# COVID-19: Analyzing Tweets for Extractive Multi-Document Summarization on News

Yu-Ting Lee[1]
*Dept. of Computer Science*
*National Chengchi University*
Taipei, Taiwan
ga11004@cs.nccu.edu.tw

Hong-Ren Chen[1]
*Dept. of Computer Science*
*National Chengchi University*
Taipei, Taiwan
ga11010@cs.nccu.edu.tw

***Abstract*—Social network enables people to connect and communicate through the Internet. The elements of social network are categorized by users, posts, and users whose posts are cited. According to research, most of the information sources from Tweets are derived from users' subjective opinions, and news reports. During the COVID-19 pandemic, opinions and news spread online in different forms. Information including panic/false information, conspiracies, physical conditions after vaccination, etc. In the aspect of users' opinions, they are recommended to the specific information preferred due to the feature from recommendation system. We develop the Summary Generation Tool for News during Important Time Duration (SGIT) takes into account the confidence towards pandemic from most users, and explores how news affects the public opinion. For news researchers, SGIT generates summary of news of interest within a specific time range. For users, SGIT provide more objective news content about pandemics and vaccination data to help users get rid of the subjective information suggested by recommendation system.**

***Index Terms*—COVID-19 pandemic, Social Network, Opinion Mining, Topic Modeling, Sentiment Analysis, Multi-Document Summarization**

## I. Introduction

Twitter is a social media platform provides users with publishing opinions in microblog-format. The time duration ranging from minutes to years, allows researchers to examine short-time/long-time duration events for observing cause and effect relationship. An event is mostly defined as happened at a specific time range. Long-time duration events are discussed by traditional media, and now recorded by electronic media. Thanks for the help for real-time update functions equipped by Twitter, people share not only personal opinions but also relevant rising events. In the aspect of personal opinions, more and more people receive information from online news media and posts published on social networks applications(e.g. Twitter, Facebook, Instagram, YouTube, etc.). Recommendation systems delivers preferred contents to users, which is common applied by social networks, and news application on end devices. However, the personalized recommendations only provides user preferred contents which might include false, subjective, and conspiracies. Thus, we develop a tool (Summary Generation Tool for News during Important Time Duration, SGIT) in purpose to assist users receive more objective news. Furthermore, SGIT also enables news researchers to examine how news affect public opinions on daily vaccinations and related news during COVID-19 pandemic.

### A. Proposal

Both facts and fake news flow on Twitter, especially when the fake news is recommended to users by recommendation system, users will share the false information in various forms of posts. However, for users who want to receive more objective information, it's hard to clarify the reachable social posts is facts or other people's subjective thoughts. To assist users and news researchers dive into how news affects the public confidence on COVID-19 pandemic and vaccine brands, we develop a tool (SGIT) to perform sentiment analysis and topic modeling for the collected tweets, trying to find out which keywords are discussed more when emotions fluctuate and the number of vaccinations vary greatly. Then, using multi-document summarization on news searched with the collected keywords analyzed through topic modeling within the variation duration.

### B. Results

After implementing sentiment analysis and topic modeling to collected tweets, performing multi-document summarization to news scrapped with variation time duration and related keywords, we have output summarization to interpret how news affect public confidence on COVID-19 pandemic. For example, we found that the emotional changes on November 3, 2021 fluctuates greatly from the collected tweets, then topic modeling was performed to the tweets within the time range, then collect relevant news using the keywords and time intervals obtained from sentiment analysis. Through the multi-document summarization for the relevant news, we are able to interpret the positive changes on sentiment was because the U.S. CDC allowed children between 5 and 11 years old to be vaccinated with low-dose Pfizer vaccine. SGIT provides more latent information of how news affects public opinions and public confidence towards COVID-19 pandemic.

## II. Related Work

## III. Method

In this section, we first discuss data collection, then give an overview of complete process.

### A. Data Collection

We started this data collection on 2021-11-01, leveraging Twitter's streaming application programming interface (API) and Tweepy to query certain keywords and hashtags for related tweets. Time duration of data collection coverages from 2021-11-01 to 2021-11-20 and the time range from 2021-11-30 to 2021-12-14 after the outbreak of variant COVID-19, named Omicron. In the experiments, we referred the keywords selection from a public coronavirus Twitter dataset (Emily et al., 2020). Around 33,168 and 12,692 English tweeets with users located in the Unite States were collected on two stages of duration respectively. The daily vaccination data refers to Our World in Data, recorded with dates, brands, and quantities. Since the doses of Johnson & Johnson only accounts for 3.6%, the discussed vaccines brand includes Moderna and Pfizer.

### B. Fine-tune the Pre-trained Language Model

Transformer-based language model have changed the prospects of natural language processing nowadays. Models such as BERT, ALBERT, RoBERTa all follow the same application principle - training the bidirectional transformer models with multiple huge unlabelled corpus. This process is completed with masking language models (MLM) and next sentence prediction (NSP) (J.Devlin et al., 2018). Then researchers are able to fine-tune the pre-trained models on downstream tasks. On the stage of fine-tuning language model for fitting with sentiment tasks, we fine-tuned BERT with tweets dataset labelled with sentiments.

### C. Sentiment Analysis

Sentiment analysis is a common technique in natural language processing used to identify emotions changes associated with written texts. Most used cases of sentiment analysis include monitoring users' feedback online, monitoring sentiments from stock market, etc. In this research, we applied sentiment analysis to monitor public confidence towards COVID-19 pandemic and different brands of vaccines.

### D. Topic Modeling

To take deeper dive in texts, sentiment analysis is not enough. In this part, we apply topic modeling to understand the tweets data, and those extracted topics help to analyze the contexts of tweets. This approach also further help researchers label documents and define topics for clusters of documents, making the analysis more interpretable. In common cases, topic modeling tasks uses Latent Dirichlet Allocation (LDA), a generative statistical model for unobserved variables via observed variables. In this research, we leverage BERTopic (Maarten Grootendorst, 2020) to produce dynamic topic modeling using BERT's embeddings on monitoring the possible discussed topics, and find the frequent mentioned keywords during the time of interest.

### E. Multi-Document Summarization on News

The goal of this process, we find out how news affects public confidence towards COVID-19 pandemic. Traditional news data are recorded on papers, and not real-time updated. Thanks to the advanced of streaming online news, a vast quantity of news is generated online. However, it's hard to digest and filter out important information by humans. To observe news happened which might affects public confidence towards COVID-19 pandemic, using neural networks for extractivec multi-document summarization on news is essential. Therefore, we apply fine-tuned Distil-BART for summarization tasks through Huggingface libraries API.

## IV. Experiments

In this section, we first formulate various types of problem we intend to solve. Then we give an overview of the solution and simulate the interpretation of relevant views.

### A. Problem Formulation

For users on social networks, recommendation system benefits human lives in most aspects. However, it's dangerous when happened during the pandemic. Since each human being holds different perspectives towards everything. In the result of our simulation, recommendation is more likely provides fake news if users prefer certain false information, especially during the COVID-19 pandemic. To earn with more objective information instead of preferred news, SGIT takes into the vaccinations facts and public opinions for further information outputs. With the extractions news which might affect public opinions, news researchers are able to detect the fake news flow in social networks, reporters are also benefited during the process on collecting eye-catching information.

### B. Overview of SGIT

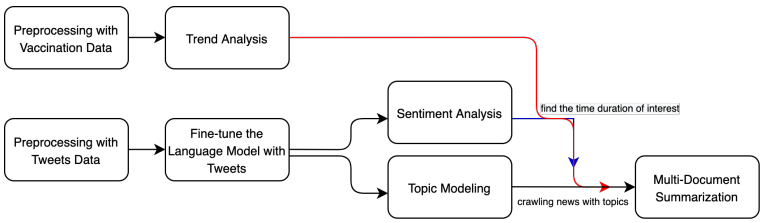


Fig. 1. Overview of SGIT, tweets inputs and experiments flow

Inputs and Outputs: With raw tweets inputs, SGIT provides text cleaning, then analysis contexts with fine-tuned language model. To capture the time of interest, we calculate the daily change rate of sentiments and vaccinations. The dates varies higher than average changes are included in time of interests. By leveraging topic modeling to monitor most discussed topic in the duration, summaries of related news are generated with topics and time inputs.

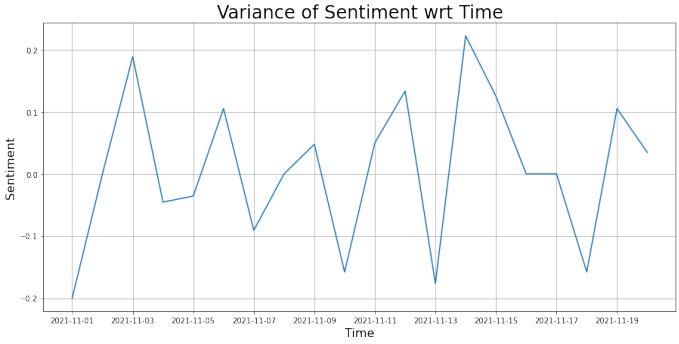


Fig. 2. Sentiments Analysis during the first time stage in the USA

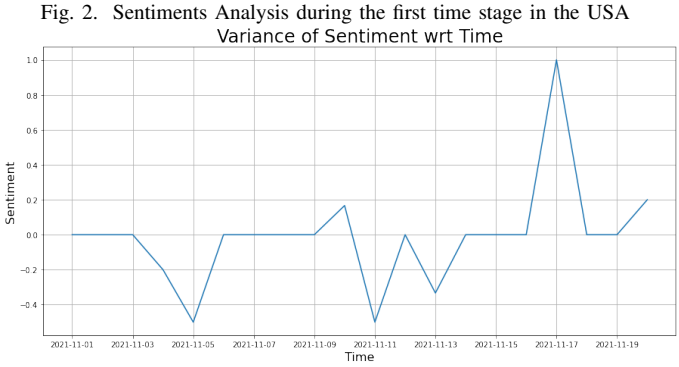


Fig. 3. Sentiments Analysis towards Pfizer vaccines during the first time stage in the USA

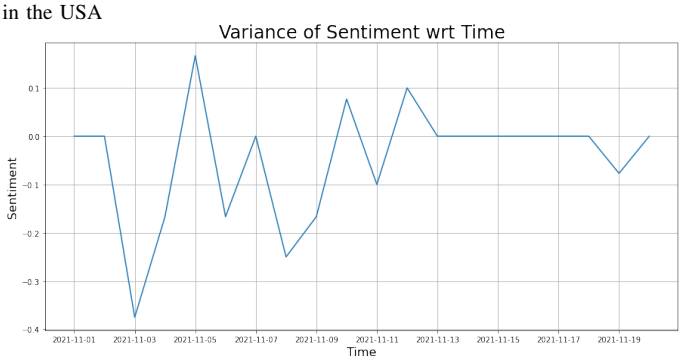


Fig. 4. Sentiments Analysis towards Moderna vaccines during the first time stage (Nov.)

### C. Visualization on Sentiment Analysis

In this subsection, we first discuss the visualization of sentiment analysis in the United States from two stages of duration. Next, we extract further more information leveraging topic modeling and multi-document summarization. We take the data of the first stage as examples:
Figure.2, Figure.3, and Figure.4 show the results of sentiment analysis according to U.S.A region and specific vaccine brands respectively. Researchers are able to choose interested time duration, then leverage SGIT to understand more information about events happened.

### D. Topic Modeling

After applying sentiment analysis to collected tweets, we use topic modeling on finding the most discussed topics during

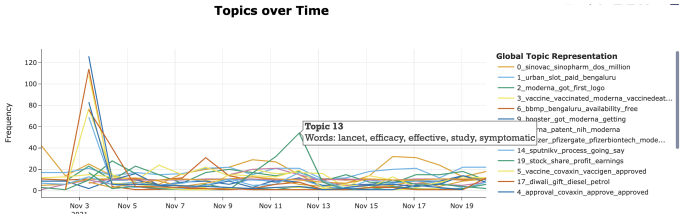


Fig. 5. Topics discussed during the first time stage in the USA

the specific time duration. In Figure.5, discussed keywords clustered into topics, represents daily frequent mentioned contents. The result also further indicates the trends of topics, in assist to help researchers observe latent aspects of public opinions. Keywords extracted by topic modeling algorithms BERTopic are further used to query related news. Thus, researchers are able to earn more information about the interpretation of sentiments and how news affects public opinions.

### E. Find the Time Duration of Interest

Find the time duration of interest is important because it's complex to analyze tweets posts in a long time duration. Therefore, we take into the probabilities of daily variance on sentiments analysis and daily vaccinations in Figure.6 and Figure.7. In SGIT, we calculate the probability of daily changes, then filter those dates with more variance changes than average changes, including sentiment analysis and vaccinations data. According to the two brands of vaccines (Moderna & Pfizer), we choose the time with the variance duration during the intersection of both vaccines.

### F. Multi-Document Summarization

In previous subsections, we have the extracted keywords, time duration of interest, and the result of sentiment analysis, now we are able to leverage the materials to obtain more information. Figure.8 shows the summarization process in SGIT, inputs including time duration and related keywords, output with the summarization of queried news titles. For demonstration, we only use the titles, then make summarization of returned news titles by saving researchers time on reading multiple news sources. In future work, the feature of making summarization utilizing contents of news will be added, but pre-trained models might consume more time on summarizing texts.

## V. Results

We present SGIT, an analytic tool including sentiment analysis, trend analysis on vaccinations data, topic modeling and multi-document summarization on news. In the aspect of news researchers, SGIT provides the information of what people discussed these days. In the aspects of normal users, SGIT is able to give more objective contents rather than the preferred contents from recommendation system. Empirically, we show that our approach leads to improvements on giving users more objective information during the COVID-19 pandemic. In future work, by filtering the websites which might include

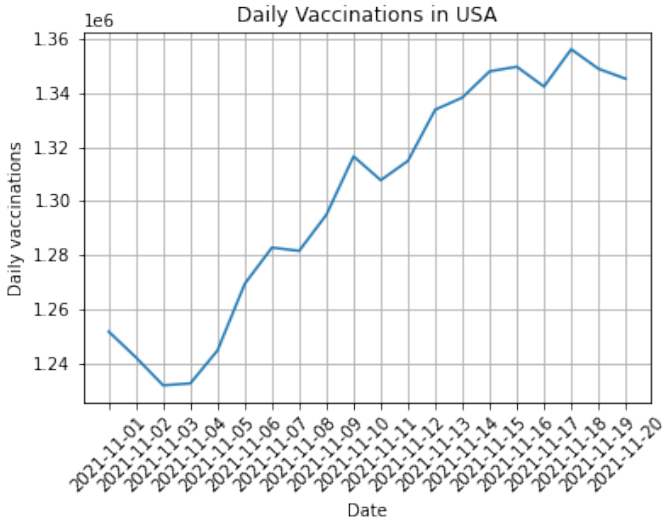


Fig. 6. The variance probability of daily vaccinations in the U.S.A

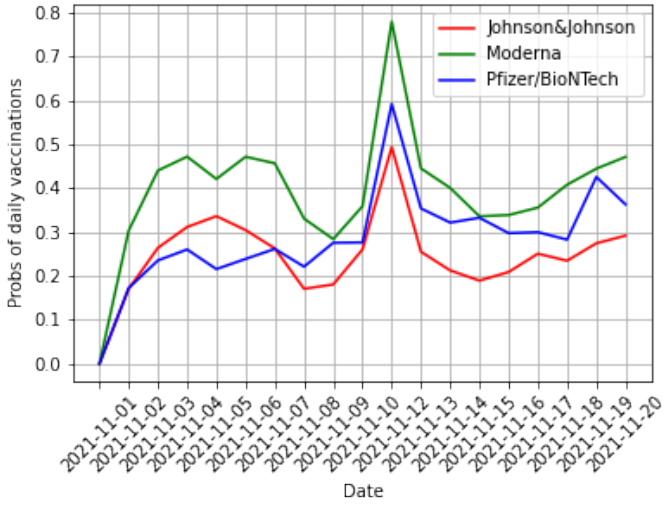


Fig. 7. The variance probability for daily vaccinations of different brands in the U.S.A

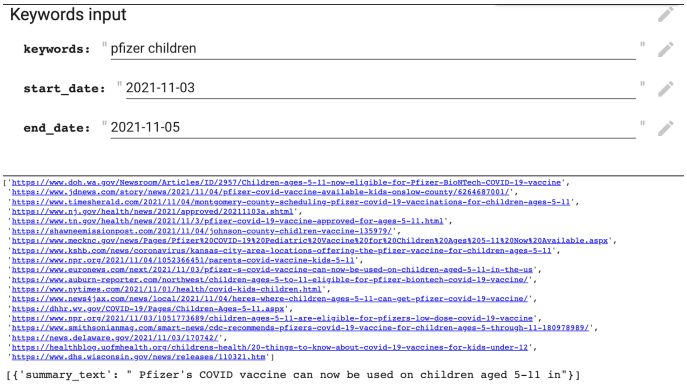


Fig. 8. Overview of SGIT's summarization process

fake news in summarization process is essential to provides factual information towards users.

## REFERENCES


[1] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding "*arXiv:1810.04805*

[2] M. Müller, M. Salathé and P. E. Kummervold, "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter"*arXiv:2005.07503*

[3] M. Munikar, S. Shakya and A. Shrestha, "Fine-grained Sentiment Classification using BERT"*arXiv:1910.13461*

[4] A. Glazkova, M. Glazkov and T. Trifonov, "Exploiting CT-BERT and Ensembling Learning for COVID-19 Fake News

[5] K. Sharma, S. Seo, C. Meng, S. Rambhatla and Y. Liu, "COVID-19 on Social Media: Analyzing Misinformation in Twitter Conversations "*arXiv:2003.12309*

[6] V. Mazzeo, A. Rapisarda and G. Giuffrida, "Detection of fake news on CoViD-19 on Web Search Engines", doi:10.3389 Detection"*arXiv:2012.11967*

[7] B. Chen, B. Chen, D. Gao, Q. Chen, C. Huo, X. Meng, et al., "Transformer-based Language Model Fine-tuning Methods for COVID-19 Fake News Detection"*arXiv:2101.05509*

[8] E. Chen, K. Lerman and E. Ferrara, "Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set"*arXiv:2003.07372*

[9] R. Vijjali, P. Potluri, S. Kumar and S. Teki, "Two Stage Transformer Model for COVID-19 Fake News Detection and Fact Checking"*arXiv:2011.13253*

[10] M. Fernández, A. Bellogín and I. Cantador, "Analysing the Effect of Recommendation Algorithms on the Amplification of Misinformation"*arXiv:2103.14748*

[11] V. Mazzeo, A. Rapisarda and G. Giuffrida, "Detection of fake news on CoViD-19 on Web Search Engines"*arXiv:2103.11804*

[12] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al., "HuggingFace's Transformers: State-of-the-art Natural Language Processing"