Scalable Social Media Analytics*

Big-Data Analytics Technology

Jianwei Luo[†]
School of Computing
National University of Singapore
Singapore
e0997986@u.nus.edu

Rui Xue[†]
School of Computing
National University of Singapore
Singapore
e0998037@u.nus.edu

Ziyu Zhou[†]
School of Computing
National University of Singapore
Singapore
da.zhou38@gmail.com

Leyi Du[†]
School of Computing
National University of Singapore
Singapore
duleyi64@gmail.com

Wang Guangyu[†]
School of Computing
National University of Singapore
Singapore
Guangyu@u.nus.edu

Abstract

Social media platforms such as Facebook and Twitter generate huge amounts of data every day, leading to plenty of interesting phenomenons and analysis. Our group will be focusing on the specific topic – echo chamber for this study. In this study, we analyzed the echo chamber on social media using a dataset from tweets related to Covid-19. Our analysis focused on three approaches: frequently used words with strong opinion, sentiment analysis, and polarity over time with our hypothesis to address the phenomenon of echo chamber. As the sentiment label is missing in our datasets, we also applied a DistilBERT model, a compressed version of Bert for NLP, to predict sentiment labels for analysis purposes in the study.

Introduction

Background

Echo chambers represent situations in which individuals primarily encounter information and opinions that coincide with their existing beliefs, often within online environments such as social media platforms, forums, or blogs. Scholarly research on echo chambers suggests that potential effects may encompass confirmation bias, polarization, diminished critical thinking, groupthink, among others. The rapid spread

information on social media platforms has created an environment where echo chambers can thrive.

Literature Review

In the pioneer study "Echo Chambers on Social Media: A comparative analysis", made by M. Cinelli, G. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, the echo chamber phenomenon was thoroughly examined through a literature review. This examination utilized two fundamental components: (i) the division of opinions regarding a controversial topic, and (ii) the tendency of similar individuals to interact with one another. These concepts were translated into measurable observables, which allowed for a comparative analysis of echo chambers on various social media platforms. The researchers studied more than 1 billion pieces of content created by 1 million users across four platforms (Facebook, Twitter, Reddit, and Gab), and constructed their interaction networks using different features such as shared links, followed pages, follower relationships, and commented posts. They assessed the presence of echo chambers using two main criteria: the tendency towards homogeneity in interaction networks, and the inclination towards biased information diffusion among like-minded peers.

Hypothesis

In our study, to analyze echo chambers, we did a case study on COVID19, using the two tweets data related to COVID19 from kaggle.

We proposed three hypothesis related to echo chambers, and will validate or reject it during the study:

(1) Opinions regarding covid19 are polarized.

(2) As people continue to engage in echo chambers on social media, their viewpoints tend to become increasingly polarized over time

To demonstrate our hypothesis, we will use different techniques, including word count, TFIDF, NLP model, etc.

Keywords

• Scalable

• Social media

• Analytics

• Sentiment

• Big data

Data

Dataset1: We get this dataset from kaggle. (https://www.kaggle.com/datasets/bansodesandeep/covid19-tweets-data?select=Corona_NLP_train.csv)This dataset1 includes around 40,000 tweets data with sentiment labels from October, 2022 to April, 2022, such that we can use it directly for analysis, and also use it for training NLP. This dataset contains 6 columns, including 'UserName', 'ScreenName', 'Location', 'TweetAt', 'OriginalTweet' and ' Sentiment'. In addition, this dataset will be used on our sentiment labels distribution analysis.

Dataset2: We get this dataset from kaggle as well (https://www.kaggle.com/code/purvasingh/covid19-tweets-eda-and-sentiment-analysis/input?select=covid19_tweets.csv). It's a database of around 100,000 tweets data without the sentiment label from Jul 25, 2020 to Aug 11, 2020. his dataset contains 13 columns, but we will only take 10 of them, including 'user_name', 'user_location', 'user_description', 'user_description', 'user_created', 'user_followers', 'user_friends', 'user_favourites', 'user_verified', 'date', and 'text'. Those data will be used in our word cloud part.

To clean the datasets, mainly to clean the original text column, we did some preprocessing to clean the text:

(1) used stopword remover to remove stop words
(2) remove punctuations existing in the text
(3) remove all empty lines exiting in the text to have a more condense text
(4) convert all text to lowercase

Methodology

(1) Word Count and TF-IDF:

In this study, we used mapreduce as an approach to get frequently used words in the dataset. The objective is to find the most relevant or important words in the corpus.

The baseline approach is to get word count. For each line of the tweet, we first tokenize the sentences into words, and get the count of each word. Afterwards, we get the top 500 words among, and manually filter those words related to subjective opinions, emotions, judgements, and etc. For instance, though "COVID19" and "virus" have high rank, it's removed after manual filtering as it's a word of fact.

Simply counting the frequency of a word (TF) can lead to misleading results, as some words may appear frequently in a document without necessarily being important or meaningful in the overall context of the corpus. Therefore, we also calculated the normalized TF-IDF (Term Frequency-Inverse Document Frequency) of each word, which is a commonly used technique to measure the importance or relevance of a word in a document or a corpus.

TF-IDF (Term Frequency-Inverse Document Frequency) for a term t in a document d:

$$TF - IDF(t, d) = tf(t, d) * idf(t)$$

Normalized TF-IDF for a term t in a document d:

$$Normalized\ TF - IDF(t, d) = (tf(t, d) / maxtf(d)) * idf(t)$$

After getting the normalized TF-IDF, we used mapreduce to get the sum of the normalized TF-IDF. The sum of normalized TF-IDF scores across documents has a similar interpretation to the sum of TF-IDF scores. However, the normalization step ensures that the scores are on a consistent scale and allows for more direct comparisons between documents. Specifically, normalized TF-IDF scores represent the relative importance of a term within a document compared to its importance across all documents. Therefore, the sum of normalized TF-IDF scores across documents represents the collective relative importance of a term across all documents. The higher the sum, the more important the term is considered to be in the overall corpus. Afterwards, we also manually filter those words we are interested in.

(2) NLP model with sentiment labels

In order to predict the sentiment labels for the Covid-19 dataset of dataset2 which lacks sentiment labels, it is necessary to train a model on a dataset with labeled sentiment information. In this study, the dataset1 containing sentiment labels was used to train a pre-trained DistilBERT model for Natural Language Processing (NLP) tasks. DistilBERT, a smaller and faster version of the Bert model, was utilized for data preprocessing and feature engineering.

To fine-tune the DistilBERT model and make final predictions, a Neural Network was employed. The model consists of two parts: the first part uses the pre-trained DistilBERT model to process the input data, while the second part involves the Neural Network carrying out the prediction task. The pre-trained DistilBERT model with a base architecture and uncased vocabulary was loaded for the first part. For the second part, a Neural Network with three layers was trained. The first layer is a linear layer that reduces the dimensionality of the output from the DistilBERT model. The second layer is a dropout layer that randomly drops out a portion of the input units during training to prevent overfitting. Finally, the third layer is a linear layer that maps the output of the previous layer to a single scalar output.

Result and interpretation

(1) Word Cloud

The word cloud is conducted over the dataset2. To view the top relevant words in our corpus, we utilized word clouds to visualize the result.

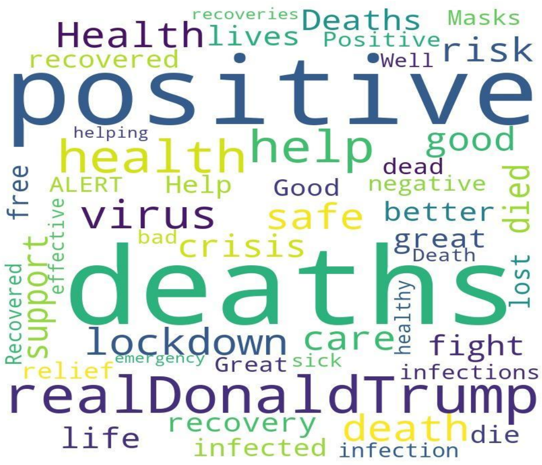

Figure 1: word cloud1 based on word count

Figure1 shows the first word cloud. Those top words are selected based on word count. From the word cloud, we can see that most words convey strong emotions and opinions, with some positive cases: health, good, better, and some negative cases: deaths, lockdown, fight, etc. Those words can be a representation of polarity.

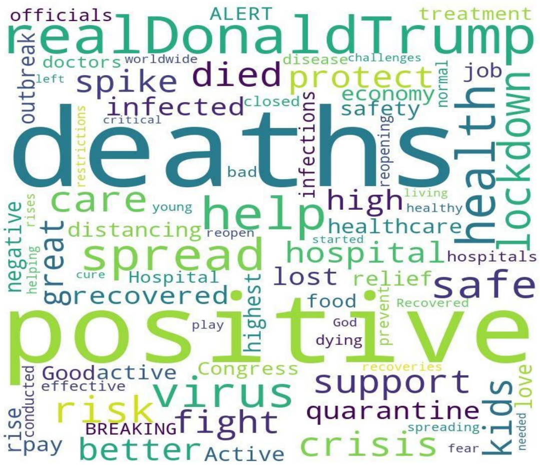

Figure 2: word cloud2 based on normalized TF-IDF

Figure 2 shows the word cloud, selected based on the sum of their normalized TF-IDF scores across all documents. This normalized score is a measure of the importance of the word within that document, and can eliminate some possible bias compared to the result we get from word count. From this figure, we can see most words convey strong emotions and opinions, similar to the word cloud1.

Overall, these two-word clouds provide similar results, in which we can infer opinion polarity in our corpus.

(2) Overall Sentiment Label Distribution

Using sentiment labels, we get the bar plot of count to visualize the sentiment polarity distribution. We then compare the result we get from dataset1 and dataset2. In this study, we use both 5 sentiment labels (neutral, positive, negative, extreme positive, extreme negative) and 3 sentiment labels (neutral, positive, negative) for analysis.

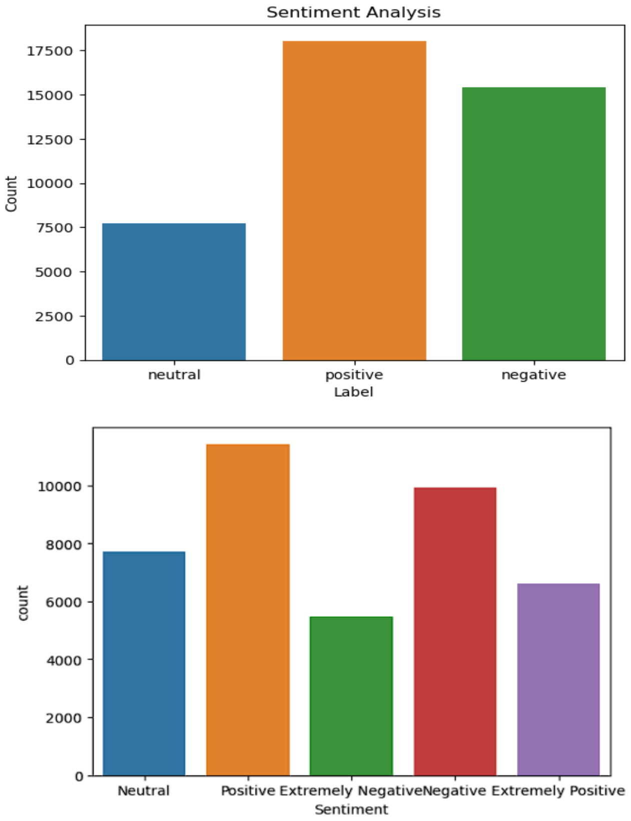

Figure 3: Sentiment Polarity distribution with 5 labels and 3 labels respectively on dataset1

Figure 3 shows the sentiment polarity distribution on dataset1. The two count plot shows the distribution of each sentiment label in the document. It's very obvious to get insight from the plot: as there are around 7500 neutral labels, 17,500 positive labels, and 15,000 negative labels, the most percentage of sentiment labels is positive, then negative, and least neutral. The highly imbalanced distribution of labels can be viewed as a representation of polarized opinion.

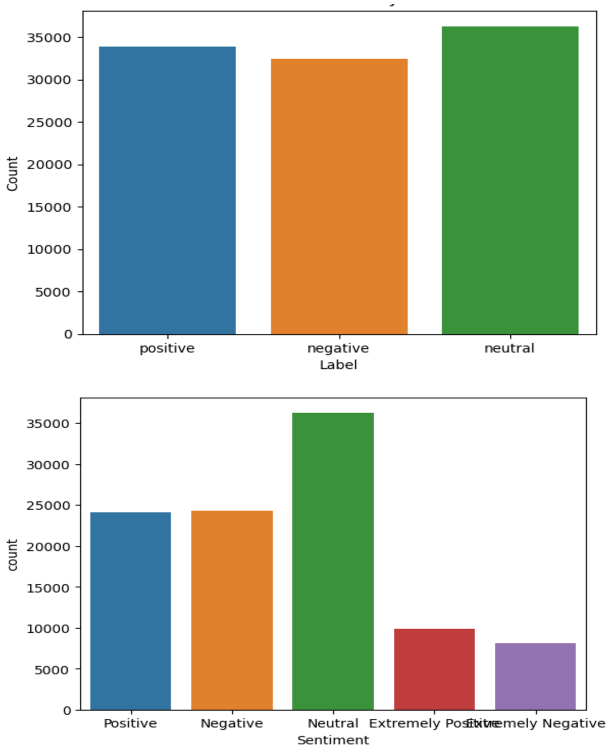

Figure 4: Sentiment Polarity distribution with 3 labels and 5 labels respectively on dataset2

Figure 4 shows the result after our model has been run which predicted sentiment for the future months to analyze the trend and polarity of the sentiment changes. It seems tweets were changed to neutral sentiment as it has the largest component in the charts below. Negatives and positives are kind of closer to each other. Extreme cases are even lower than before. Neutral sentiment increased and negative sentiment decreased as compared to the Figure 3 chart. The difference could be caused by the real polarity change, or by the NLP model prediction mislabelling.

(3) Sentiment Label Distribution Over Time

As stated in the hypothesis, we are interested in how the opinion polarity changes over time, as people engage in echo chambers. Therefore, we analyzed the sentiment label with percentage over time. As we are interested in the polarity distribution, we are more interested in the percentage of each label instead of the count of each label over time.

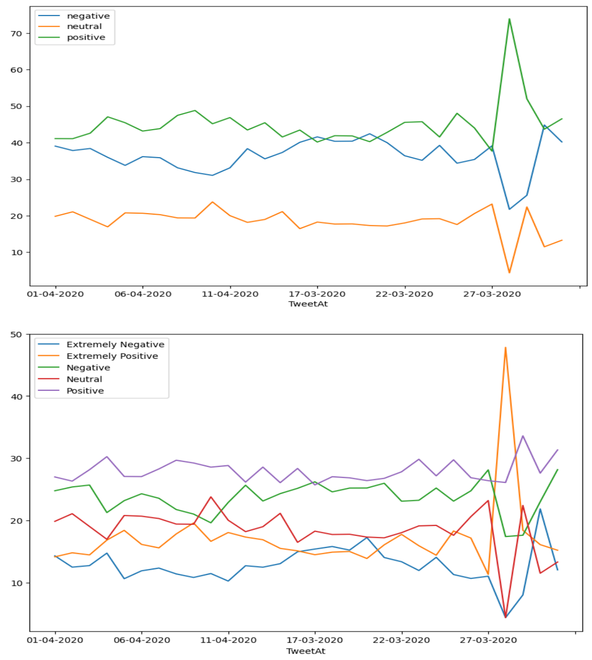

Figure 5: Sentiment Polarity Trend with 3 labels and 5 labels respectively on dataset1

Figure 5 shows the trend chart of the sentiment polarity changes over time. X coordinate represents the timeline, and Y coordinate represents the percentage of count, for each sentiment in each day. As we can see from the graph, the three labels largely leveled off over time, with a very noticeable boost in positive labels around March 27, 2020, followed by a fall back. Meanwhile, neutral labels and negative labels have a very significant drop followed by a rebound. When we take a step closer and subdivide the labels from three to five, we can see at the time point of March 27, 2020, extremely positive labels had a maximum boost followed by a fall back. positive labels had a nice boost but not as big as extremely positive labels and then fell back, while neutral The neutral labels and negative labels, extremely negative labels have a very significant drop followed by a rebound, where neutral labels have the largest drop.

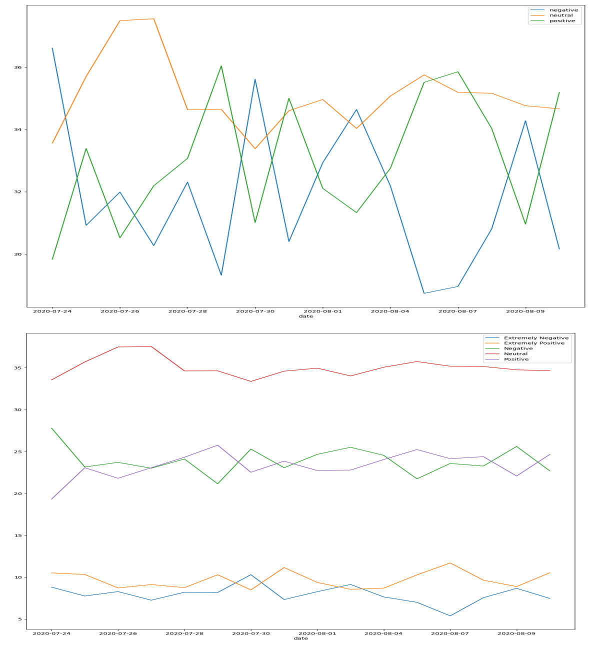

Figure 6: Sentiment Polarity Trend with 3 labels and 5 labels respectively on dataset2

Figure 6 is a similar plot as figure 5, showing the sentiment polarity trend on the dataset2. From the figure, we can see that the change of the three kinds of labels fluctuates very much with time. We can see an obvious decline of negative between August 4, 2020 and August 7, 2020, while the downward path of positive is also very obvious between August 1, 2020 and August 4, 2020, and neutral tends to be more stable compared to the other two. However, we did not see a general rule of how polarity changes over time, which related to our hypothesis 2.

Conclusion

In conclusion, our study aimed to investigate the phenomenon of echo chambers in the context of COVID19 by analyzing tweets related to the pandemic. We proposed three hypotheses related to the polarization of opinions in echo chambers, and used various techniques such as word count, TFIDF, and NLP models to validate these hypotheses.

Our findings show that using dataset1 (the one from kaggle), which contains sentiment labels, echo chambers do exist and our three hypotheses are validated. However, when using dataset2, where we obtain sentiment labels using an NLP model, we were only able to validate hypothesis 1, while we found no evidence to support hypotheses 2 and 3.

Our study suggests that echo chambers can thrive in the context of COVID19, but further research is needed to better understand this phenomenon and its implications.

Limitation

The first limitation in our study is the limited size and diversity of our dataset. In this study, we included only one case of covid19 as the controversial topic. More data can be included and analyzed for a deeper and more reliable conclusion.

The second limitation is on the NLP model, from which we get the sentiment label. As the sentiment label is not the direct feature we get, the analysis using it might be inaccurate and impair our study.

Another limitation is the difficulty to interpret the information of the tweets. Emotions and opinions can often be expressed through the use of emojis in text-based communication, such as in social media posts, messages, or comments. However, most text analysis techniques focus solely on analyzing the textual content of a message, and may not be able to pick up on the emotional or opinionated meaning conveyed by emojis.

For example, if someone wrote the sentence "I had a great time at the party 😊", a text analysis tool might be able to identify the sentiment of the message as positive, but it may not recognize the significance of the 😊 emoji in conveying the speaker's happiness and positive emotions.

Thus, when analyzing text data, it's important to keep in mind that emotions and opinions may be conveyed through non-textual elements such as emojis, and that text analysis techniques may not be able to fully capture this information.

Reference

[1] M. Cinelli, G. Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "Echo Chambers on Social Media: A comparative analysis," in Proceedings of the 2020 Conference on ResearchGate , doi: arXiv:2004.09603v1. Available:https://www.researchgate.net/publication/340826673_Echo_Chambers_on_Social_Media_A_comparative_analysis/citation/download

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, Oct. 2019. Available: https://arxiv.org/abs/1910.01108.

Data sources:

https://www.kaggle.com/code/purvasingh/covid19-tweets-eda-and-sentiment-analysis/input?select=covid19_tweets.csv

https://www.kaggle.com/datasets/bansodesandeep/covid19-tweets-data?select=Corona_NLP_train.csv