Sentiment analysis of Twitter comments related to COVID-19

Group Member

Dai Dong, Sijia Han, Frank Shi

Abstract

Twitter has one of the most significant microblogging social media in the world. It has 330 million registered users up to the first quarter of 2019 (Tankovska, 2021). For Twitter users, it has become a norm to use tweets and comments limited by the number of characters to express their lives and emotions. During the COVID-19 pandemic that started in December 2019, a significant number of related comments are spread every day through different social media and Twitter. With the trend of the epidemic, analyzing the public sentiment data in the comments can also help understand the public's emotional changes during the epidemic, helping understand the public's perception of the epidemic. In our project, We plan to make a sentiment analysis of tweets based on the monthly comparison. We use the python programming language to extract tweets about the coronavirus from August 2020 to the end of March 2021 from the coronavirus-related tweets database. We use Naive Bayes' algorithm and a database of known positive and negative sentiment tweets to train the classifier to distinguish between positive tweets and negative tweets to a certain extent. When completing sentiment analysis, our classifier accuracy is about 67%.

1.Introduction

The rise of the person-to-person contact administrations(Social Network Service, SNS) has drastically changed the way for individuals to connect and communicate to their friends and the public. SNS paved the way for individual communication and contributed to implementing enormous amounts of data to do comparative research. With five hundred million users on Twitter, it has become a popular communication tool and a data research tool. The tweet provides millions of data for the sentiment analysis of goods, companies, reviews and news.

During the last year, the coronavirus outbreak was first founded in December 2019 in Wuhan City, Hubei Province, China. In March 2020, the virus will be officially labelled COVID-19. The patient will experience fever, cough and shortness of breath within 2 to 14 days of exposure. There were 132 million confirmed cases of Covid-19 reported globally as of April of 2021, and 2.87 million deaths were reported [3]. In acknowledgment of Covid-19's widespread global dissemination, the World Health Organization on March 11, 2020, declared Covid-19 a pandemic.

The change of the communication "The number of "monetisable" daily users on the platform increased by 34% to 186 million in the second quarter compared to a year earlier - the fastest growth since Twitter first started using the metric in 2016".

In this research, all data is collected from Tweeter's database, preprocessed by the Panacea Lab. The Panacea lab provides the COVID-19 related tweets, which is extremely helpful for the experiment. Python was chosen for the programming language to perform sentiment analysis due to python's conveniences. Python provides functions that could access these tweets from Twitter's developer API and tweepy library. In summary, part two contains the relative work, part three has the research detail, and the result has been provided in part four.

2.Related Work

2.1. Sentiment Analysis

There are two types of sentiment classification strategies in the Sentiment Analysis, machine learning approach and lexicon-based approach, which both rely on the bag-of-word. The machine learning approach's classifiers are using the unigrams or N-gram as features to determine the polarity of the text. The lexicon-based approach assigned a polarity for each Unigram and produced the overall polarity by adding each Unigram together.

2.2. Machine learning approach

A Machine Learning Approach for text classification is a supervised algorithm that analyzes data labelled as positive, negative or neutral. Extracts features that model the differences between the various classes and construes a function that can be used for classifying new examples.[ ] The entire process of machine learning and text classification is: Data preprocessing, feature generation, feature selection, learning the algorithm and model evaluation. Before the sentiment classification, it is necessary to do the data model smoothing. The first step is to incorporate the Parts-of-Speech tagging(POS), The words in the text are divided into classes according to their function, and these classes are called Part-of-Speech classes or POS classes for short. The POS classes are also sometimes called syntactic categories, grammatical categories, or lexical categories.[ ] Tokenization is text processing in which the plain text is broken into words. It may not be a simple process, depending on the type of text and the kind of tokens we want to recognize. Stemming is the type of word processing in which a word is mapped into its stem, which is a part of the word representing the primary meaning of the word. It is used in Information Retrieval due to the property that if two words have the same stem, they are typically semantically very related. Hence, if their stems replace words in documents and queries, the resulting indices are smaller, and words in a query can be easily matched with their morphological variations. Lemmatization is a word processing method in which a surface word form, i.e., the word form as it appears in text, is mapped to its lemma, i.e., the canonical form as it appears in a dictionary. For example, the word working would be mapped into the verb work, or the word semantically would be mapped to the lemma semantics.

2.3. Machine Learning algorithms

The most well-known machine learning algorithms for text classification are Naive Bayes, Support Vector Machines (SVMs), Maximum Entropy (MaxEnt) and Decision Trees [ ]. The Naive Bayes classifier is the fundamental foundation of this research. We assume that all variables are independent except one distinguished variable, usually called the class variable, since the model is used for classification. The other variables are called features or attributes. Since the features are used as input and the class variable produces the classification result or output in the classification task, we also call the feature variables the input variables, the class variable, and the output variable.[ ]

$$\arg\max_{x_1} \frac{\mathrm{P}(V_2|V_1) \cdot \mathrm{P}(V_3|V_1) \cdot \ldots \cdot \mathrm{P}(V_n|V_1) \cdot \mathrm{P}(V_1)}{\mathrm{P}(V_2, V_3, \ldots, V_n)} =$$

$$\arg\max_{x_1} \mathrm{P}(V_2|V_1) \cdot \mathrm{P}(V_3|V_1) \cdot \ldots \cdot \mathrm{P}(V_n|V_1) \cdot \mathrm{P}(V_1)$$

Figure.1

2.4 Possible research

One of the possible approaches is rules-based sentiment analysis. It is a simple model with a bunch of preset words labelled positive or negative. Then the model would use the word lists to fit in the data and calculate the total of positive, negative and neutral. Then it would determine the sentiment of the sentence (L.I. Tan 2015).

Apply a sentiment dictionary. Since some finished sentiment dictionaries, it would be a better approach than the previous one since we do not need to define the current word list ourselves. These words have already been given a better score for sentiment analysis (Reagan 2017).

Combine the sentiment dictionary with machine learning. Since the sentiment dictionaries may not perfectly fit with all circumstances, we might need to reset the score of different words in the dictionary. To shoulder the load, LSTM-RNN based ML model could be a good solution (Agrawal, 2020).

In 2017, the International Workshop on Semantic Evaluation opened several tasks on semantic evaluation for participants to work out. Task 4 included one subtask that "Given a tweet, decide whether it expresses the POSITIVE, NEGATIVE or NEUTRAL sentiment." We planned to use and improve the model with the best f1-score in the competition, provided with open code. It is an LSTMs and CNNs model (Cliche, 2017).

3. Problem Definition and Processes

In sentiment analysis, the accuracy of the analysis model is the essential part of the entire project. We searched many Twitter sentiment analysis models, and their highest accuracy is 0.685 [1]. Therefore, we hope to find a way to improve the accuracy of the model.

3.1. Data Collections

The Twitter API platform provides extensive access to public Twitter data that users choose to share with the world. Using the Twitter API application, we can extract real-time tweets from the Twitter database. Panacea Lab is a publicly accessible resource and contains the Covid-19 Twitter chatter dataset for scientific use. This dataset includes a dataset of COVID-19-related tweets obtained by Panacea Lab from Twitter. And due to the gradual increase in people's awareness of the virus and the spread of the virus, starting from March 11, 2020, the data set can collect more than 4 million tweets per day.

The Covid-19 Twitter chatter dataset will collect tweets in all languages in this data set. The more popular ones are English, Spanish and French. Due to language barriers, we only analyzed the tweets in the English language in the collection. The data set also only provides the tweet ID of each tweet related to the coronavirus. When we need to use tweet data, we first need to match each tweet ID stored in the data set with the actual tweet data through the Twitter API to read the corresponding tweet data. This process is also called Hydration.

On the other hand, there is a Project-level Tweet cap limiting the number of Tweets you can retrieve from several Twitter API v2 endpoints. It is set to 500,000 Tweets per month for Standard Projects at the Basic access level. We can't obtain all the data from the Covid-19 Twitter chatter dataset due to Twitter API Basic access level application restrictions. So we chose to filter the tweets extracted from the coronavirus-related tweet data set. First of all, because the data for the first five months, from March 2020 to July 2020, is multilingual mixed data, we skipped the data extraction for these months to reduce the workload. Secondly, the number of daily tweets collected in the Covid-19 Twitter chatter dataset can reach more than 4 million, which means that we will have an enormous data size from August 2020 to the end of March 2021.

Based on the consideration of these factors, we finally chose to randomly sample 100 tweets from the daily data set in the Covid-19 Twitter chatter dataset to represent the tweet data of the day. The 100 tweets extracted from daily tweets in the current month are integrated to represent the tweet data of this month. At the same time, we repeated this progress five times to delete five monthly random data as our test data to ensure that we have enough test data and training data while ensuring the diversity of the data.

3.2 Data Preprocessing

Data preprocessing is an indispensable step in machine learning because the quality of the data and the valuable information derived from the data will directly affect our model's

learning ability. Therefore, it is imperative to preprocess the data before inputting the data into the model.

Since we plan to analyze the text content of the tweet itself, we do not include other tweet attributes such as favourite count, retweet, creation date. Therefore, after completing the tweets data collection, we first need to extract the pure text content from each tweet. The second process is to clean the data. Although we have collected data from the contents of tweets, there is still a bunch of useless information in the data, which is not helpful to the sentiment analysis we plan to implement. Useless information in the content of the tweet includes URL pattern, mentions pattern, emoticons pattern, digital numbers, smileys pattern and punctuation pattern. We use regular expression operations to remove the useless elements listed above in each text content data.

After removing the noise in the data set, we still need to tokenize and lemmatize the data for subsequent processing. To this end, we use the TweetTokenizer() function and the WordNetLemmatizer() function in the nltk package to tokenize and lemmatize the data.

3.3 Dictionary

For better model training and to improve the accuracy, we apply a known open sentiment dictionary containing a dataset of positive words, which has a total number of 2006words, and a dataset of negative words, which has a total number of 4783 words [2].

3.4 Naive Bayes Algorithm

As we introduced before, we use Naive Bayes as a classifier to distinguish whether each tweet is positive or negative. We hope that by putting the processed tweets into the classifier, the classifier can analyze accurate results, positive or negative. Before making the classifier reach what we expect, we first need to train the Naive Bayes classifier. In the previous section, we obtained a sentiment dictionary consisting of positive and negative lists. We use this dictionary as the training data to train our Naive Bayes classifier model. After completing the classifier's training procedure, we have an excellent result for the test data set. So，we put the five datasets that were randomly extracted and done processed before into the classifier. We put each dataset into the classifier and then calculate the average overall accuracy to reduce possible errors.

4. Result

The figure below shows the entire process of the experiment, including the preprocessing stage-extract tweets and training the naive Bayes classifier. The processing stage of classifying the tweets then gives results by the month.

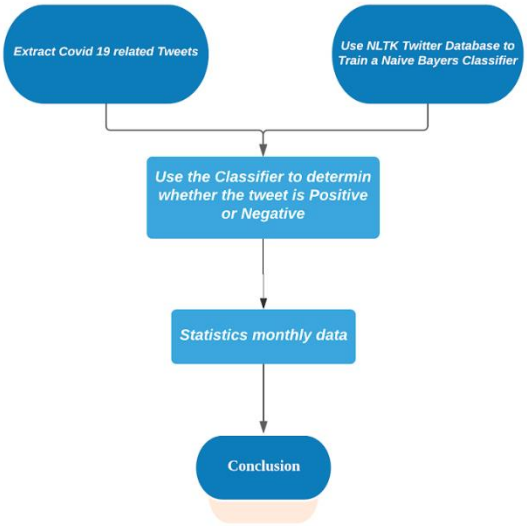

Figure.1 Process chart

During the training classifier stage, we have compared the accuracy for the Unigram, bigram and trigram. Due to the decrease in the accuracy, we drop the bigram, trigram and hybrid for the classifier.

Unigram	**0.7133**
Bigram	**0.4923**
Trigram	**0.5073**
Hybird	**0.7133**

Figure 2. The accuracy of the classifier

After choosing the best performing classifier, apply it to the analysis of the data after the preprocessing. The result shows that tweets' positive rate is significantly higher than the negative rate, which is in line with forecasts from related work done by other research papers. The detailed polarity rate sees figure 3 and visualized in figure 4.

	Positive	Negative
2020-08	0.6755	0.3245
2020-09	0.6738	0.3262
2020-10	0.6787	0.3213
2020-11	0.6625	0.3375
2020-12	0.6548	0.3452
2021-01	0.6495	0.3505
2021-02	0.6685	0.3315
2021-03	0.6804	0.3196

Figure 3. The result of positive and negative rates.

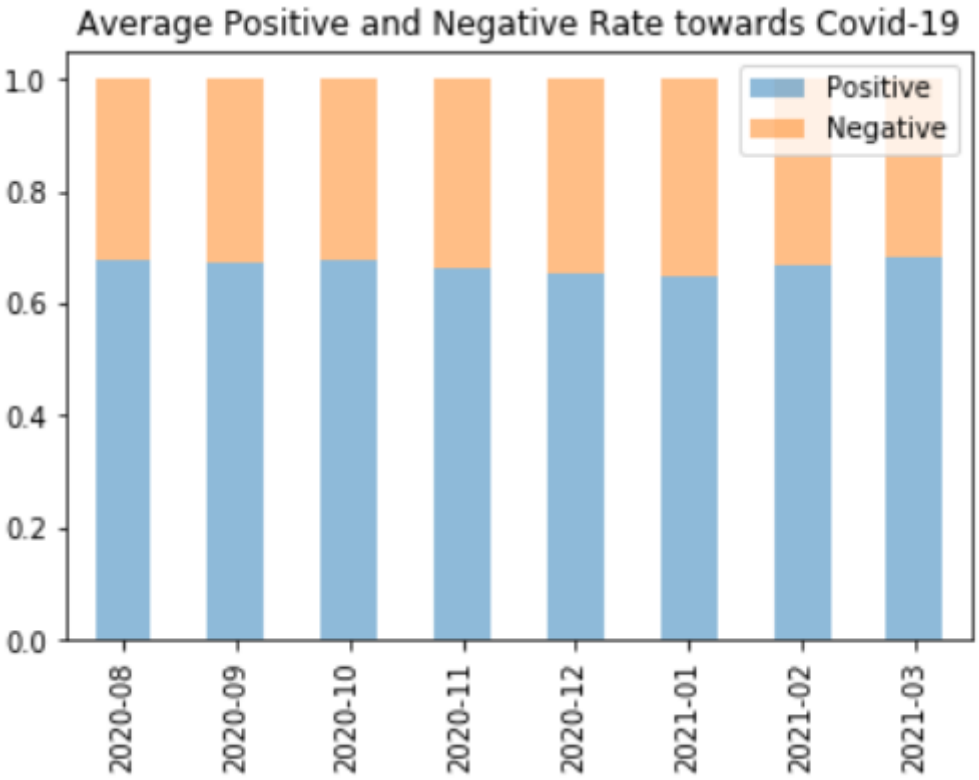

Figure 4. Visualize the polarity rates

The most common words also are extract and visualize in figure 5

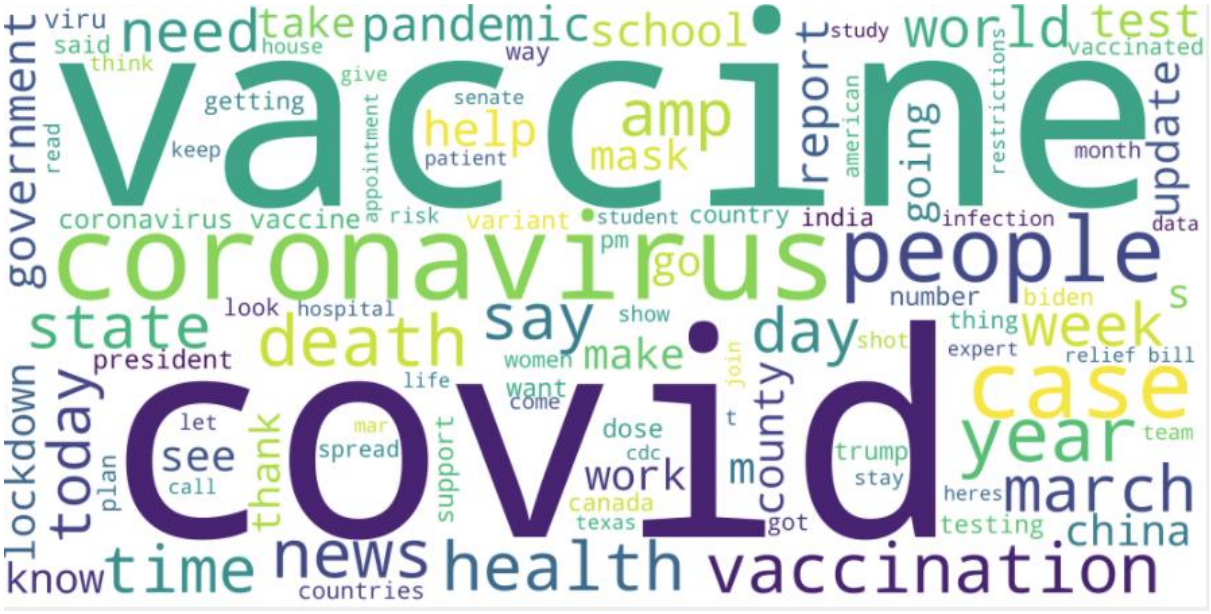

Figure 5. Word cloud

5. Conclusions

This study performed sentiment analysis on Twitter data related to the coronavirus. The research used a public sentiment dictionary and a new database of coronavirus-related tweets to help. The data uses Twitter API and python's tweepy library and collected tweet content data between August 2020 and March 2021 based on the data set provided by Panacea Lab. The sentiment dictionary contains about 6,800 English positive and negative words. Finally, the naive Bayes algorithm is used to obtain the accuracy of judging positive and negative. According to the results, using Unigram will provide a higher accuracy rate for the classifier. Likewise, overall about 67% of tweets are positive. Because the collected data missed the beginning of the new coronavirus, the most widespread period of the virus, the fluctuations are not obvious. However, comparing our project with known similar projects, the accuracy did not have any improvement. There are several possible reasons. There are typos in some tweets that cannot be recognized by the classifier, resulting in lower accuracy. Since only positive and negative words are identified in the project, neutral words may become biased. In the future, we plan to establish a comparison between sentiment analysis of tweets and the number of people infected with the new coronavirus and determine whether there is an association. We use Naive Bayes in the project. But there may be better ways to improve accuracy. We will try other methods to enhance the project and improve accuracy.

Reference
Not finished yet

[1] Leelaylay, “SemEval-2017 Task 4: Sentiment Analysis in Twitter,” *GitHub*. [Online]. Available: https://github.com/leelaylay/TweetSemEval. [Accessed: 06-Apr-2021].

[2] Leelaylay, “leelaylay/TweetSemEval,” *GitHub*. [Online]. Available: https://github.com/leelaylay/TweetSemEval/tree/master/dataset/dict. [Accessed: 07-Apr-2021].

[3] “Coronavirus (COVID-19),” *Google News*. [Online]. Available: https://news.google.com/covid19/map?hl=en-US&mid=%2Fm%2F015jr&gl=US&ceid=US%3Aen. [Accessed: 07-Apr-2021].