Sentiment Analysis of Long-term Social Data during the COVID-19 Pandemic

Sophanna Ek, Marco Curci, Xiaokun Yang[‡], Beiyu Lin[†], Hailu Xu
Department of Computer Engineering and Computer Science, California State University, Long Beach
[‡]College of Science and Engineering, University of Houston, Clear Lake
[†]Department of Computer Science, University of Texas, Rio Grande Valley
Email:{sophanna.ek, marco.curci}@student.csulb.edu, yangxia@uhcl.edu,
beiyu.lin@utrgv.edu, hailu.xu@csulb.edu

***Abstract*—The COVID-19 pandemic has brining the "infodemic" in the social media worlds. Various social platforms play a significant role for instantly acquiring the latest updates of the pandemic. Social medias such as Twitter and Facebook produce vast amounts of posts related to the virus, vaccines, economics, and politics. In order to figure out how public opinion and sentiments are expressed during the pandemic, this work analyzes the long-term social posts from social media and conducts sentiment analysis on tweets within 12 month. Our findings show the trend topics of long-term social communities during the pandemic and express people's attitudes towards progress of major actions during the pandemic. We explore the main topics during the prolonged pandemic, including information surrounding economics, vaccines, and politics. Besides, we show the differences in gender-based attitudes and propose future research questions refer to the "infodemic" in the social world. We believe that our work contributes to attract public attention to the "infodemic" of the social crisis.**

***Index Terms*—COVID-19; Pandemic; Social Network Data;**

I. Introduction

The global outbreak of COVID-19 pandemic has thoroughly changed and disrupted the human's lives in the past year. By March 2021, it has spread to 223 countries, reached to 119 million cases and more than 2 million deaths [8]. This disease has overwhelmed the entire world and changed the human's normal activities. However, despite of the spreading of real virus worldwide, social medias, as a key tool that majority people receive and post news or information in the daily, have been inevitable overwhelmed by the COVID-19 related information. According to the director of World Health Organization, COVID-19 is not only a global pandemic, but also an infodemic that threatens everyone in the social world [6]. Social posts such as tweets in Twitter and posts in Facebook instantly spread strong emotions, such as fear, stress, anger, despair, or hope, and these emotions are usually accompanied by vaccine-related news or government activities. Interaction via social medias tights everyone in the world during this special pandemic time.

Social networks show strong emotional reactions towards the pandemic in the past year. During the pandemic, a large number of social posts related to rumors, hate speech, racist conspiracy, and negative sentiments had quickly proliferated on the social networks. For example, studies tracked the discriminatory comment spread on Twitter through March 2020 and showed that related racial attacks significantly increased during April 2020 [13]. Many studies have shown that the infodemic greatly threatens the normal social life and create directly affects to the economic and politics [13], [17].

Many previous studies have provided various insights into the analysis of sentiment in a specific timeline during the COVID-19 pandemic. For example, they analyzed the sentiment engagement of social bots in Twitter by following the data from January and March, 2020 [17]; Kruspe et al. [11] explored the cross-language sentiment of COVID-19 related tweets during December 20219 to April 2020; By collecting the data from March 2020, Lyn et al. [13] presented significant differences between Twitter users based on the controversial/non-controversial terms. However, previous work generally focused on a short period of time, usually ranging from several days [13], [14] to a few months [9]–[11], where they only provided views on a single stage of the pandemic. In this pandemic, which has lasted for more than a year, a single stage analysis of sentiment cannot thoroughly reveal the overall perspectives of the infodemic in social medias.

Different from previous studies, we analyze the social media data from a long-term perspective. In this work, we focus on the sentiment analysis of Twitter' tweets throughout the one year. First, we work with a collection of 12 months of Twitter data related to COVID-19 during the pandemic. Second, we provide a long-term analysis of social sentiment from the COVID-19 social data and display the overview of infodemic in the past year. Third, we provide various views of the social data by classifying social data into different topics, including vaccine-related, politic-related, and economic-related. And we analyze the variances in opinions of gender groups. Finally, we discuss the possible future research questions.

II. Related Work

Many previous papers focused on the various perspectives of social data from different scopes. For example, existing work targeted on the activities of misinformation or controversial terms in social medias [13], [18], [19]. They effectively identified the malicious activities by utilizing the content analysis [18], [19], social connections [13], various sentiment analysis [17], and the supports from various platforms [20]. During the pandemic, many works had conducted to provide various

perspective of sentiment analysis for COVID-19 related data or other social data in different levels. Pennycook et al. [15] investigated the individual's inattention plays a key role in the spreading of Covid-19 misinformation, where content-neutral intervention can benefit for discerning the true or false news. Shi et al. [17] proposed that for almost all topics refer to the pandemic, social bots and real users shared a similar trend on sentiment polarity and the former can be worse in the negative topics. Pano and Kashef [14] investigated the various text strategies for correlating the tweet with Bitcoin prince during the pandemic from May to July 2020. They proposed that the Bitcoin price correlate well with the tweet sentiment over shorter timespans, usually a single day. Brennen et al. [9] analyzed the misinformation in social media relevant to COVID-19 between January and the end of March 2020. They found that independent fact-checker rises quickly during the early of pandemic and around 59% of misinformation has been reconfigured manually. Kuchler et al. [12] analyzed the geographic spread of pandemic relevant data in Facebook by March 2020. They proved that the geography of social connections can be an important predictor of outbreaks during the pandemic and can improve the out-of-sample predictions of the COVID-19 spreading [12]. Shahsavari et al. analyzed the narrative framework of rumors and conspiracy related to the pandemic and they found the dynamics of storytelling can help to monitor the spreading of misinformation [16].

III. Data Collection and Analysis Methods

In this section, we discuss the method that we used to collect relevant data from Twitter and introduce the analytical methods that are used in the sentiment analysis of tweets.

A. Data collection

We collect COVID-19 relevant tweets for sentiment analysis by creating custom tweet scraper using the Twitter APIs [7]. We do not choose to use the existing datasets by the following reasons: first, the existing online datasets are scatted and have no strong relationship with the other data sets. They are not continuous in time and their main topics are different; second, the existing datasets usually cover only a very short period of time, such as a few days or a month, which hardly provide a thorough view of the long-term situation of the pandemic. The related tweets were collected using the Tweepy API [7]. The tweets selection was filtered by using the relevant keywords including "covid-19" and "vaccine" or any hashtags such as "#covid-19" or "#coronavirus". The data was collected on weekly basis from the 1st - 7th, 8th - 14th, 15th - 21st, and 22nd - 28th/30th periods for 12 months. The volume of the weekly collection was varied between 25,000 to 35,000. The total collected tweets up to approximately 1,300,000 tweets from February 2020 to February 2021.

B. Analysis methods

We performed the data preprocessing on each data instance before performing the data analysis. All collected tweets had been preprocessed by removing url link, user references, punctuations and hashtags symbol. The text was then tokenized and have common English stop words removed. The cleaned text was then applied the stemming process using Porter Stemmer to remove the morphological affixes from words and leaves only the word stem for our analysis. The text was then applied the lemmatization to keep the word to its meaningful base form. We also use the stop words, Porter Stemmer, and Lemmatization from the NLTK Library (Natural Language Toolkit) [3] in python for the data processing.

1) TF-IDF: TF-IDF method was used to find more important topics from the collected tweet data using the TF-IDF score. Instead of giving every word with equal importance, TF-IDF gives more importance to the words that occur more frequently in one document and less frequently in other documents. TF-IDF score is determined by conducting the word's term frequency and its inverse document frequency. TfidfVectorizer in sklearn python library [4] was used to learn the pre-processed tweet data and score each term appear in the tweet corpus. A couple of important parameters were used to fit the tweet data with TfidfVectorizer. The ngram_range, max_features, and max_df. The ngram_range determines the range of n-vlaues for different n-grams to be extracted. We use the unigrams and bigrams in our sentiment analysis. Max_features parameter determines the top-n vocabulary that will be built for this fitting. Max_df ignore terms that have a document frequency strictly higher than the given threshold. Max_df is useful since it will ignore terms that are too common and less important. For this experiment, the top 20 max_features was used. After getting weight for each top 20 keywords/topics, 8 topics were then used to get their term frequencies in our COVID-19 tweet dataset. Figure 1 show the proportion of each topic appearing the tweet dataset from February 2020 to February 2021.

2) Sentiment Analysis: The sentiment analysis is performed on the same pre-processed tweet dataset. TextBlob [5] is a python library that can perform the common natural language process tasks such as sentiment analysis, classification, noun phrase extraction, and more. It assigns individual scores to all the words, then takes an average of all the sentiments to calculate the final sentiment. It takes the pre-processed tweet content and gives the polarity score which is used to determine the sentiment of the tweet data. The score is range from -1 to 1. The content is said to be negative if the polarity score is less than 0, positive if it is greater than 0, and neutral if it is equal to 0.

IV. Sentiment Analysis

In this section, we perform the sentiment analysis with the collected data and investigate the sentiment patterns of content-level, political and social cognitive attributes. We provide perspectives on social data in three categories: vaccine-related, politic-related, and economic-related. Besides, we explore the differences of views among gender effects.

We first analyze the frequency of various keywords among all the collected tweets. Figure 1 displays the frequency of each

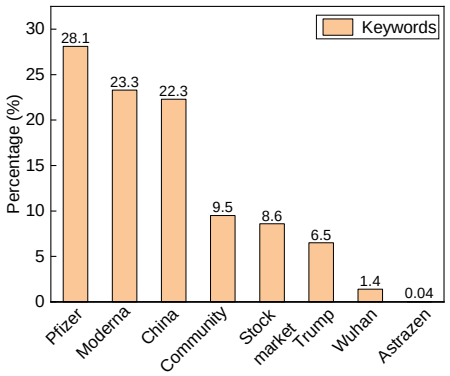

Fig. 1: Percentage of various keywords appears in the tweets from Feb 2020 to Feb 2021.

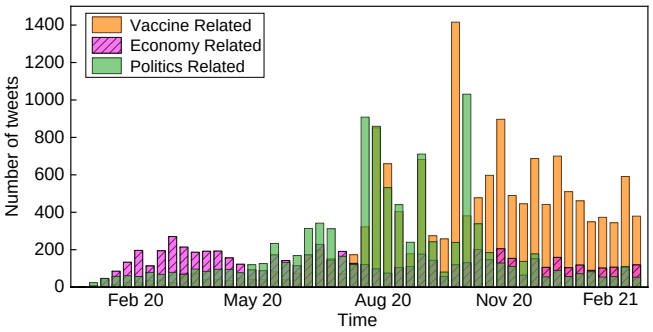

Fig. 2: Trends of positive, negative, and neutral tweets during the pandemic.

keyword in the tweets. Note that we remove the keywords refer to "covid19" or "corona virus" since all the tweets contain such terms during the data collection process. From Figure 1, we can observe that the vaccine information takes the most important role in the past of year. The two types of major vaccines such as Pfizer and Moderna appear in half of the tweets in the past year. Except the vaccine related tweets, around 30.3% tweets refer to the politics or region such as "China", "Trump", and "Wuhan". Community spreading of virus also takes an important role in the social world since around 9.5% tweets focus on community spread. Besides, the economic topics such as "Stock market" take around 8.6% in all tweets.

Next, we investigate the attitudes of users during the pandemic, where we separate the tweets into three categories according to Bayes classification: positive, negative, and neutral. We calculate the polarity of a tweet by dividing the sum of polarity of all the words by the total number of words in the sentence. Figure 2 displays the trends of three categories over a 12-month period, where each data point presents the topics within a single week.

From Figure 2, we can observe that a large portion of tweets fall into the negative category in most weeks from Feb 2020 – Feb 2021. Generally the negative posts take around 15% among all tweets during the pandemic. It takes a large portion compared to other normal topics that refer to the entertainment or fashion. Besides, the positive and neutral tweets take the most portion in the social media.

To learn more about the tweets during the period from February 2020 to February 2021, we study three key topics that refer to (1) economic during the pandemic, (2) political discussion related to the U.S presidential election, and (3) the procedure of mRNA vaccines. We analyze the three topics based on the weekly data from February 2020 to February 2021.

The tweet is identified to be vaccine relevant if it contains any words of the name of vaccine company, and key words such as "vaccine", 'vaccine distribution', "mutation", "injection", "mRNA", etc. The economy relevant tweet can be classified if it contains any words of 'job', 'nasdaq', 'economy', 'stock', 'S&P500', 'stimulus', and 'opening'. Tweets refer to politics can be identified if it contains 'election', 'political', 'campaign', 'elected', 'politics', and the name of candidates. Figure 3 shows the variances of topics during the 12 months. From Figure 3, we can observe that during the early time of the pandemic, such as from February to April 2020, the economy relevant tweets have higher volume that other. This is because the stock market crash suffered a sudden global stock market crash that began on 20 February 2020 and ended on 7 April [2]. During that period of time, there had multiple severe daily drops, such as "Black Monday II" of 12-13% in most global markets on March 16th, "Black Monday I" on 9 March, and "Black Thursday" on 12 March [2]. Therefore, large amounts of tweets correlated to COIVD-19 were posted during the period of time.

From June to early November 2020, political relevant tweets dominate the social world due to the significant activities refer to the presidential election in USA. There were huge volume of topics that had been raised during the election and most of these topics were related to the COVID-19 situation since USA had rapid increment of infected cases. After November 2020, vaccine related tweets become the largest part among all tweets. This is because the stock market had been recovered and the USA president election had finished, and normal

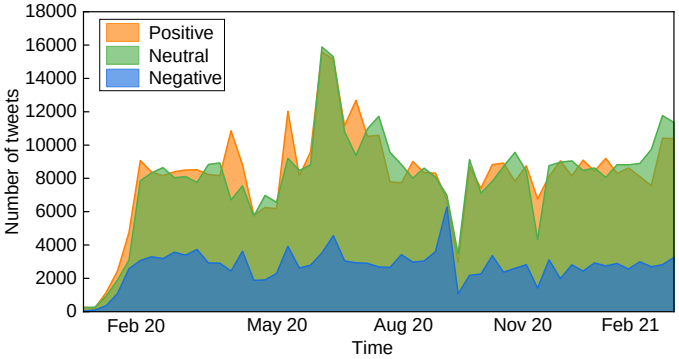

Fig. 3: Frequency of vaccine, economy, and politics related tweets across all the tweets.

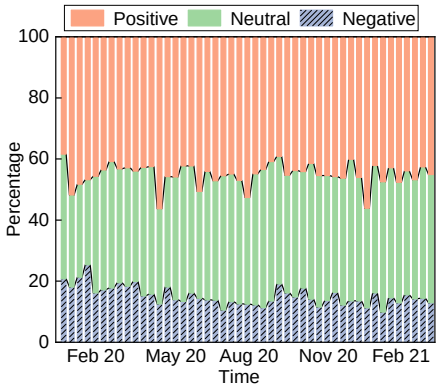

Fig. 4: Sentiment of vaccine-related tweets

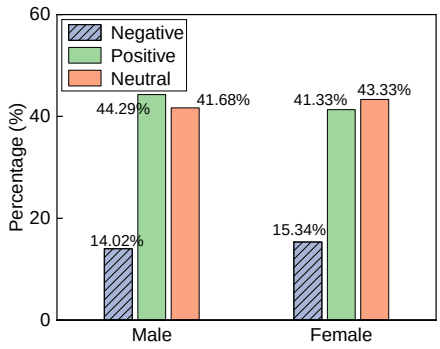

Fig. 5: Attitudes of gender groups.

people turned to focus on the updates of various vaccines that can help to stop the pandemic.

Next, we analyze the attitudes of users for tweets that relate to vaccine and genders. Figure 4 shows the distribution of negative, positive, and neutral tweets during the 12 months' period of time. We can observe that the negative tweets have the highest volume at the beginning of the pandemic and decline with the time. This is because users start to get used of the pandemic and the progress of vaccines reduces the negative emotions during the long period of time. Demographic data such as gender and age are not available from the Twitter API due to the restriction of its users' privacy rule, therefore we identify Twitter users' gender by using the gender classification based on tweet content. The gender classification was trained with available training dataset on Kaggle [1] by using the multinomial Naïve Bayes classifier [21] in Sklearn library. To better train the classifier, only data with 99% gender-confidence were chosen to include in our data set. Figure 5 shows the attitudes among different genders. Around 44.29% of male users express positive emotion during the pandemic, that is a bit higher than the neutral tweets that take around 41.68%. Besides, negative tweets in female users a little bit higher that male users and the portion of positive rate is lower that male users. It leaves more spaces that we can explore the deeper causes in the future study.

V. Conclusion

As the pandemic continues to threaten the entire world by affecting life in the real and virtual worlds, the "infodemic" continues to severely challenge the way people obtain and access the truth about the world. Many precious studies present various viewpoints of the analysis of social information, however, they mainly focus on a relatively short period of time. The long-term analysis of social information is necessary, especially during this special world health crisis that leads to a global social crisis. In this work, we conduct a long-term analysis of social posts that related to COVID-19 within 12 months. We analyze the collected social data in three categories: vaccines, politics, and economics and analyze the sentiment attitudes of tweets from various perspectives. Our analysis shows that negative tweets occupy an important position during the pandemic and decline with the development of vaccines. Besides, the evolution of various topics tightly follow the hot discussions on economics, politics, and vaccines.

In the future, we will analyze the tweets by targeting on deeper and more comprehensive perspectives. We will continue to explore the variances among different topics and show the roles that genders feeds back at different stages of the "infodemic". Besides, we will characterize sentiment patterns to extend the understanding of the impacts of online social "infodemic".

References

[1] Twitter user gender classification, https://www.kaggle.com/crowdflower/twitter-user-gender-classification
[2] 2020 stock market crash (2020), https://en.wikipedia.org/wiki/2020_stock_market_crash
[3] Natural language toolkit (2020), https://www.nltk.org
[4] Scikit learn api reference (2020), https://scikit-learn.org/stable/modules/classes.html
[5] Textblob: Simplified text processing (2020), https://textblob.readthedocs.io/en/dev/
[6] Un tackles 'infodemic' of misinformation and cybercrime in covid-19 crisis (March 2020), https://www.un.org/en/un-coronavirus-communications-team/un-tackling-infodemic-misinformation-and-cybercrime-covid-19
[7] Tweepy documentation (2021), https://docs.tweepy.org/en/latest/index.html
[8] Who coronavirus (covid-19) dashboard (March 2021), https://covid19.who.int
[9] Brennen, J.S., Simon, F., Howard, P.N., Nielsen, R.K.: Types, sources, and claims of covid-19 misinformation. Reuters Institute 7, 3–1 (2020)
[10] Dyer, J., Kolic, B.: Public risk perception and emotion on twitter during the covid-19 pandemic. Applied Network Science 5(1), 1–32 (2020)
[11] Kruspe, A., Häberle, M., Kuhn, I., Zhu, X.X.: Cross-language sentiment analysis of european twitter messages duringthe covid-19 pandemic. arXiv preprint arXiv:2008.12172 (2020)
[12] Kuchler, T., Russel, D., Stroebel, J.: The geographic spread of covid-19 correlates with the structure of social networks as measured by facebook. Tech. rep., National Bureau of Economic Research (2020)
[13] Lyu, H., Chen, L., Wang, Y., Luo, J.: Sense and sensibility: Characterizing social media users regarding the use of controversial terms for covid-19. IEEE Transactions on Big Data (2020)
[14] Pano, T., Kashef, R.: A complete vader-based sentiment analysis of bitcoin (btc) tweets during the era of covid-19. Big Data and Cognitive Computing 4(4), 33 (2020)
[15] Pennycook, G., McPhetres, J., Zhang, Y., Lu, J.G., Rand, D.G.: Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. Psychological science 31(7), 770–780 (2020)

[16] Shahsavari, S., Holur, P., Wang, T., Tangherlini, T.R., Roychowdhury, V.: Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. Journal of computational social science 3(2), 279–317 (2020)
[17] Shi, W., Liu, D., Yang, J., Zhang, J., Wen, S., Su, J.: Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter. International Journal of Environmental Research and Public Health 17(22), 8701 (2020)
[18] Xu, H., Guan, B., Liu, P., Escudero, W., Hu, L.: Harnessing the nature of spam in scalable online social spam detection. In: 2018 IEEE International Conference on Big Data (Big Data). pp. 3733–3736. IEEE (2018)
[19] Xu, H., Hu, L., Liu, P., Guan, B.: Exploiting the spam correlations in scalable online social spam detection. In: International Conference on Cloud Computing. pp. 146–160. Springer (2019)
[20] Xu, H., Hu, L., Liu, P., Xiao, Y., Wang, W., Dayal, J., Wang, Q., Tang, Y.: Oases: an online scalable spam detection system for social networks. In: 2018 IEEE 11th International Conference on Cloud Computing (CLOUD). pp. 98–105. IEEE (2018)
[21] Xu, S., Li, Y., Wang, Z.: Bayesian multinomial naïve bayes classifier to text classification. In: Advanced multimedia and ubiquitous engineering, pp. 347–352. Springer (2017)