Journal of Electronics and Information Science (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2022.070102
ISSN 2371-9524 Vol. 7 Num. 1

Guide Social Media to Play the Correct Direction Based on Text Classification

Zhiyuan He[a,*], Yaqiao Li[b]

School of Informatics, The University of Edinburgh, Edinburgh, United Kingdom

[a]979705195@qq.com, [b]704878999@qq.com

**Corresponding Author*

Keywords: COVID-19, Text classifiers, Social media, Deep learning

Abstract : Due to the global spread of COVID-19, it has caused many negative comments on the Internet. However, there is very little research on the spread and changes of negative speech during the epidemic, and very little is known about it. In this work, we use twitter as the source of social media voice, and use twitter api to collect related tweet data. We use existing manually labeled data sets, train different text classifiers and compare their performance, and use the best-performing text classifier to classify our collected tweets as negative and positive. Through these data analysis, we found that from the beginning of the pandemic to the outbreak, there are more negative tweets and greater impact than positive tweets. We need to understand these changing trends, so as to guide the online media to play a correct role in guiding society.

1. Introduction

Due to the outbreak of COVID-19, information full of panic and worry on social networks began to increase, which may increase or amplify the harm and impact of COVID-19 [1]. If we grasp the development trend of this information and the harm it causes, we can re-make social media a positive energy dissemination tool.

This paper aims to investigate the potential harm and impact of positive and negative tweets on American social media platforms, and analyze how these related tweets have changed since the COVID-19 explosion. To facilitate our research, we will use the Twitter API to collect relevant tweets, and use the existing relevant annotated COVID-NEGATIVE data set to train the classifier to help classify the collected tweets. Tweet tagging problems can be classified as text classification problems [5]. Deep learning models have outstanding performance in text classification tasks due to their multi-layer networks and non-linear characteristics[6], but an important disadvantage is that a large amount of data is required to train the network. However, there are currently very few data sets related to the subject of our research, although there are lots of annotated datasets for negative tweets on social media, they are not in the context of COVID-19 pandemic. Related research mostly collects relevant tweets, and then manually label these tweets to train the classifier[7, 8]. In the past few decades, the problem of text classification has been extensively studied and solved in many practical applications[9]. The research and solution of text classification problems are mainly

Divided into these stages: feature extraction, dimensionality reduction, classifier selection and evaluation. Among them, as the main research objectives, feature extraction and classifier selection will be introduced in this chapter. We chose BERT [16], random forest [12], logistic regression [5] and SVM [10] as the classifier model and compare their performance of on the specific tweets dataset[8] we used. Moreover, we will use the traditional TF-IDF algorithm [21] and the deep learning model BERT[16, 22] for feature extraction[19, 20], and compare their extraction performance. Finally, we also briefly discussed the possible model bias in the data and the data set bias between the COVID-NEGATIVE dataset we used and the collected tweets.

2. Methods

The main objective of this paper is to use the COVID-NEGATIVE data and Self-Annotated data to train and test an excellent classifier to classify the collected Tweets data. Then, we would research and analyze the shift in different types of tweets during the COVID-19 outbreak.

In this process, two different Text Feature Extraction methods, BERT and TF-IDF, would be compared for their performance and influence on tweets classification in this project. Classic and popular ML algorithms: Support Vector Machine, Random Forest and Logistic Regression would be used as baseline classifiers to classify processed text data. Moreover, the BERT model would be adopted and compared with three other classic ML classification models. Due to the imbalance of the data in the COVID-NEGATIVE data set, it may lead to the problem of model bias. Therefore, We randomly selected 200 collected data and manually labeled the labels, and used these data as the second test set to test the performance for the classifiers trained on COVID-NEGATIVE dataset. By testing different feature extraction methods combined with different models, we would select the “best classifier” to classify the collected tweets.

2.1 Data Preparation

2.1.1 Data Collection and Exploration

To collect related tweets in the specific location: the US, the query parameter were set as: keywords+”place country:US lang:en”. keywords means the keywords combi- nation we used below, place country located tweets’ location in the US, where “lang” were used to specified collected language was English. The Twitter API stipulates that the maximum number of tweets searched per month is 10 million, the number of tweets related to COVID-19 is huge. So, we made a keyword-search approach strategy. Three keywords sets were set up, to make the collected tweets are related with topics COVID-19 and negative informations. Based on keyword-search approach, we collected 16076 tweets from January 23, 2020 to May 31, 2020 in the US.

To give a more detailed perspective on our collected data, some data exploration approaches were adopted. Hashtags can be a powerful indicator of the final category of tweets, such as tweets with “stopAapiNegative” or “StopAsianNegative” hashtags are more likely to be of the “CounterNegative” in general. However, Hashtags can also be misleading, like some of negativeful tweets will use a lot of hashtags (including anti-negative hashtags like “AAPI”) in order to make them hot. Another useful extraction method is Wordcloud[23].

We applied data cleanning to our data and get the Wordcloud in Figure 1. From our observation, most large words are tend to be “negativeful”. Some words are meaningless and neutral, like “know” ,”one” et al. However, “counternegative” words were hard to find, only “aapi” in the lower right corner could be found. We extracted all hashtags from raw collected tweets data. Figure 2 presents the top 25 frequent hashtags. Similar with Wordcloud, most hashtags are supposed to be “Asian-related” and only “AAPI” is likely to be “Counternegative”.

Fig.1 : Tweets Wordcloud

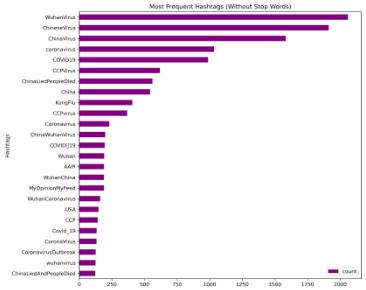

Fig.2 : Most Frequent Hashtags

Based on the distribution of hashtags in tweets, we can find that most tweets are likely to be “Asian-related”, while only a small number of tweets are “counternegative”. Further analysis is based on our Asian-related list and Counternegative list in the keyword list and our understanding. Our top 20 hashtags are going to be negativeful, Counternegativeful, and Neutral. Only “AAPI” were “Counternegativeful”,and the left were classified as “neutral”. From the counting result, our tweets dataset contained 7990 “Negativeful” Hashtags,4373 “neutral” Hashtags and 190 “Counternegativeful” Hashtags.

2.1.2 Annotated COVID-NEGATIVE Dataset

This proejct used existing manually labeled COVID-NEGATIVE dataset[8] to train Tweets classification model to help us classify the collected tweets. This dataset contains 2,319 tweets data, labeled as 678 “Negative” tweets, 359 “Counter Negative” tweets , 961 “Neutral” tweets and 321 “Non-Asian Negative” tweets. For the research purpose of this project and the research population, we ignored and discarded the tweets marked as “Non-Asian Negative”. Moreover, in order to verify the performance of our model in our collected tweets, we manually labelled some tweets collected through the Twitter API in the follow-up. Based on the original data set, to label our collected tweets, we had made some modifications to the labeling principles and definitions of these labels based on the basis of original definition in [8]:

Negative: Tweets contain negative information such as abuse, derogation, discrimina- tion, and rumors directed at Asians or Asian organizations, China, the Chinese govern- ment, Wuhan and other Asian-related groups or individuals.

Counter Negative: Oppose and criticize racial discrimination, insults, violence and other acts against Asian groups or individuals, or encourage and support Asian indi- viduals or groups against racial discrimination.

Neutral: There is no explicit hatred or anti-negative attitude towards Asians, but it contains content related to COVID-19 and Asian-related, most of which are news or pure spam tweets.

Through these data, we would verify the performance of the classifier(trained on the COVID-NEGATIVE dataset) on the tweets data set we have collected. But collected raw tweets data were

unstructured, it contained redundant information and noise. This data was not suitable for text classification and fed to models. In order to make the algorithm perform better, the tweets needed be transformed into a more digestible , a series of data cleaning steps would be adopted. Although data cleaning can reduce the noise in the data and improve model's accuracy in most cases. However, it may be a double-edged sword[27], as useful context may be deleted in the process of data cleaning. The actions of deleting useless words or condensing words into stems may be counterproductive. For the BERT model, which uses attention mechanism that learns contextual relations between words (or sub-words) in a text, preserving the originality of the data may be a better strategy. Therefore, in the follow- up, we will test whether not using pre-processing can actually improve the performance of the model.

2.2 Implementation

2.2.1 Feature Extraction

The performance of Feature Extraction Methods is mainly reflected in the classification score of the classifier. Classification accuracy, F1 score, accuracy and recall rate will be used to evaluate their performance. The same data input, partitioning, and model parameters will be set to ensure that comparisons are not affected by other factors. COVID-NEGATIVE will be divided into 80% train data and 20%test data, the random seed were set as 99 to ensure the same data set partition.

TF-IDF Feature Extraction :TF-IDF calculates the different weights of words in the text to form a number vector. We convert the COVID-NEGATIVE text data into the corresponding number vector, and obtained the 1998*8369 vector. When there is a large amount of text data, it is easy for TF-IDF to generate a large corpus, which will generate feature vectors with large dimensions, this may increase the probability of overfitting of the model[28]. Intuitively, regarding documents number, 8,369 features are really large dimen- sions. Moreover, after observing features, we found that TF-IDF generated a lot of features that are meaningless, such as the pure number "1000","10000", and the very long and strange word "zhangyixingsingle". These words are obviously rarely used, and perhaps only one document uses these complex and awkward words, which may lead to model overfitting problems. The disadvantage of bag- of-words is that some words with very low document frequency are also included, and these meaningless words(features) may lead to overfitting of the model. In order to solve it, a dimensionality reduction techniques will be applied.

BERT Feature Extraction :For deep learning feature extraction, it was implemented by pre-trained BERT mode. bert-base-uncased, bert-base-cased and bert-large-uncased model were three pre-trained models were to be used in this section. In most cases, bert-base-uncased performs better than bert-base-cased. However, this is reversed when the case information of the text is important. In Twitter, people may capitalize certain words and add punctuation to emphasize or highlight certain words. And there's reason to believe that capital letters don't have to be completely meaningless for text messages on Twitter. Therefore, we assume that BERT cased model may have better performance compared with BERT uncased model. The other model we chose was BERT Large versus BERT Base. The main difference between them is the structure of the model. BERT Large has more Encoder layers, which means more parameters and more attention heads. However, we intuitively think that deeper network structure may have the ability to extract more useful features as its nonlinear character and deeper structure. The 1998 tweets in COVID-NEGATIVE dataset were to be fed into BERT. Normal text data cannot be recognized and used by BERT. Each tweet sentence were split into tokens and added the special token [CLS] in the beginning and [SEP] at the end of each sentence. Tokenized tweets will be converted into corresponding BERT token IDs. As ID tokens are used for token-based authentication to cache user

profile information and provide it to client applications, thereby providing better performance and experience. In addition, all input sequences should be a constant length. Therefore, the input of tweets of different lengths will be modified to the same fixed length by padding or truncation. The default max length of Huggingface is 512. We found through analysis of all tweets, the average sentence length of the tweets is only 50.28 tokens. When max length is longer than most sentence lengths, this will only increase the training time and waste resources, so we set max length to 64. Then we created attention masks, which is used to clearly distinguish the real tokens and those filled in padding[PAD] tokens, the actual tokens were replaced by 1 and padding tokens were replaced by 0. The input IDs and attention masks for the tweets text data were to be fed to pre-trained BERT to extract embeddings, and finally we got the feature vectors extracted by the pre-trained BERT model.

2.2.2 Models

The Classic Machine Learning Models SVM, RandomForest and Logistic Regression were implemented to build the models. To find the best hyperparameters combination for models, grid search and 5-fold cross-validation methods were applied in our experiments.

The parameter tuning strategies of different models are different for different data sets.As mentioned earlier, we processed our input data into different forms.Therefore, there is no absolute uniform parameter setting that can be applied to these different processed data. It is worth mentioning that our purpose is to select the classifier with the best performance and compare the performance difference between the traditional ML models and the BERT model. For the poorly performing combina- tion of dataset and models, we will not over-adjust the model parameters to achieve better but not the best results, we decided to default to use uniform model parame- ter settings for the poorly performing combination of the data and model to reduce meaningless work. The hyperparameters of several models are as follows:

Logistic Regression : “solver” using “newton-cg” and regularization strength hyperparameter “C” were set as 2.

Random Forest : max depth=20, n estimators=105, criterion=“entropy”.

Support Vector Machine : ’C’: 2, ’gamma’: 0.01, ’kernel’: ’linear’.

BERT : Batch size=16 epochs=4 learning rate= 2e-5.

Evaluation metrics play an important role in evaluating classification performance and guiding classifier modeling. We used f1-score and accuracy as our evaluation metrics in our experiments. Considering that the imbalance of our data set may lead to model bias problem in the trained model, we not only evaluated the overall classification performance of the model, but also evaluated the classification perfor- mance of the model in each category.

3. Evaluation

3.1 TF-IDF Dimension Reduction

In above, we mentioned that the shortcoming of TF-IDF is that it is easy to generate large-dimensional feature vectors when processing a large amount of text data, which increases the probability of model overfitting.We will use two ways of dimension reduction, LSA and dimension reduction by word frequency, to process the original TF- IDF feature data. In principle, it is possible to have more components than samples, but redundant com- ponents would be useless noise.

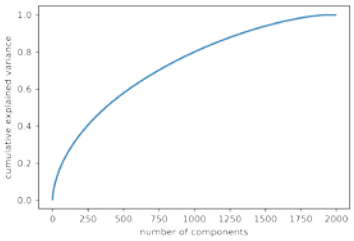

Fig.3 (Left): Data Information Preserved after Applying Lsa. Figure

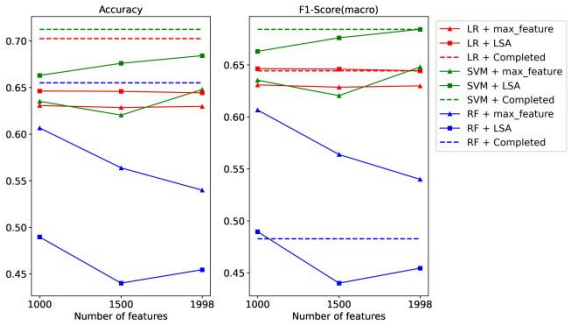

Fig.4 (Right): Classification Performance Applying (LSA)

Latent Semantic Analysis dimen- sion reduction, word frequency dimensionality reduction(max feature) and no dimen- sionlity reduction(Completed). Used models: Logistic Regression(LR),Support Vector Machine(SVM) and Random Forest(RF). Left: Classification performance on Accuracy evaluation. Right: Performance on F1-Score evaluation

We used LSA to reduce the dimension of 1998x8369 TF-IDF raw data to obtain 1998x1998 vector data. Figure 3.quantified how much of the total variance of the 1998 dimension is con- tained in the first N components. Seen from the figure, we can observed that the first 1500 components contain about 90% of the variance, which means that about 90% of the information is stored in the first 1500 dimensional components. We reduced the dimension of TF-IDF raw feature data to 1998, 1500 and 1000 dimensions, and com- pared their performance.

When we used Accuracy as the evaluation metrics, all di- mensionality reduction methods reduced the classification performance of the models. When we used f1-score (macro) as the evaluation metrics. Interestingly, we found that for models LR and SVM, the performance of the model was the same when using LSA to reduce the complete data from the 8369 dimension to the 1998 dimension. This is consistent with the 1998 components in Figure 4.that contain 100% of the variance, that is, 100% of the information is saved after the dimensionality is reduced to the 1998 dimension, which makes the completed data(8369 dimension) has the same clas sification performance with data after LSA dimensionality reduction(1998 dimension). Generally, the use of LSA will not enhance the classification performance of the model. For the model Random Forest, when f1- score is used as evaluation metrics, the overall classification performance of the clas- sifier is enhanced by reducing the dimensionality of the word frequency(max feature).

3.2 Different Pre-Trained Bert in Text Feature Extraction

We put forward several conjectures and assumptions. First of all, bert-base-uncased has better performance in most cases than bert-base-cased, except for cases where case information is important. The case of text in Twitter text should not be simply meaningless. Based on this analysis, our first hypothesis was that bert-base-cased as a text fea- ture extraction method would

perform better than bert-base-uncased. Deeper network with more complex structures may be able to dig deeper and more useful features from the text, although this is also accompanied by the risk of overfitting, Another hypoth- esis was that the more complex structure of the pre-trained model, bert-large-uncased, may have better performance than bert-base-uncased. We used the COVID-NEGATIVE dataset, feeding the above 3 pre-trained bert models to generate feature embeddings, and used SVM, Logistic Regression and Random Forest for classification training in our experiments.

Pre-trained BERT&Feature Extraction	Model	COVID-NEGATIVE Data		
		Accuracy(Test data)	F1-score(macro)	Accuracy(Train data)
bert-base-uncased	Logistic Regression	0.7850	0.7580	0.8980
	SVM	0.7750	0.7445	0.9074
	Random Forest	0.7225	0.6633	1.0
bert-base-cased	Logistic Regression	0.7600	0.7244	0.8686
	SVM	0.7250	0.6626	0.7372
	Random Forest	0.7075	0.6504	1.0
bert-large-uncased	Logistic Regression	0.7750	0.7473	0.9712
	SVM	0.7425	0.7108	0.9881
	Random Forest	0.6950	0.6508	1.0

Table 1 : Classification Performance Applying Different Types of Pre-Trained Bert Mod- Els. the Highest Scores on Each Model Are in Bold

Table 1 shows the classification performance on the test set when we use dif- ferent pre-trained BERT models to do feature extraction and combined with different models. The performance of bert-base-uncased model is obviously better than the other two pre-trained models, as bert-base-uncased model has the highest classifica- tion scores for any algorithm model. Contrary to our expectation, bert-large-uncased performance does not match our expectation, which indicates that case information in the COVID-NEGATIVE tweet data set is not important enough for classification. Another result contrary to our hypothesis is that the performance of Bert-Large is similar to but slightly worse than that of Bert-base. After observing their accuracy performance on train data, we found that Bert-Large had an obvious tendency of overfitting, and its train accuracy is almost 100%. The more complex network architecture does not give the model a performance boost. We believed that the model does extract deeper and more features, but this situation is not applicable to the case of a small amount of data. Feature extraction for a small amount of data in the deep network structure is likely to increase the risk of overfitting.

3.3 Classifiers Comparison

Model	Feature	Data	COVID-NEGATIVE		Manually Annotated	
			Accur	F1-	Accur	F1-
Logistic	TF-IDF	No	0.702	0.6443	0.730	0.5316
SVM	TF-IDF	No	0.712	0.6842	0.655	0.5346
Random Forest	TF-IDF	No	0.665	0.4988	0.725	0.4522
Logistic	BERT	No	0.785	0.7580	0.600	0.5246
SVM	BERT	No	0.775	0.7445	0.585	0.5322
Random Forest	BERT	No	0.722	0.6633	0.615	0.4445
BERT		No	0.723	0.6952	0.770	0.5494
Logistic	TF-IDF	YES	0.710	0.6573	0.730	0.5264
SVM	TF-IDF	YES	0.722	0.6886	0.660	0.5113
Random Forest	TF-IDF	YES	0.650	0.4916	0.725	0.4410
Logistic	BERT	YES	0.767	0.7310	0.625	0.4952
SVM	BERT	YES	0.782	0.7436	0.610	0.5070
Random Forest	BERT	YES	0.722	0.6450	0.610	0.5070
BERT		YES	0.703	0.6691	0.780	0.5405

Table 2 : Classification Results on COVID-NEGATIVE Dataset and Manually Annotated Dataset for Combinations of Different Models, Feature Extraction Methods and Data Cleaning

Our text feature extraction part adopted non-dimension reduction TF-IDF and the pre-trained BERT, bert-base- uncased. We divided the results into whether processing data cleaning and compared their performance. However, there is the bias between data sets. To test the influence of datasets bias problem to our classification, we randomly sampled 200 samples from the collected tweet data, manually annotated these samples according to the labeled principles and definitions. These data were used as another test set to test the categor- ical performance of the models trained on COVID-19 on the tweet data we collected. The classification performance are shown in Table 2.In Manually Annotated Dataset, fine-tuned BERT without data cleaning had the highest F1-score, and fine-tuned BERT with data cleaning had the highest accuracy. SVM + BERT text feature extraction had the highest F1-score and Accuracy in the COVID-NEGATIVE dateset, but it has ordinary performance in the mannually annotated dataset. On the contrary, fine-tuned BERT performed mediocre in the COVID-NEGATIVE dataset, but it performed very well in the self-annotated dataset. The main difference in these results was the difference in the data sets used, this may indicated that the dataset bias resulted in these results. Our original hypothesis was that data cleaning may delete some useful text information, while BERT, who depends on context information, may be affected deeply. However, out of our expectation, the result showed that data cleaning has little impact on classifier (slightly increasing or decreasing classification scores).

On the whole, the F1-score of all classification models in manually annotated dataset is lower than those in COVID-NEGATIVE, the main reason for this phenomenon is due to the counternegative data in the Manually Annotated dataset. There are only 10 “Counternegative” data. Any wrong judgment will cause a big change of f1-score in the Counternegative class. Since we use “macro” F1-score, The change of F1-score of Coun- ternegative also has a considerable impact on the overall F1-score. Another reason is on the datasets bias. In our manual tagging process, some tweets implicitly express the “Counternegative” perspective were also labelled as “Counternegative”, even though they don’t have explicitly related “counternegative” keywords. However, the large proportion of trained classi- fiers classified them as ’Neutral’. When we looked at tweets labeled “Counternegative” in COVID-19 NEGATIVE, we found that most of those

tweets had strong positions and opinions to express “Counternegative”. With this finding, we paid attention to the “Counternegative” tweets with strong expression in the Manually Annotated Dataset, interestingly, most of the classifiers did an accurate classification of this kind of tweets. Because of the bias in training data and the bias between two datasets, this double bias results in the poor performance in the Manually Annotated dataset.

3.4 Analysis on Social Media

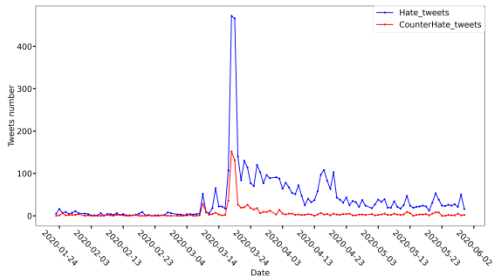

Fig.5 (Left): Tweets Number Shifts for ’Negative’ and ’Counternegative’ Tweets from January 23, 2020 to May 31, 2020 in the Us.

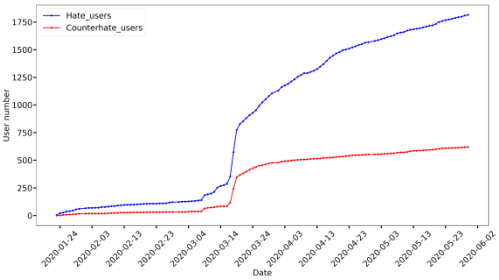

Fig.6 (Right): User Accumulation Number from January 23, 2020 to May 31, 2020 in the Us

Based on the results of Figure 5. , we selected the fine-tuned BERT without data clean- ing as our classifier to classfy the 16076 tweets we collected. 5,098 (31.7%) tweets were classified as Negative, 803(5.0%) were Counternegative, and 10,175 (63.3%) were Neu- tral. Figure 4.3 is the line chart present every day’s ‘Negative’ and ‘Counternegative’ tweets number. Our most intuitive finding is that Asian-related tweets are more frequent and more numerous than ’Counternegative’ tweets. In addition, in order to analyze the “Asian-related” and “Counternegative”’s influence on social media, we also ana- lyzed their number of retweets. We found that, on average, each negative tweet had 1.22 retweets, while the “Counternegative” had nearly half the retweets of the “Negative” one, 0.68.

We further analyzed the user’s role in our research. Users who had posted “Negative” or “Counternegative” tweets were labeled as “Negative user” and “Counternegative User”. If a user had posted both tweets, the user identity would be determined according to the category which had larger number of tweets. Figure 6. shows the accumulation number of new “Negative users” and “Counternegative users”. Similar to Figure 5, there was a big spike in both types of users during mid-March. As of May 31, there were 1,814 “Negative” users and 620 “Countenegative” users. Moreover,Every day’s new “Negative” user is 0.55, and new “Counternegative” user only have 0.16. Based on data analysis, we obtained that that average Negative tweets published by per negative user were 2.81, for average Counternegative tweets published by per Counternegative user were 1.30. This ratio was much smaller compared with the ratio for whole Negative tweets(5098) and Counternegative tweets(809). One hypothesis we put forward about the high number of

"negative" tweets was whether it is because some robots or users repeatedly post "Negative" or "Counternegative" tweets to create rumors or guide public opinion. We found that 61 out of 1814 Negative Users posted more than 10 negative tweets, among which the highest user posted 307 tweets. For Counternegative Users, there were 26 users posted more than 10 Counternegative tweets, with the highest user posting 223 tweets. The sheer number of 223 repeat tweets, compared to a total of only 803 "counternegative" tweets, is very huge. We removed users with these special circumstances, those who had posted more than 100 tweets, we found that the number of Asian-related and Counternegative tweets went from 5,098 to 4,573, and from 803 to 268, respectively. The ratio between Negative tweets and Counternegative tweets was larger compared before.

4. Conclusion

In the process of comparison of text classification performance, our research results show that, compared with TF-IDF in combination with LSA and TF-IDF Dimension Reduction by word frequency, TF-IDF without dimension reduction performed better in most cases. Furthermore, we compared the performance of text feature mining of different types of pre-trained BERT models on our data set. It is found that the deeper bert-large-uncased network structure, text feature mining does not bring better performance to the model, but increases the risk of overfitting. In comparison, the simpler Bert-base-uncased structure enjoyed better performance in most cases. We thought it is mainly confined to the data volume of training data. When the data volume is large enough, the deeper bert-large-uncased network structure should have better performance to extract more useful features. Compared with the traditional feature extraction method, TF-IDF, the pre-trained BERT had better performance when combined with most models. Moreover, when the classifier trained from one data set is used to predict other data sets, the model bias is a problem that cannot be ignored. In addition to the bias caused by the data imbalance in the training data set itself, the difference between two data sets may also aggravate the model's classification bias. The inherent bias of training data can be improved by modifying algorithms and models, etc. However, we have not found an effective way to solve the bias caused by the differences in data sets.

The paper also have limitations. First, our research on different models and algorithms is mainly limited to specific COVID-NEGATIVE data sets. The experimental results verified under the specific data are not universal. Secondly, only one researcher manually annotated the tweets data we collected, which made the annotated data were not reliable enough, as personal bias and understanding would influence the labeling results. If the circumstances permit, we suggest that two researchers can label respectively to improve labeling reliability. Moreover, our data and analysis are limited to the specific region of the Twitter platform in the United States. For the follow-up research, it is very useful to expand the social media platforms and regions of the research, and increase the scope of the research. Finally, what we study and pay attention to is the spread and influence of negative speech on social media based on twitter text. However, the spread of negative information is not limited to text. It may also be spread through videos, pictures and other forms. Follow-up research can focused on researching and analyzing the spread of negative information through pictures or other forms on social media.

References

[1] World Health Organization et al. The true death toll of covid-19: Estimating global excess mortality, 2021.

[2] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. Information, 10(4):150, 2019.

[3] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In Proceedings of the 40th international ACM SIGIR conference on research and development in

information retrieval, pages 115–124, 2017.

[4] Bertie Vidgen, Austin Botelho, David Broniatowski, Ella Guest, Matthew Hall, Helen Margetts, Rebekah Tromble, Zeerak Waseem, and Scott Hale. Detecting east asian prejudice on social media. arXiv preprint arXiv:2005.03909, 2020.

[5] Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian negative and counternegative in social media during the covid-19 crisis. arXiv preprint arXiv:2005.12423, 2020.

[6] Charu C Aggarwal and ChengXiang Zhai. A survey of text classification algorithms. In Mining text data, pages 163–222. Springer, 2012

[7] William S Noble. What is a support vector machine? Nature biotechnology, 24(12):1565–1567, 2006.

[8] Leo Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[10] Vandita Singh, Bhupendra Kumar, and Tushar Patnaik. Feature extraction techniques for handwritten text in various scripts: a survey. International Journal of Soft Computing and Engineering (IJSCE), 3(1):238–241, 2013.

[11] Samina Amin, M Irfan Uddin, Saima Hassan, Atif Khan, Nidal Nasser, Abdullah Alharbi, and Hashem Alyami. Recurrent neural networks with TF-IDF embedding technique for detection and classification in tweets of dengue disease. IEEE Access, 8:131522–131533, 2020.

[12] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation, 2004.

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. nature, 521(7553):436–444, 2015.

[14] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, Michelle X Zhou, and Huamin Qu. Context preserving dynamic word cloud visualization. In 2010 IEEE Pacific Visualization Symposium (PacificVis), pages 121–128. IEEE, 2010.

[15] Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf. Journal of documentation, 2004.

[16] Robert Dzisevi˘c and Dmitrij S˘es˘ok. Text classification using different feature extraction approaches. In 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), pages 1–4. IEEE, 2019.