 | 0.90 | 0.90 | 0.90 |
the remainder of the features (shown in Table IX, row 2). We continue to run experiments and discover that even the least significant feature, Hashtags, contributes to the model’s improvement. These results indicate that all features are both valuable and necessary for detecting Misleading tweets.

2) Correlation and the SHAP Ranking: Is there any association between the features’ correlation values and the SHAP ranking? The hypothesis is that the highly correlated features should be close in the SHAP ranking. The correlation between each feature pair is shown in Fig. 8. The dark color denotes a strong correlation between the two features based on the absolute value. The correlation between the features’ correlation values and the SHAP ranking? The hypothesis is that the highly correlated features thus, farther from each other in SHAP ranking, demonstrating that our hypothesis is indeed true.

VIII. CONCLUDING REMARKS

In this article, we carried out an exploratory analysis and meta-data associated with tweets pertaining to COVID-19 vaccines to determine the characteristics of both Misleading and Non-Misleading tweets. The topic detection aspect of our study helped establish the main themes of discourse across these categories, as well as identify potentially distinguishing characteristics. The latter were explored as features to carry out a classification task, where the observed outcomes support explainability. We observe that this explainability property, coupled with the aforementioned identification of the topic of tweets, actionable intelligence can be generated, which determines a principal thrust of our future work.

In future, we also want to explore whether the approach laid out can be generalized to identify Misleading tweets on other topics beyond COVID-19 vaccination. The current work, given its limited scope, consequently dealt with a simple notion and dichotomy of non-misleading information. Precisely defining more complex forms of misleading information or fake news itself is a challenge, amplifying, in turn, the challenges of the task of classification. Beyond the extension of the work to span other topics, there is also an opportunity to refine the techniques by carrying out an analysis that is fine-grained in geographic, temporal, and linguistic dimensions. For example, to analyze which Misleading tweets are more prominent and specific to certain regions, which of them persist over what span of time, and doing so in languages beyond English.

REFERENCES

[1] J. Bosman, J. Hoffman, M. Sanger-Katz, and T. Arango, Who are the Unvaccinated in America? There’s no One Answer. New York, NY, USA: New York Times, 2021.
[2] The Disinformation Dozen Center for Countering Digital Hate, London, U.K., 2021.
[3] S. Machingaidze and C. S. Wiysonge, “Understanding COVID-19 vaccine hesitancy,” Nature Med., vol. 27, no. 8, pp. 1338–1339, Aug. 2021.
[4] D. O. Mesa. (2021). Report 43 Quantifying the Impact of Vaccine Hesitancy in Prolonging the Need for Non-Pharmaceutical Interventions to Control the COVID-19 Pandemic. [Online]. Available: https://www.imperial.ac.uk/mrc-global-infectious-disease-analysis/covid-19/report-43-vaccine-hesitancy/
[5] W. Wu, H. Lyu, and J. Luo, “Characterizing discourse about COVID-19 vaccines: A Reddit version of the pandemic story,” Health Data Sci., vol. 2021, pp. 1–11, Aug. 2021.
[6] L. G. Malagoli, J. Stancioli, C. H. G. Ferreira, M. Vasconcelos, A. P. C. da Silva, and J. M. Almeida, “A look into COVID-19 vaccination debate on Twitter,” in Proc. 13th ACM Web Sci. Conf., Jun. 2021, pp. 225–233.
[7] F. Pierri et al., “VaccinItaly: Monitoring Italian conversations around vaccines on Twitter and Facebook,” 2021, arXiv:2107.03757.
[8] M. R. DeVerna et al., “CoVaxxy: A collection of English-language Twitter posts about COVID-19 vaccines,” in Proc. Int. AAAI Conf. Web Social Media, vol. 15, May 2021, pp. 992–999.
[9] E. Chen, K. Lerman, and E. Ferrara, “Tracking social media discourse about the COVID-19 pandemic: Development of a public coronavirus Twitter data set,” *JMIR Public Health Surveill.*, vol. 6, no. 2, May 2020, Art. no. e19273.

[10] S. Sharma, E. Agrawal, R. Sharma, and A. Datta, “FaCov: COVID-19 viral news and rumors fact-check articles dataset,” in *Proc. Int. AAAI Conf. Web Social Media*, vol. 16, 2022, pp. 1312–1321.

[11] S. Sharma and R. Sharma, “Identifying possible rumor spreaders on Twitter: A weak supervised learning approach,” in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2021, pp. 1–8.

[12] S. Butt, S. Sharma, R. Sharma, G. Sidorov, and A. Gelbukh, “What goes on inside rumour and non-rumour tweets and their reactions: A psycholinguistic analyses,” *Comput. Hum. Behav.*, vol. 135, Oct. 2022, Art. no. 107345.

[13] M. Dhawan, S. Sharma, A. Kadam, R. Sharma, and P. Kumaraguru, “GAME-ON: Graph attention network based multimodal fusion for fake news detection,” 2022, arXiv:2202.12478.

[14] M. Mayank, S. Sharma, and R. Sharma, “DEAP-FAKED: Knowledge graph based approach for fake news detection,” in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, 2022, pp. 1–8.

[15] R. Jagtap, A. Kumar, R. Goel, S. Sharma, R. Sharma, and C. P. George, “Misinformation detection on YouTube using video captions,” 2021, arXiv:2107.00941.

[16] E. Ferrara, “What types of COVID-19 conspiracies are populated by Twitter bots?” 2020, arXiv:2004.09531.

[17] R. Kouyi et al., “Coronavirus Goes Viral: Quantifying the COVID-19 misinformation epidemic on Twitter,” *Curves*, vol. 12, no. 3, pp. 1–9, Mar. 2020.

[18] J. Ma and L. Stahl, “A multimodal critical discourse analysis of anti-vaccination information on Facebook,” *Library Inf. Sci. Res.*, vol. 39, no. 4, pp. 303–310, Oct. 2017.

[19] H. Rosenberg, S. Syed, and S. Rezaie, “The Twitter pandemic: The critical role of Twitter in the dissemination of medical information and misinformation during the COVID-19 pandemic,” *Can. J. Emergency Med.* vol. 22, no. 4, pp. 418–421, Jul. 2020.

[20] M. D. Molina, S. S. Sundar, T. Le, and D. Lee, “‘Fake news’ is not simply false information: A concept exposition and taxonomy of online content,” *Amer. Behav. Scientist*, vol. 65, no. 2, pp. 180–212, 2021.

[21] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, “Detection and resolution of rumours in social media: A survey,” *ACM Comput. Surv.*, vol. 51, no. 2, pp. 1–36, 2018.

[22] A. Cossard, G. De Francisci Morales, K. Kalimeri, Y. Mejova, D. Paolotti, and M. Starnini, “Falling into the echo chamber: The Italian vaccination debate on Twitter,” in *Proc. Int. AAAI Conf. Web Social Media*, vol. 14, May 2020, pp. 130–140. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/7285

[23] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2009.

[24] M. L. McHugh, “Interrater reliability: The Kappa statistic,” *Biochem. Med.*, vol. 22, no. 3, pp. 276–282, 2012.

[25] K. Xu, F. Wang, H. Wang, and B. Yang, “Detecting fake news over online social media via domain reputations and content understanding,” *Tsinghua Sci. Technol.*, vol. 25, no. 1, pp. 20–27, Feb. 2020.

[26] U. Yaqub, “Tweeting during the COVID-19 pandemic: Sentiment analysis of Twitter messages by president Trump,” *Digit. Government, Res. Pract.*, vol. 2, no. 1, pp. 1–7, Jan. 2021.

[27] J. Waters, N. Nicolaidou, D. Stefanidis, H. Efthathiades, G. Pallis, and M. Dikaiakos, “Exploring the sentiment of entrepreneurs on Twitter,” *PLoS ONE*, vol. 16, no. 7, Jul. 2021, Art. no. e0254337.

[28] V. Bonta, N. Kumares, and N. Janardhan, “A comprehensive study on lexicon based approaches for sentiment analysis,” *Asian J. Comput. Sci. Technol.*, vol. 8, no. S2, pp. 1–6, Mar. 2019.

[29] C. A. A. Cruz and F. F. Balahadha, “Analyzing public concern responses for formulating ordinances and lawsusing sentiment analysis through VADER application,” *Int. J. Comput. Sci. Res.*, vol. 6, pp. 842–856, Jan. 2022.

[30] S. Dhar and I. Bose, “Emotions in Twitter communication and stock prices of firms: The impact of COVID-19 pandemic,” *Decision*, vol. 47, no. 4, pp. 385–399, Dec. 2020.

**Shakshi Sharma** received the master’s degree from the National Institute of Technology (NIT), New Delhi, India, in 2017. She is currently pursuing the Ph.D. degree with the Computational Social Science Group, Institute of Computer Science, University of Tartu, Tartu, Estonia.

From July 2018 to May 2020, she worked as an Assistant Professor at the Department of Computer Science and Engineering, The NorthCap University, Gurugram, Haryana, India. After earning her master’s degree, she spent around a year working as a Programmer at one of the multinational companies (MNC): Fidelity International Company, Gurugram. Her research interests include the problem of misinformation on online social media platforms. Specifically, analyzing mis(dis)information from various data sources and utilizing multiple AI and natural language processing (NLP) techniques. In addition, she is working in the field of AI ethics, focusing on the interpretability of black-box models and data biasness.

**Rajesh Sharma** received the master’s degree from the Indian Institute of Technology (IIT) Roorkee, Roorkee, India, in 2006, and the Ph.D. degree from Nanyang Technological University, Singapore, in December 2013.

He joined the University of Tartu, Tartu, Estonia, in August 2017, and worked as a Senior Researcher (equivalent to Associate Professor) till December 2020, where he has been working as an Associate Professor and leads the Computational Social Science Group, Institute of Computer Science, since January 2021. From January 2014 to July 2017, he has held a Research Fellow and a Post-Doctoral Researcher positions at the University of Bristol, Bristol, U.K., Queen’s University, Belfast, U.K., and the University of Bologna, Bologna, Italy. He has also worked in the IT industry for about 2.5 years after completing his master’s degree. His research interests include understanding users’ behavior, especially using social media data. His group often applies techniques from AI, natural language processing (NLP), and most importantly, network science/social network analysis.

**Anwitaman Datta** is currently an Associate Professor with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. His core research interests span the topics of large-scale resilient distributed systems, information security, and applications of data analytics. Presently, he is exploring topics at the intersection of computer science, public policies, and regulations along with the wider societal and (cyber)security impact of technology. This includes the topics of social media and network analysis, privacy, cyber-risk analysis, and management, cryptocurrency forensics, the governance of disruptive technologies, as well as impact and use of disruptive technologies in digital societies and government.