Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230935

Sentiment analysis to COVID-19 vaccination based on bert and LSTM

Shuyu Fang

Shanghai Jiao Tong University, Shanghai, Minhang District, 201100, China

858167250@sjtu.edu.cn

Abstract. The novel coronavirus (COVID-19) was defined as a pandemic in March 2020, which has brought great harm to the overall economy and people's health. Efforts to suppress and end the COVID-19 have led to an unprecedented vaccine rush, stemming from extensive research by experts and authoritative institutions. Previous research has found that significant numbers of people are hesitant to get vaccinated, worrying there may be some hurt for heath. To this end, a timely understanding of people's sentiments about the COVID-19 vaccination is crucial to the popularization of vaccines. Thanks to the popularity of the mobile Internet, people have become accustomed to expressing their opinions by posting comments on the Internet, which provides a quick way to obtain data for analyzing people's sentiment changes on the COVID-19 vaccine. In this paper, to gain an intuitive understanding of people's acceptance of a particular vaccine, we propose a tweet sentiment analysis method based on Bert and LSTM. Using N-grams to evaluate the model results, an accuracy rate of 74.14% can be obtained, which verifies the effectiveness of the proposed method. Extensive experiments show that our method can provide some new insight for the later vaccination policy promotion to a certain extent.

Keywords: Sentiment Analyses; n-gram; Bert; LSTM

1. Introduction

In March 2020, the World Health Organization (WTO) defined the novel coronavirus (COVID-19)[11] as a pandemic, which has brought great harm to the overall economy and people's health. Efforts to suppress and end the COVID-19 have led to an unprecedented vaccine rush, stemming from extensive research by experts and authoritative institutions. Although vaccination is considered as an effective measure to curb the COVID-19, previous studies have found significant numbers of people are hesitant to get vaccinated. They believe that limited by the research and development cycle and the lack of clinical trials, these short-term developed vaccines are relatively unstable and may cause some unknown harm to the human body. In this case, to boost vaccination rates, it is imminent to know people's emotional changes about whether to be vaccinated against the new crown vaccine in a timely manner.

Sentiment analysis[5] is a natural language processing (NLP)[1,2] technique that determines the sentiment of text by automatically identifying essential points in the text. The sentiment of the text can be positive (e.g. "I think vaccines can prevent the novel coronavirus"), negative (e.g. "I refuse to be vaccinated"), or neutral (e.g. "Today is Monday"). Sentiment analysis has been applied in various fields. For example, in the business world, it is possible to understand how customers feel about goods and services and make improvements based on understanding customer sentiment on social media, product

DOI: 10.54254/2755-2721/6/20230935

reviews, and survey responses. In the field of management, sentiment analysis can help managers monitor employee morale. Through employee sentiment analysis, managers can identify issues that may need to be addressed and understand employee reactions to changes within the company. In the political arena, sentiment analysis can help policymakers understand what policy audiences think, thereby promoting or dismantling policy implementation. Thanks to the rapid development of the Internet, since people are accustomed to sharing their attitudes and emotions towards social events on social media, such as the new crown vaccine, using sentiment analysis technology to determine the emotion of the text provides a way to quickly investigate people's attitudes towards vaccines. a feasible solution.

In this study, we propose a Twitter sentiment analysis method based on Bert and Long Short-Term Memory (LSTM), which aims to mine users' potential sentiments for the COVID-19 vaccine from massive social data. Specifically, we employ Bert to encode the input text and use LSTM to learn the global dependencies of different word vectors. Three text representation methods of unigram, bigram and trigram are used to split the total data set, where 80% is used for training and 20% is used for testing. Compared with other representative methods, such as MultinomialNB, Decision Tree Classifier, Random Forest Classifier, and Bagging Classifier, the accuracy of our method improves significantly, up to 74.14% on the validation set. Compared with using LSTM alone, we outperforms 5.9%. Extensive experiments have verified the effectiveness of our method, showing that it can help policy implementers to quickly and intuitively understand people's acceptance of specific vaccines, and to a certain extent assist in later policy promotion or policy changes.

The rest of the paper is organized as follows: section 2 gives the details of data collection and preprocessing. The proposed framework and methodology are described in Section 3. Section 4 presents the experimental results and the comparison between classification techniques. The conclusions of this paper and future work are defined in Section 5.

2. Data collection and preprocessing

2.1. Data collection

A dataset of views on COVID-19 vaccination (dataset 1) was used for the study. The dataset consists of 6000 samples. The data set is divided into User ID, Manual Annotation (1: Negative, 2: Neutral, 3: Positive), and Tweet Texts (Raw texts, URLS, Hashtags). The contents of each tag are as follows in Figure 1.

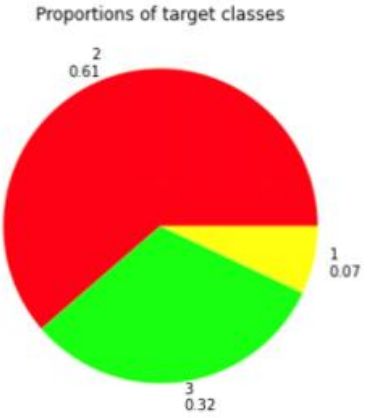


Figure 1. Tags distribution of data set 1.

Another dataset on sentiment analysis of tweets (Dataset 2) was used for the study. The dataset consists of 39,827 samples. The categories of this data set are: User ID, Tweet Texts, Sentiment Labels (12 detailed categories, 3 general categories), in which the proportion of each label is in Figure 2.

DOI: 10.54254/2755-2721/6/20230935

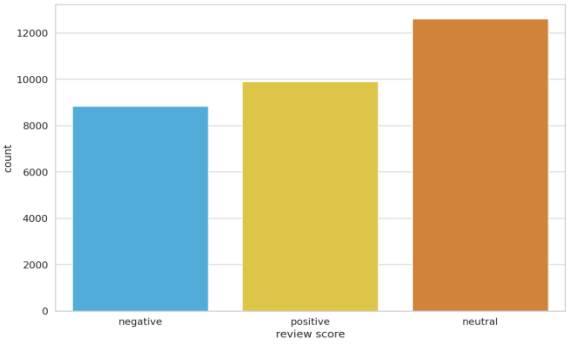


Figure 2. Proportion of each label in data set 2.

2.2. Feature selection

The goal of feature selection is to identify the most important features in a problem domain, through which the accuracy of prediction can be improved. In this paper, for data set 1, we first used Tokenizer, Filter step word, Extract step to pre-process the text content, and then used a)MultinomialNB,b)DecisionTreeClassifier,c)RandomForestClassifier,d)BaggingClassifier method to identify the most relevant features. For dataset 2, after preprocessing, Encode of bert was used, and CLS characters were used in LSTM to represent sentence features (that is, the beginning of each sentence has [CLS] character, which has no obvious semantics and will not be similar to any character in the sentence, so the corresponding vector of this character is used to represent the sentence).

2.3. Data splitting

Datasets 1 and 2 are split into a training set (80%) and a testing set (20%) at this step. The test set is used to evaluate the models after they have been developed using the training set.

3. Methodology

The workflow of this research mainly contains three different steps in this experiment. First, data set 1 was obtained from Kaggle. This data set had 6000 total samples, and the categories included three labels (positive, negative, neutral) and their corresponding texts. Secondly, use tokenizer, filter step word and extract step to preprocess the data. Then, using n-gram for the text according to the byte size of n(n= 1,2,3) sliding window operation, forming a sequence of n byte fragments, for the formation of each byte fragment frequency statistics, and according to the preset threshold filtering, forming a key gram list, that is, the feature vector space of the text. Finally, with 80% of the total data for training and 20% for testing, we can obtain the results of different representative methods including MultinomialNB, Decision Tree Classifier, Random Forest Classifier, and Bagging Classifier.

In order to better study the emotion analysis of tweets, we construct the model based on the Bert and LSTM on the data set 2 with a larger total sample size. There were more than 30,000 data sets in data set 2, which included 12 labels representing emotions and corresponding texts. Data preprocessing included tokenizer, filter step word, extract step, and label mapping (mapping 12 emotional labels to positive, negative, and neutral). The processed data were tested using Bert, LSTM and Bert+LSTM models to compare the advantages of each model.

3.1. N-gram

N-gram[4] is an algorithm based on statistical language model. It makes the contents of the text in bytes by the size of N sliding window operation, forming a length of N byte fragment sequence. A "gram" is a single fragment of bytes. A key gram list, which is the vector feature space of the text, is created by counting and filtering the frequency of occurrence of each gram in accordance with the predetermined threshold. A feature vector dimension is assigned to each gram in the list. The model is predicated on the idea that the Nth word's occurrence is only related to the N-1 words that came before it, and not to

DOI: 10.54254/2755-2721/6/20230935

any other words, and that the probability of the entire sentence is the sum of the probabilities of each word's occurrence. You can determine these probabilities by counting the simultaneous occurrences of N words in the corpus. Binary Bigram and ternary Trigram are frequently employed in this model.The Unigram, Bigram and Trigram can be formulated by equation (1)-(3), respectively.

$$P(w_1, \ w_2, \ \ldots, w_m) = \prod_{i=1}^{m} P(w_i) \quad (1)$$

$$P(w_1, \ w_2, \ \ldots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-1}) \quad (2)$$

$$P(w_1, \ w_2, \ \ldots, w_m) = \prod_{i=1}^{m} P(w_i|w_{i-2}w_{i-1}) \quad (3)$$

Where w_i means an individual word in a sentence.

3.2. Bert

The full name of Bert Model is Bidirectional Encoder Representations from Transformers, which is obtained by training the Masked Language Model and predicting the next task. In this study, the bert-base-uncased model was used.

3.3. LSTM

A special type of recurrent neural network (RNN) called long short-term memory (LSTM)[3,6,9] has a "gate" structure that decides whether data is updated or discarded under the logical control of a recurrent neural network (RNN). The forgetting gate, input gate, and output gate are the three gates that make up an LSTM, and they each control how much information is stored and forgotten at any given time. The amount of fresh information entered to the cell is determined by the input gate, whether it is forgotten at any given time is determined by the oblivion gate, and whether it is output at any time is determined by the output gate. The LSTM structure and examples used in this study are shown as follows.

For example, if you enter the string: The scenery here is very beautiful. The string will be divided into: [' (CLS) ', 'the', 'scenery', 'here' and 'is',' very ', 'beautiful', '. ', '[SEP]]. Their ids are: [101, 1996, 17363, 2182, 2003, 2200, 3376, 1012, 102]. In this experiment, each id corresponds to an embedding layer. After the embedding layer, these 9 ids will become 9×768 2D data input. The input data is normalized by layers, that is, any one of the 9 token embedding is normalized by means of a minus mean value divided by standard deviation. For example, the nine token embedding are represented by e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, each e_n (n=1, 2, 3, ..., 9) is 768 dimensions. Take e_1 as an example, subtract the mean and divide it by the standard deviation. The result is (e_1 – mean (e_1))/std (e_1). After that, the data enters the dropout layer, and in doing so, 20% of the neurons are randomly selected and the values of the selected neurons return to zero. Then through bidirectional LSTM, pooling (take '[CLS]' corresponding feature vector to represent sentence vector), full connection layer, and softmax. Finally, cross entropy and loss are calculated.

4. Experiment and performance analysis

4.1. Evaluation matrices

In dataset 1, the four models are tested using the three text representation methods in N-gram (i.e. unigram, bigram and trigram), including MultinomialNB, Decision Tree Classifier, Random Forest Classifier, Bagging Classifier Results were evaluated using Accuracy. In dataset 2, Bert, LSTM and Bert+LSTM models were used to compare their advantages and disadvantages. The results were evaluated using Accuracy and Confusion Matrix.

4.2. Experiment setting

To execute the proposed methods and models, the Jupiter notebook was used to obtain CPU and GPU help for Python research projects in the cloud. To implement the tests for the four models, the PYTHON

DOI: 10.54254/2755-2721/6/20230935

scikit-learn library was called. Bert's model uses bert-base-uncased. LSTM is designed using a neural network designer.

4.3. Result of N-gram model

As shown in FIG. 3, when the text representation method of unigram is used, the accuracy rate of the random forest-based model can reach about 0.72 at the highest. However, for the same random forest-based model, the accuracy of the other two text representation methods ranged from 0.62 to 0.64. When using unigram text representation method, the accuracy of the model is mostly higher than 0.7, while the accuracy of bigram and trigram text representation method is both lower than 0.65. As a result, it can be seen that the text representation method based on unigram has a higher accuracy in the test set than bigram and trigram. In unigram, the random forest-based model has the best accuracy.

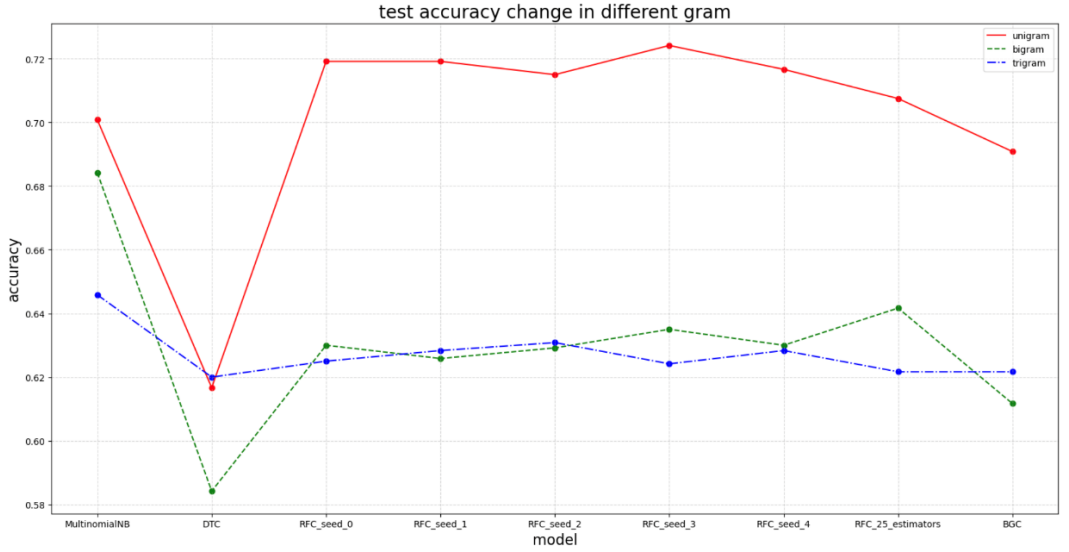


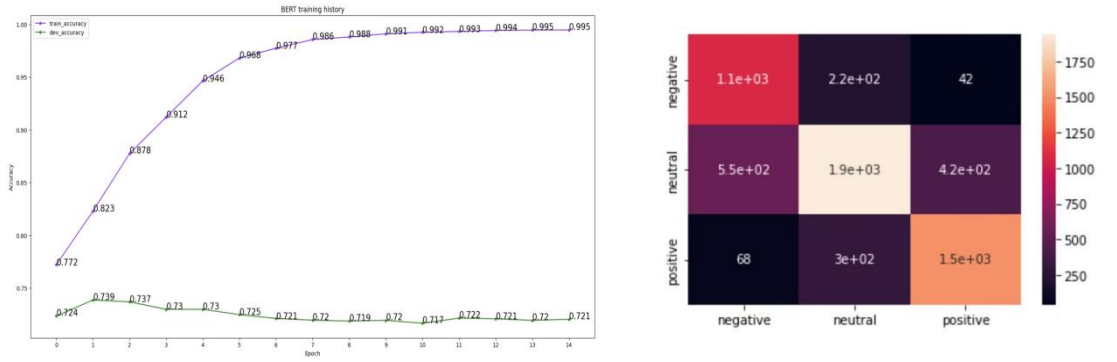
Figure 3. Performance comparison based on different text representation methods.

4.4. Performance comparisons

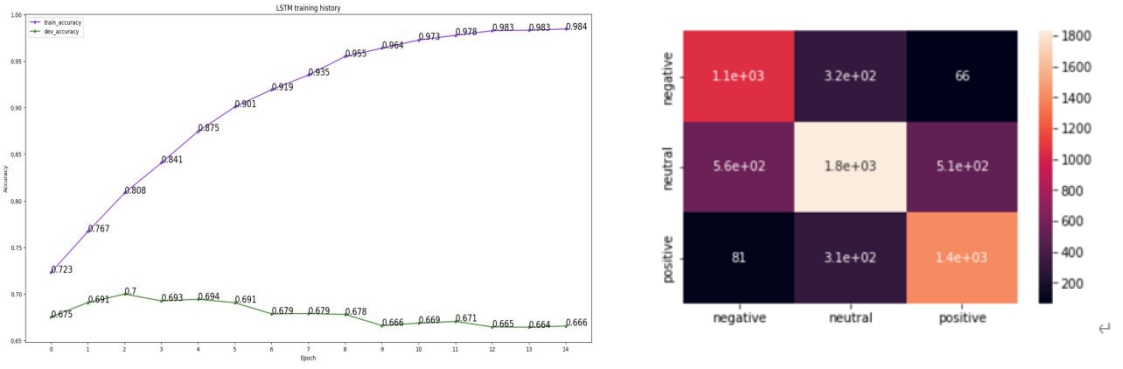
In this section, to verify the effectiveness of our proposed method, we compare the results of Bert, LSTM and Bert+LSTM models on the second data set, whose results can be seen in Figure 4. The best valid accuracy of Bert model is 73.88%, while the confusion matrix suggests that the positive and negative information will be probably recognized as neutral and the neutral ones are difficult to classify. For the LSTM model, its best valid accuracy is 69.99%. The confusion matrix suggests that the positive and negative information will be probably recognized as neutral and the neutral ones are difficult to classify. Compared with Bert model, LSTM has lower accuracy.

As shown in the figures above, the best valid accuracy of Bert+LSTM model is 74.14%. and the confusion matrix suggests that the positive and negative information will be probably recognized as neutral and the neutral ones are difficult to classify. Compared with Bert, the accuracy is improved by 0.35% and compared with LSTM, the accuracy is improved by 5.9%. Since LSTM is a special recurrent neural network, it is naturally suitable for encoding positional features due to the properties of network cycling. Although the BERT has a position embedding to encode the location features, adding another LSTM after the BERT can further enrich the location features. The final effect shows that adding LSTM after the BERT can further improve the validation set effect.

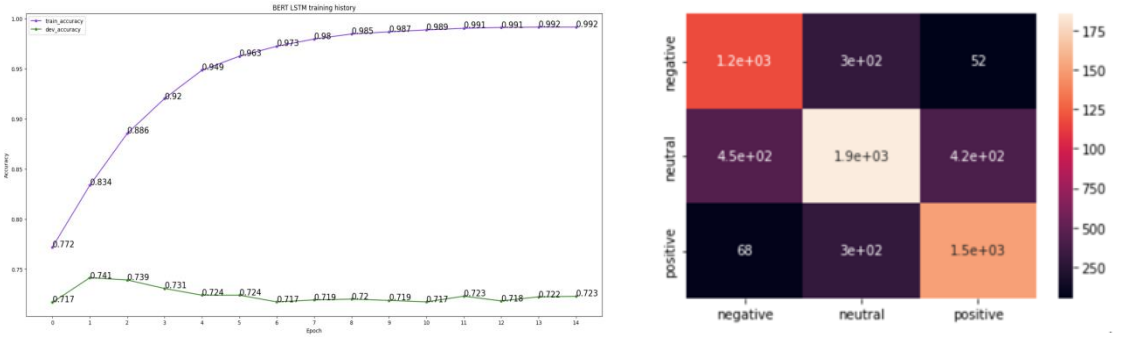
DOI: 10.54254/2755-2721/6/20230935



(a) Accuracy and confusion matrix of BERT model.



(b) Accuracy and confusion matrix of LSTM model.



(c) Accuracy and confusion matrix of BERT and LSTM.

Figure 4. Performance comparisons of different models.

5. Conclusion

After using N-gram to evaluate the previous research methods of machine learning, this study proposed the deep learning method of Bert and LSTM to analyze the sentiment of tweets and attained 74.14% accuracy.

In the study, n-gram was used to test multiple feature selection methods and classification techniques. Finally, it was concluded that the text representation method based on unigram was more accurate on the test set, and the RF-based model had better performance in unigram.

In addition, it can be seen from Bert, LSTM and Bert+LSTM models that LSTM has the characteristics of network circulation, so adding LSTM after Bert can further enrich the location characteristics. The confusion matrix suggests that Positive and Negative texts are easily identified as Neutral, while Neutral texts are difficult to identify.

DOI: 10.54254/2755-2721/6/20230935

5.1. Discussion

In this paper, we propose a tweet sentiment analysis method based on Bert and LSTM, which aims to gain an intuitive understanding of people's acceptance of a particular vaccine. According to the results, Bert+LSTM has the highest accuracy. However, Low accuracy, Overfitting and High demand on features still occur during the experiment. These problems may be caused by errors such as too little total data in the data set and unclear emotion of tweets obtained from the data set. The accuracy of test results may be improved by obtaining a larger sample size or by using other models (such as Deberta or Boruta)[7,8,10]. Since the model of LSTM is designed by neural network designer, changing the structure and parameters of LSTM may further improve the test results. Future work could look at how attitudes toward vaccines have changed over time, or using more models to study people's perceptions and emotional changes about different vaccines, or how people in different places and different cultures react to policies.

References

[1] A. Soni, B. Amrhein, M. Baucum, E. J. Paek and A. Khojandi, "Using Verb Fluency, Natural Language Processing, and Machine Learning to Detect Alzheimer's Disease," 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Mexico, 2021, pp. 2282-2285.doi: 10.1109/EMBC46164.2021.9630371

[2] Bose P, Roy S, Ghosh P. A Comparative NLP-Based Study on the Current Trends and Future Directions in COVID-19 Research. IEEE Access. 2021 May 20;9:78341-78355. doi: 10.1109/ACCESS.2021.3082108. PMID: 34786315; PMCID: PMC8545210.

[3] C. Li, G. Zhan and Z. Li, "News Text Classification Based on Improved Bi-LSTM-CNN," 2018 9th International Conference on Information Technology in Medicine and Education (ITME), Hangzhou, China, 2018, pp. 890-893.doi: 10.1109/ITME.2018.00199

[4] D. Nagalavi and M. Hanumanthappa, "N-gram Word prediction language models to identify the sequence of article blocks in English e-newspapers," 2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2016, pp. 307-311. doi: 10.1109/CSITSS.2016.7779376

[5] K. Khan and S. Yadav, "Sentiment analysis on covid-19 vaccine using Twitter data: A NLP approach," 2021 IEEE 9th Region 10 Humanitarian Technology Conference (R10-HTC), Bangalore, India, 2021, pp. 01-06.doi: 10.1109/R10-HTC53172.2021.9641515

[6] M. Sushmitha, K. Suresh and K. Vandana, "To Predict Customer Sentimental behavior by using Enhanced Bi-LSTM Technique," 2022 7th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2022, pp. 969-975.doi: 10.1109/ICCES54183.2022.9835947

[7] Q. Mi, Y. Gao, J. Keung, Y. Xiao and S. Mensah, "Identifying Textual Features of High-Quality Questions: An Empirical Study on Stack Overflow," 2017 24th Asia-Pacific Software Engineering Conference (APSEC), Nanjing, China, 2017, pp. 636-641.doi: 10.1109/APSEC.2017.77

[8] S. S. Kumar and T. Shaikh, "Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest," 2017 International Conference on Computer and Applications (ICCA), Doha, Qatar, 2017, pp. 227-231.doi: 10.1109/COMAPP.2017.8079769

[9] T. S. N. Ayutthaya and K. Pasupa, "Thai Sentiment Analysis via Bidirectional LSTM-CNN Model with Embedding Vectors and Sentic Features," 2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), Pattaya, Thailand, 2018, pp. 1-6.doi: 10.1109/iSAI-NLP.2018.8692836

[10] X. Ye and S. Manoharan, "Performance Comparison of Automated Essay Graders Based on Various Language Models," 2021 IEEE International Conference on Computing (ICOCO), Kuala Lumpur, Malaysia, 2021, pp. 152-157.doi: 10.1109/ICOCO53166.2021.9673585

[11] Yasmin F, Najeeb H, Moeed A, Naeem U, Asghar MS, Chughtai NU, Yousaf Z, Seboka BT, Ullah I, Lin CY, Pakpour AH. COVID-19 Vaccine Hesitancy in the United States: A

DOI: 10.54254/2755-2721/6/20230935

Systematic Review. Front Public Health. 2021 Nov 23;9:770985. doi: 10.3389/fpubh.2021.770985. PMID: 34888288; PMCID: PMC8650625.