Public attitudes analysis to Covid-19 vaccination based on natural language processing

Yibo Sun

Computer Science, University of Nottingham Ningbo China, Ningbo, Zhejiang province, 315199, China

Scyys9@nottingham.edu.cn

Abstract. The COVID-19 epidemic has spread globally since 2020, seriously affecting the order of social and economic development and endangering the lives of the people. With the continuous efforts of medical research institutions, the Covid-19 vaccine has been gradually launched, bringing hope for the prevention and treatment of the new coronavirus. However, many people still have doubts about the safety of these rapidly developed vaccines. In order to better promote vaccines to the rest of the public, it is important to determine their attitudes toward providing appropriate vaccines. Thanks to the rapid development of social networks and natural language processing technologies, collecting and analyzing attitude data from social media has proven to be an alternative solution. In this paper, we search for some of the most popular methods to test their efficiency and correctness in analyzing public attitudes toward vaccines. Extensive experimental results have verified the effectiveness of our work, which can provide new insights into automated surveys of attitudes toward Covid-19 vaccines.

Keywords: natural language processing.

1. Introduction

The novel coronavirus pneumonia has become prevalent around the world since 2020, seriously affecting the order of social and economic development and endangering the safety of people's lives. The covid-19 is highly infectious and pathogenic, which can cause serious respiratory illness and endanger the safety of people's lives [1]. According to the report from the World Health Organization, till November 2022, there are 629 million cases and 6.5 million deaths have been reported worldwide. The incredible numbers show that covid-19 has been one of the deadliest pandemics of the last two centuries. How to effectively prevent Covid-19 has become a very challenging task, which has attracted the research attention of a large number of medical experts and government organizations.

Recently, thanks to the continuous efforts of medical research institutions, the vaccine for Covid-19 has been gradually launched, bringing hope for the prevention and treatment of the new crown virus. The basic theory behind the development of the Covid-19 vaccine is to teach the body's immune system how to resist potential future infections, especially to eliminate the memory of the virus in T lymphocytes that appear after infection with Covid-19. Vaccines typically contain several substances: antigens, preservatives, stabilizers, surfactants, residues, diluents, adjuvants, and other substances. Among all these substances, the antigen is the most important ingredient for a vaccine to work. According to the difference in the antigen, vaccines can be divided into six kinds, live attenuated vaccines, killed or inactivated vaccines, toxoids, subunit and conjugate vaccines, mRNA vaccines, and viral vectors. Getting vaccinated does not mean that a person will never develop that disease, but it will undoubtedly greatly eliminate the possibility of becoming seriously ill.

After a vaccine is developed, it usually needs to go through three stages of different tests to verify its safety. In the first phase, a small number of young, healthy adult volunteers will be selected to test the vaccine. In the second phase, several volunteers of different ages will be vaccinated to assess behavior at different ages. In addition, an unvaccinated group will be added. The group aimed to demonstrate that the effects were due to the vaccine, rather than chance. The final phase, building on the second phase, expanded the scope of volunteers to thousands, possibly from different regions. However, most vaccine research and development institutions have scaled back the development process to produce vaccines to fight epidemics. Although in order to promote vaccination against covid-19, most countries have introduced promotional policies to use government credibility to encourage people to get vaccinated, many people still have doubts about the safety of these rapidly developed vaccines. What's more, some people express their concern about a certain medical corporation.

To better promote vaccine to the rest of the public, identifying their attitudes to delivering appropriate vaccine are very important. In this case, we urgently need a method to survey people's attitudes toward different vaccines. Since people are used to sharing life and attitude towards various events on social networking software, like Twitter and Facebook, collecting and analyzing the attitudes data from social media has been proven an alternative solution. The technology of Natural language processing aims to let the computer understand sentences in a natural language situation, which can be further applied to mine the users' hidden attitudes and sentiments about COVID-19 vaccination. On this occasion, the government can make a change to the original vaccination promotion. Thereby can rise the full vaccination rate and may help humans to defeat COVID -19.

NLP has a very long history, only several decades after the invention of the computer. At the early stage of research, scientists usually used hand-written rules to deal with NLP tasks [2]. Although this kind of method can deal with some common situations, its size of it is usually very tremendous. Besides, handwritten rules can't deal with all the possible situations in a natural communicating situation. For example, fragmented information may not influence the understanding of human beings, but it means nothing to computers based on handwritten rules. Besides, different rules may impact each other when there are some equivoques in the task. To solve these issues, scientists put forward the statistical NLP to supplement the shortcomings of the original structure. The statistical NLP mainly aims to build broader, fewer rules with the frequency of words to minimize ambiguity [2]. In recent years, with the development of computer hardware and deep learning, neural networks begin to be implemented in NLP tasks [3].

Focusing on the above goal that analyzing people's attitudes towards vaccination, in this paper, we search for some of the most popular methods to test their efficiency and correctness in analyzing the public attitudes toward a certain kind of vaccine. The preparation steps of this research can be roughly divided into data collection and preprocessing, model training based on different representative NLP methods, and results analysis, which we will introduce in detail in the following sections.

2. Methods

In this section, we will introduce in detail the representative methods we selected and compared, mainly including decision tree, bootstrap, random forest, Naïve Bayes classifier, and logistic regression.

2.1. Decision tree

In computer science, the decision tree is a non-parametric learning method under supervision, which is usually used for classification and regression. Compared to other neural network methods, the result can be understood by a human. Besides this method can deal with the loss of some eigenvalue [4].

The model of the decision tree will intake a list of parameters and their goal classes to create a tree. Each intermediate node in the resulting tree represents a decision. Each leaf indicates the result reached by a series of choices from the root of a tree to the bottom. From a geometric point of view, the n input data build a space of n dimensions. When applying this method to build decision surfaces to split the space [5]. The final minimal spaces represent the result after all related decisions. In this way, the smaller amount of input parameters, the smaller the decision tree is. Normally the smaller tree is, the higher accuracy and readability it has [5]. In the process of building a tree if tree temps to satisfy all the situations of input. It may lead to an overfitting problem. To avoid this cutting off the tree and limiting the maxim depth is important [6].

There are several available mature algorithms for building decision trees. For example, Chi-Squared Automatic Interaction Detection (CHAID) [7], Classification and Regression Trees (CART) [8-9], and Quick, Unbiased, Efficient, Statistical Tree (QUEST) [10]. Among all the algorithms, we directly choose the algorithm implemented based on the scikit-learn library.

2.2. *Bootstrap aggregating*

The bootstrap aggregating method is also known as bagging, which is part of the machine learning ensemble meta-algorithm. Bagging is one of the most intensive procedures used in high dimensional data set problems since it can significantly improve classifiers or unstable estimators in terms of computation [11]. The core idea of this bagging method is to generate several versions of predictors and train the predictors with randomly generated subsets [3]. When predicting the outcome with the function, the results over different versions will be aggregately averaged. When using regression trees and classification with a subset in linear regression of real data, bagging can significantly increase the accuracy.

There is a well-packaged Python function based on the bagging theory from scikit-learn, which provides the bagging classifier function based on four theories: Pasting, Bagging, Random Subspaces, and Random Patches. In this paper, we also chose to use this function instead of building an own function.

2.3. *Random forest classifier*

Random forest classifier was first proposed by Breiman et al. in 2001 [8], which is a method of ensemble learning. The random forest can be applied to a wide range of prediction problems with few parameters needing to be changed [12], especially showing excellent performance when the number of variables is much larger than the observations. When generating the model, the input dataset will be divided into two parts. The major part of the data is used to train the trees, and the rest will be used in the internal cross-validation technique to result in the performance [13]. If the model is used for classification tasks, the result made by most trees will be returned. When used for regression tasks, it will return the average result of all the trees instead. There is a well-packaged Python function based on random forest theory from scikit-learn. Because of the limited time, here researcher chose to use this function instead of building an own function.

2.4. *Naïve bayes classifier*

Naïve Bayes classifier is a kind of classifier applying the theorem of Bayes. The theorem can be simplified as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

Where A, B are events and P(B) is not equal to 0. In this model the input parameter can be seen as a vector $\mathrm{x} = (\mathrm{X}_1, .., \mathrm{X}_\mathrm{n})$. The result of this model is C. The equation can be seen as $\mathrm{P}(\mathrm{C}|\mathrm{X}_1, .., \mathrm{X}_\mathrm{n})$. The result $P(C|x) = \frac{P(x|C)P(C)}{P(x)}$. Use the relevant methods of discrete mathematics to reach the conclusion of the equation. The previous study has built some mature models. To avoid duplication of effort, one model from scikit-learn based on the Naive Bayes classifier is chosen to implement. This pre-build model is suitable for classifying discrete features.

2.5. *Logistic regression*

The logistic regression model is a statistical model calculating the probability of a certain event by a linear combination of several independent variables. The logistic function was invented to describe the sequence of autocatalytic chemical reactions and population growth in the 19th century by Verhulst [14]. In 1973, a constant of a California public transportation project: McFadden first linked the mathematical discrete choice with the multinomial logit. This provided his Nobel prize in 2000 for providing the theoretical foundation of this logit model. The researcher found a well-packed python-based model from scikit-learn. This model can handle both sparse and dense input.

3. Data collection and analysis

3.1. Data collection

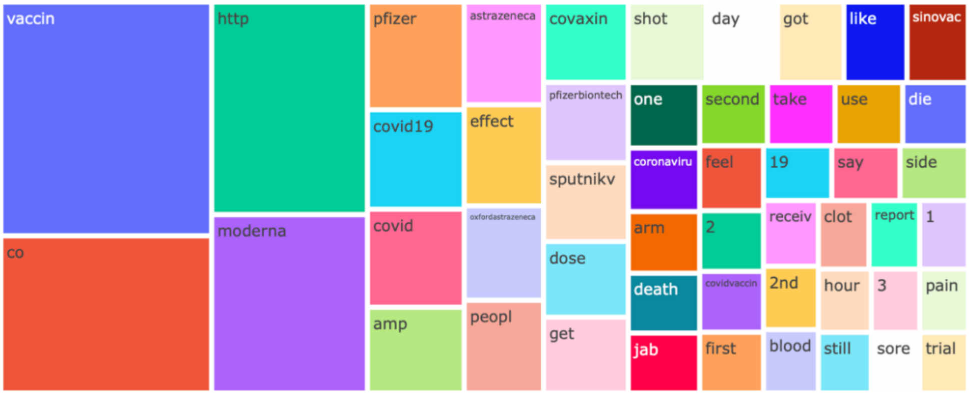
This research focuses on the message selected from online social media. Because of the time limits researcher had limited time to develop a web crawler collecting the latest and first-hand information. So searching for datasets that other scholars have collected and labeled in advance would be a good choice. In this research, researcher searched two datasets. One is smaller with the amount of 5991 values while the other is much bigger with more than 30000 values. In these two datasets only two column will be used: text and sentiment label.

3.2. Data analyses

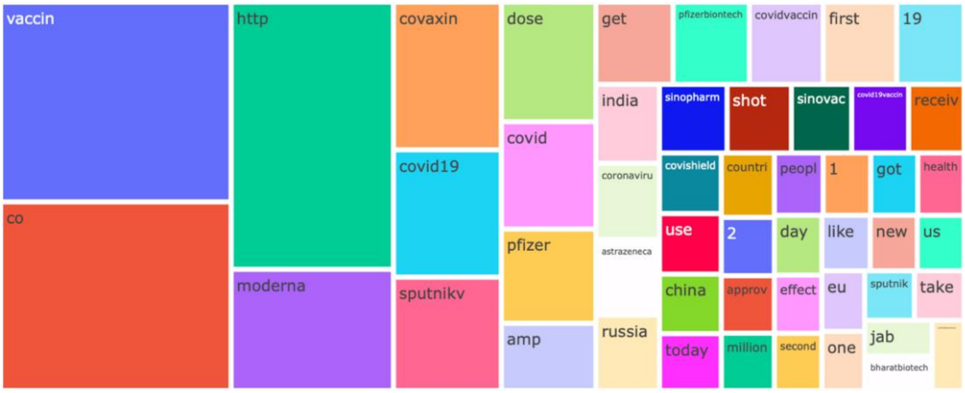
The key information in this research is the text and the pre-labeled sentiment. The sentiment does not need other special disposals. The text column is another key piece of information. The raw text contains a variety of information, including emoticons, tags, text that users type in, and web addresses that they paste. Because of encoding problems, many emoticons will not be acceptable for computers. And the web address may contain useful information for analyzing the sentiment of the writer. But due to the limited time, this kind of potential information needs to be removed from the text.

3.3. Data preprocessing

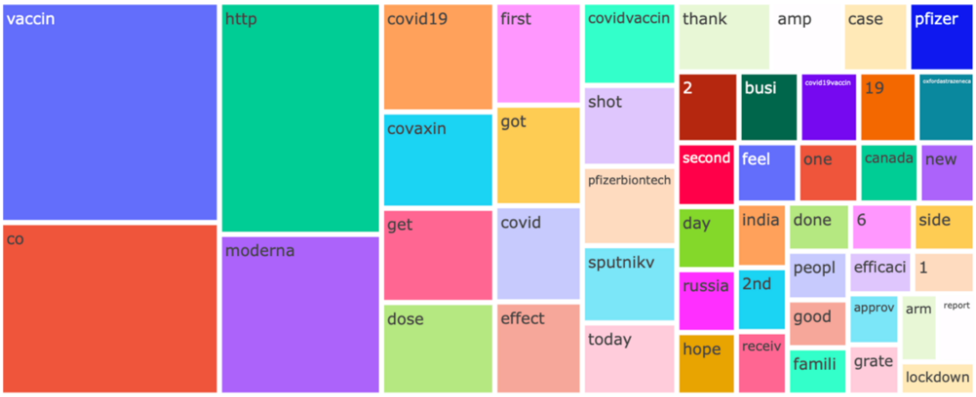
In this research, researcher uses regular expressions to simplify the text and remove useless text. After cleaning the text, use the function from the Python package sklearn to divide the whole sentence into small pieces and put them into a list as vectors for the following training. Here, we visualize the most common 50 words in our data set, as Figure 1 and 2 shown. It can be seen from those diagrams that there is no big difference between the three kinds of words with different emotions.



(a) 50 Most common negative words.

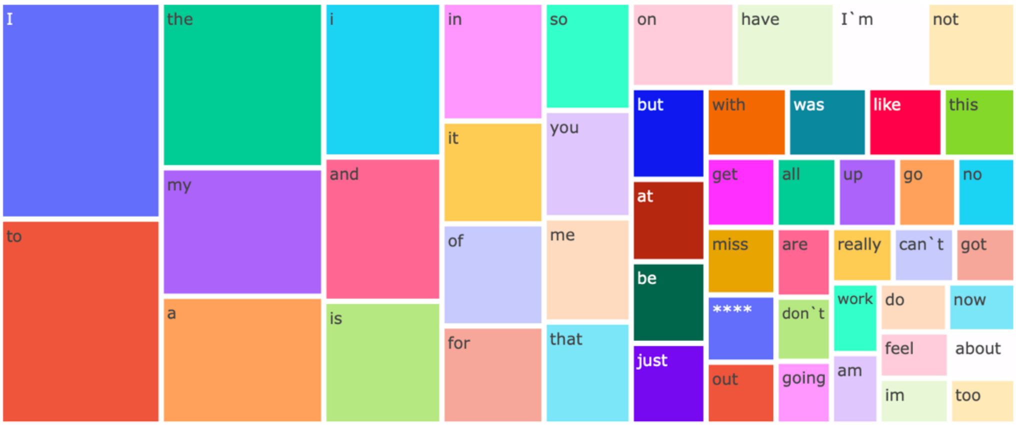


(b) 50 Most common neutral words.

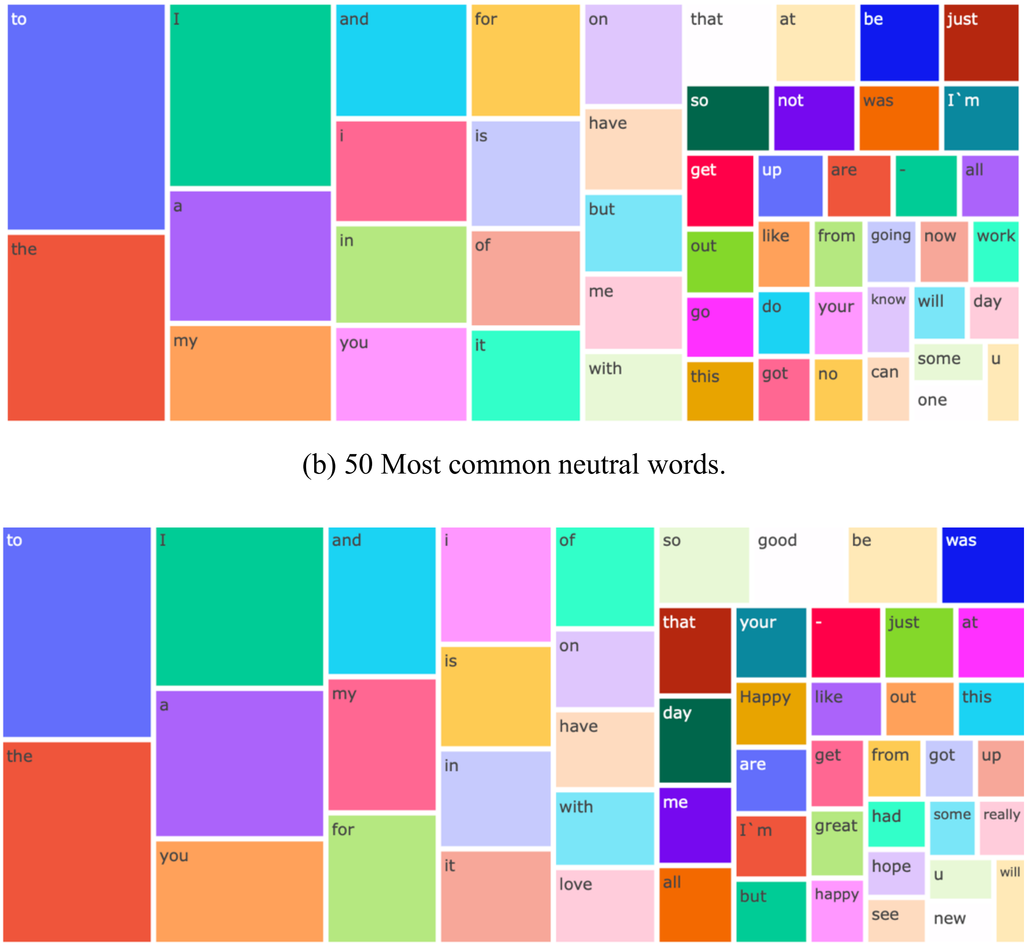


(c) 50 Most common positive words.

Figure 1. Most common 50 words of data set one.



(a) 50 Most common negative words.



(b) 50 Most common neutral words.

(c) 50 Most common positive words.

Figure 2. Most common 50 words of data set two.

4. Experiment and performance analysis

Before applying the methods, the data need to be divided into two parts. One for training the model, and another for testing the results. The rate of the training set and the testing set was settled to be 9:1. However, when the model is applied to the second dataset, the amount of data is too large to handle. To this end, we drop the rate to 5:5. When applying models on the bigger dataset two models met the problem. The linear regression model and bagging model cannot fit this dataset of its enormous quantity. Those two models took tons of time but did not return any useful results. in this case, researcher considers those two models are not suitable for this situation.

Table 1. Performance comparison of different methods on two datasets.

Methods	Data set 1		Data set 2	
	Training	Testing	Training	Testing
Decision tree	0.99	0.65	0.99	0.63
Bagging	0.97	0.68	N/A	N/A
Random Forest Classifier	0.99	0.70	0.99	0.67
Naïve Bayes classifier	0.94	0.69	0.97	0.63
Logistic Regression	0.99	0.72	N/A	N/A

Table 2. Time costs comparison of different methods on two datasets.

Methods	Data set 1 (s)	Data set 2 (s)
Decision tree	173	3054
Bagging	281	N/A
Random Forest Classifier	186	3155
Naïve Bayes classifier	21	266
Logistic Regression	65	N/A

As shown in Table 1 and Table 2, for the experiment on dataset 1, the training accuracy of every model reached 90. After applying the trained model to the remaining data, the accuracy rate turned out to be around 60% and 70%. Among these five models, the Multinomial model took the shortest time, the Bagging model took the longest time. The decision tree model presented the best, the accuracy is highest at around 73% meanwhile it took a moderate amount of time. The logistic regression model took the second longest and achieved the second highest accuracy, second only to the decision tree model. The random forest classifier model took time between the decision tree model and the bagging model. Besides the accuracy of the random forest classifier is also between those two models.

In the experiment on dataset 2, the remaining three models still reached the training accuracy of 90%. But their test accuracy normally dropped 5% each. With the size of the data tripled, the time each model took increased. The random forest classifier and Multinomial model took almost ten times as long as the previous one. The time cost of the decision tree model only doubled.

5. Conclusion

In this paper, we compare and analyze the results of different representative methods on two datasets, which aim to provide new insights into automated surveys of attitudes toward Covid-19 vaccines. Based on extensive experiments, we found that the trained model with more features will present well in forecasting. But the amount of training time will significantly increase. In this case, if other researchers want to implement the trained models to machines with limited performance, the decision tree will be recommended; otherwise, the more complicated model will show a better performance. In future work, data from various models will be further analyzed, because users are accustomed to combining pictures and text to express their opinions and emotions more fully on social media.

References

[1] D. Bajic, V. Dajic, and B. Milovanovic, "*Entropy analysis of COVID-19 cardiovascular signals*" Entropy, vol. 23, no. 1, p. 87, 2021.

[2] Prakash M Nadkarni, Lucila Ohno-Machado, Wendy W Chapman, *Natural language processing: an introduction,* Journal of the American Medical Informatics Association, Volume 18, Issue 5, September 2011, Pages 544–551

[3] Breiman L. *Bagging predictors*[J]. 1996, Machine learning, 24(2): 123-140.

[4] Kingsford C, Salzberg SL. *What are decision trees? Nat Biotechnol*. 2008 Sep;26(9):1011-3. doi: 10.1038/nbt0908-1011. PMID: 18779814; PMCID: PMC2701298.

[5] Quinlan J R. *Learning decision tree classifiers*[J]. ACM Computing Surveys (CSUR), 1996, 28(1): 71-72.

[6] Song Y Y, Ying L U. *Decision tree methods: applications for classification and prediction*[J]. Shanghai archives of psychiatry, 2015, 27(2): 130.

[7] Kass GV. *Anexploratory technique for investigating large quantities of categorical data*. Appl Stat. 1980;29: 119–127.

[8] Breiman L. Random forests[J]. *Machine learning*, 2001, 45(1): 5-32.

[9] Quinlan RJ. C4.5: *Programs for Machine Learning*. San Mateo California: Morgan Kaufmann Publishers, Inc.; 1993.

[10] Loh W, Shih Y. *Split selection methods for classification trees*. Statistica Sinica. 1997;7: 815–840.

[11] Bühlmann P, Yu B. *Analyzing bagging*[J]. The annals of Statistics, 2002, 30(4): 927-961.

[12] Biau G, Scornet E. *A random forest guided tour*[J]. Test, 2016, 25(2): 197-227.

[13] Belgiu M, Drăguţ L. *Random forest in remote sensing: A review of applications and future directions*[J]. ISPRS journal of photogrammetry and remote sensing, 2016, 114: 24-31.

[14] Pranckevičius T, Marcinkevičius V. *Application of logistic regression with part-of-the-speech tagging for multi-class text classification*[C]//2016 IEEE 4th workshop on advances in information, electronic and electrical engineering (AIEEE). IEEE, 2016: 1-5.