COVID-19 Vaccine Misinformation Detection

Richeng Zhang, Zhanfu Yang, Tengyue Chen

Course Section: Thursday 12:30pm to 3:00pm

May 8th, 2022

Contents

Introduction

Misinformation is defined as false or misleading information that attempts to imitate the content and form of news, comments or articles, and the false information misleads or deceives the public through dissemination in the media (Lazer et al., 2018). As the COVID-19 spreads globally, the fake news related to the COVID-19 appears in the media. World Health Organization warns that due to disinformation related to COVID-19, it is difficult for the public to find reliable sources and trustworthy information (World Health Organization, 2019). Fake news about the Covid-19 pandemic includes false or misleading information about Covid-19 vaccines. The confidence in a Covid-19 vaccine is threatened by misinformation. In Europe, people are highly hesitant about vaccines due to misinformation about vaccine side effects (Marco-Franco et al., 2021) Therefore, the massive spread of misinformation about the COVID-19 vaccine on social media could lead to a continuous reduction in the COVID-19 vaccination rate. This project suggests a model for detecting misinformation about the COVID-19 vaccine on social media.

Literature review

Misinformation about COVID-19 Vaccine

Misinformation is considered false or misleading information spread in the media (Lazer et al., 2018). Social media is easy to use, low cost, and fast, making it easier for misinformation to reach audiences on social media ((Shu et al., 2017) As social media becomes an essential communication tool and source of information in people's lives, the impact of misinformation is more widespread and important. COVID-19 is a pandemic disease in the world since 2019. The research shows that 47% of people in the United States expose to fake news about the COVID-19 in 2020 (Casero-Ripolles, 2020). The COVID-19 vaccine is one of the measures people take to fight the COVID-19 pandemic. Because misinformation can easily influence public decisions or sentiment,

misinformation can mislead the public about COVID-19 vaccine. The COVID-19 vaccine is considered to be one of the effective ways to mitigate the spread of the COVID-19 virus, and the government vigorously promote the vaccine (Conte et al., 2020). However, some people should refuse the vaccine. An important reason for the public to question vaccines is distrust of health providers, vaccine manufacturers, and public health agencies (Liu & Chu, 2022). Hesitation about a COVID-19 vaccine is largely driven by. Study in Nigeria found that study participants were heavily exposed to misinformation about COVID-19 and misinformation influenced their attitudes towards vaccines (Wonodi et al., 2022). The US study also found that distrust of the government and medical institutions led people to not believe in vaccines, and the information source of participants about the COVID-19 is the media (Morales et al., 2022).

Misinformation about COVID-19 affects trust in vaccines. The public affected by the COVID-19 fake news in the early stages of the pandemic perceives public health experts as exaggerating the severity of the pandemic in the US ((Laato et al., 2020). Worse yet, fake news about COVID-19 induces people to associate vaccines with conspiracy theories (Allington, 2021). The link between misinformation and conspiracy theories may explain why vaccine hesitancy is driven mainly by distrust of government. Another serious consequence is people sharing misinformation online to mislead others into not following public health advice and using dangerous fake medical methods (Morales et al., 2022). Therefore, fake news about vaccines receives more attention in the world.

COVID-19 Vaccine Misinformation Detection

Although detecting misinformation is difficult, the detection of misinformation never stops. Misinformation detection mainly relies on content analysis. Zhang and Ghorbani study characteristics of misinformation include a large number of volumes, wide variety, and fast speed (Zhang & Ghorbani, 2019). Meanwhile, they propose to detect misinformation based on creator and user analysis, news content analysis, and social context analysis (Zhang & Ghorbani, 2019). Bakir and McStay found that misinformation could manipulate the audience's emotions, and they suggested identifying misinformation through sentiment analysis of misinformation content (Bakir & McStay, 2017). Social

media platforms try to prevent and reduce the harm of fake news through content analysis. Facebook allows users to flag and report possible misinformation, and other users receive warnings when they share flagged misinformation (Mosseri, 2019). Twitter also uses hashtags and warning messages to reduce the impact of misinformation on users (Roth & Pickles, 2020). Google chooses to partner with authoritative news outlets to prevent the spread of misinformation (Google, 2022). Measures of social media platforms are built on content analysis. Social media platforms identify misinformation through user analysis of content. However, social media platforms are limited by the influence of free speech, and it is difficult for social media platforms to clean the misinformation completely. The social media platforms do not publicly announce that they would remove all misinformation.

The detection of misinformation about the COVID-19 vaccine in social media is mainly based on content and key words. One way to detect misinformation is a social network and neural network model. The social network model is to create a label as nodes about the keywords of the tweets, and the model use the Louvain community detection algorithm to build a social network to output misinformation datasets (Muric et al., 2021). However, the social network model lacks a method to verify the accuracy rate and the function of prediction, and the output is only a dataset of misinformation. Graph link prediction is another neural network method for misinformation detection. In Graph link prediction, nodes represent tweets with misinformation about the COVID-19 vaccine, while links between tweets indicate that they share the same misinformation, and evaluation of Graph link prediction shows detection outperforms classification (Weinzierl & Harabagiu, 2021). Another method of detecting misinformation is machine learning. Because the content of fake news on social media has strong emotional fluctuations, constructing a sentiment score through Topic Modeling can detect misinformation (Hu et al., 2021). The construction of the sentiment analysis model needs to construct the corresponding sentiment score for the sentiment expression of the content. XGBoost, LSTM, and BERT are commonly used models for machine learning and deep learning. In the detection of COVID-19 vaccine misinformation detection, XGBoost has the fastest training speed and BERT has better performance (Hayawi et al., 2022).

The challenge of COVID-19 vaccine misinformation detection on social media is the identification of labels. Because the habits and using of user language are inconsistent in social media, the definition of misinformation is difficult to achieve. Therefore, different narratives of fake news become one of the challenges in studying fake news detection. The previous studies (Muric et al., 2021; Weinzierl & Harabagiu, 2021; Hayawi et al., 2022) collected some common misinformation labels (see Table 1), and the labels of false information are mainly based on commonly used keywords that are obvious in society. Meanwhile, in the model training stage, because some tweets contain emotions such as humor and sarcasm, the model under training is recommended to be screened manually (Hayawi et al., 2022).

Keyword for COVID-19 vaccine misinformation
Abolish big pharma
antivaccine
Arrest Bill Gates
Between meandry doctor
Big pharma kills
Bill Gates Bio Terrorist
Bill gates evil
The COVID-19 vaccine causes infertility or miscarriages in women.
Natural COVID-19 immunity is better than immunity derived from a COVID-19 vaccine.
RNA alters a person's DNA when taking the COVID-19 vaccine

Table1. Example of keyword for COVID-19 vaccine misinformation

Research Question

Rebuttals that debunk and clarify misinformation are considered valid by some people (Wang et al., 2021). Alternatively, news media companies and statutory bodies jointly enforce incentives and sanctions and hold misinformation sources accountable (Gupta et al., 2022). However, regardless of the method used, detecting and screening

misinformation is an important step. Considering the importance of detection of COVID-19 vaccine misinformation, this study argues that a way to detect COVID-19 vaccine misinformation is needed.

This study tries to build a model to detect fake news or misinformation about the COVID-19 vaccine on Twitter. The misinformation label of this study is basic on the keyword from the previous studies. The researchers manually detect whether it is misinformation according to the label in the training phase, and the model will try to realize the automatic detection of misinformation by the model through multiple training of the model in the future. The other future research could find a solution to distinguish misinformation about the COVID-19 vaccine through the model. And the models can help people to understand whether the tweets are fake COVID-19 vaccine information.

Methodology

Data Crawling

In our project, we need two datasets, training dataset and testing dataset. What we find in GitHub is a training dataset with only labels and tweet IDs, so we need to scrape the tweets according to tweet IDs.

The other dataset is testing dataset, we use “covid vaccine” as key word to scrape tweets from Twitter. In this case, we need to filter the tweets that are retweeted because they may affect the result of our detection. In this project, we select 3000 tweets as the testing subset.

The code of these two processes is shown in file: Labeled Tweets Scraping & Tweets Analysis and COVID Vaccine Text Scraping.

Exploratory Data Analysis

As the Figure 4.1 shown, there are only two columns in our dataset, the labels of tweets and tweets’ contents. There is no null value in the training dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 12751 entries, 0 to 2483
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   is_misinfo    12751 non-null  int64
 1   text content  12751 non-null  object
dtypes: int64(1), object(1)
```

Figure 4.1

The Figure 4.2 and Table 4.1 show the distribution of normal COVID vaccine tweets and COVID vaccine misinformation tweets. Label 1 represents COVID vaccine misinformation tweets and Label 0 represents normal COVID vaccine tweets.

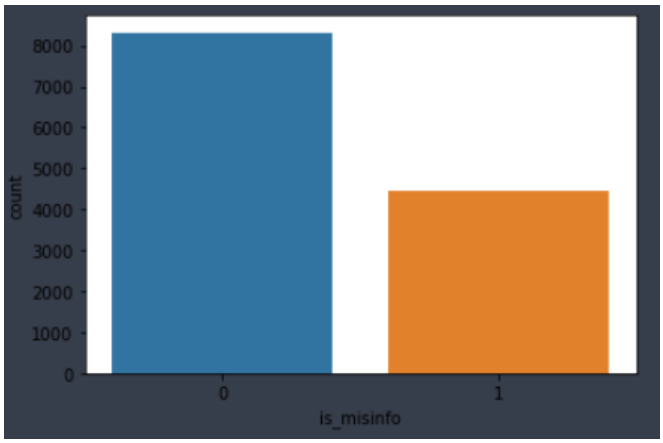


Figure 4.2

Labels	Number
1	8315
0	4436

Table 4.1

After text preprocessing, we can show the keywords contained under different labels. The Figure 4.3 shows keywords of normal COVID vaccine tweets.

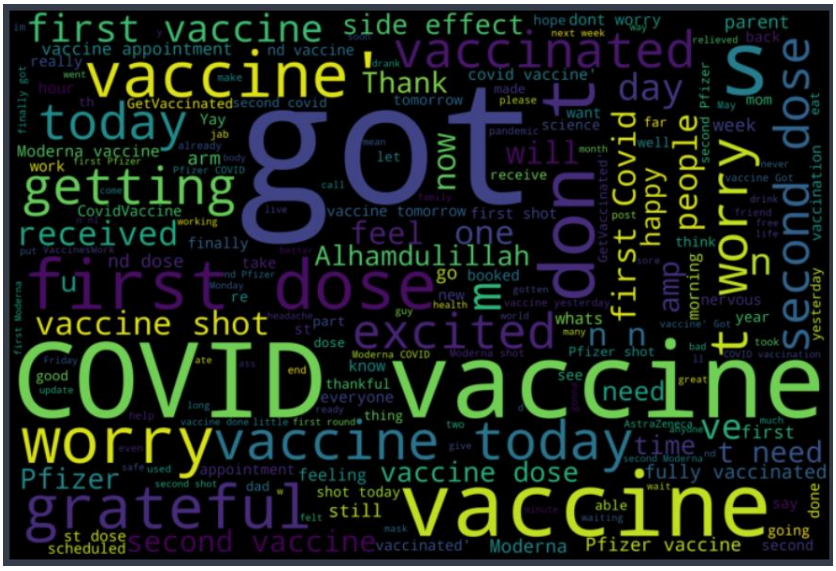


Figure 4.3

In this graph, we can see some positive words such as excited, grateful, thankful and thank.

Figure 4.4 shows keywords of COVID vaccine misinformation tweets.

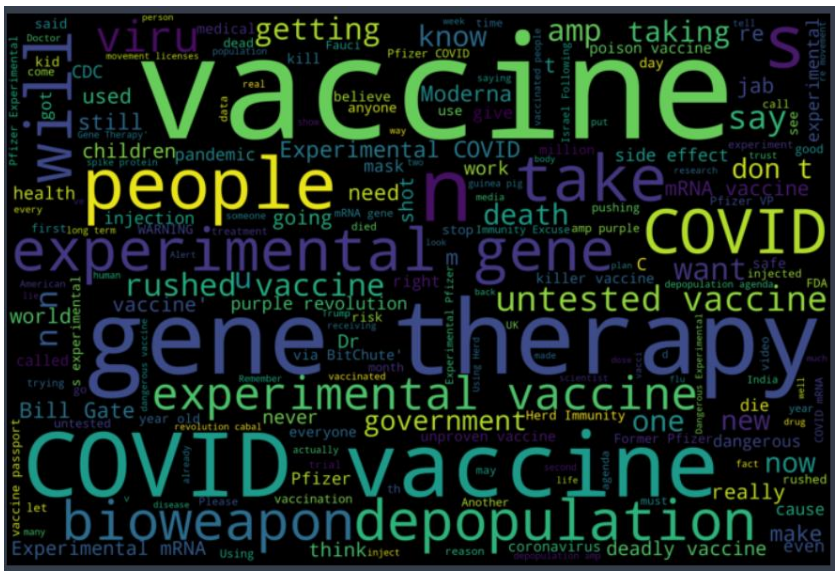


Figure 4.4

In Figure 4.4, we can find several negative words like experimental vaccine, gene therapy, bioweapon, depopulation, death, etc.

After preprocessing the test data, which is crawling from Twitter, we are able to generate a word cloud of keywords. As Figure 4.5 shown below.

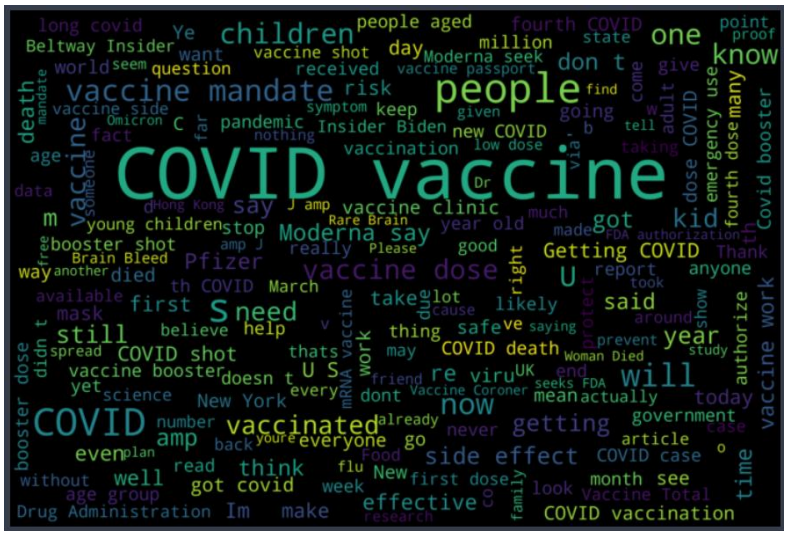


Figure 4.5

As we can see, there are not only some positive words, like believe, vaccine work, safe, but also some negative words, like death, brain bleed, died, etc.

TF-IDF & kNN Model

In order to detect which tweet we crawl from Twitter is COVID vaccine misinformation, we use TF-IDF & kNN model.

TF-IDF (term frequency–inverse document frequency) is a common weighting technique for information retrieval and data mining, commonly used to mine keywords in articles, and this algorithm is simple and efficient, often used by industry for the initial text data cleaning (Joon-Min Gil, 2019).

kNN is one of the most used classification algorithms, and a machine learning algorithm of supervised learning. The structure of the model established by kNN is determined based on data (Yun-lei Cai, 2010).

In this model, we use TF-IDF and kNN to finish the detection work. It's important for us to select a proper and effective k values, according to Figure 4.6 and 4.7, we can see that when k value equal to 9, the accuracy score is the highest. We use the features and labels to train this model.

	Score	k values
0	0.871145	1
1	0.842133	2
2	0.875327	3
3	0.859383	4
4	0.884213	5
5	0.884213	6
6	0.889963	7
7	0.876111	8
8	0.892316	9
9	0.877156	10
10	0.887872	11

Figure 4.6

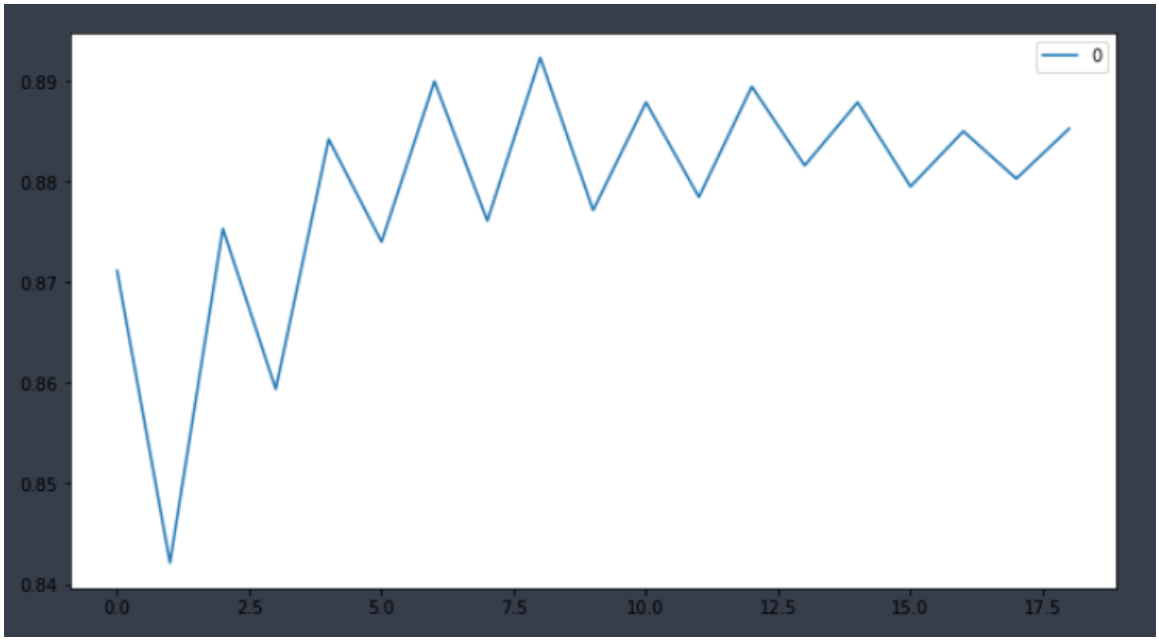


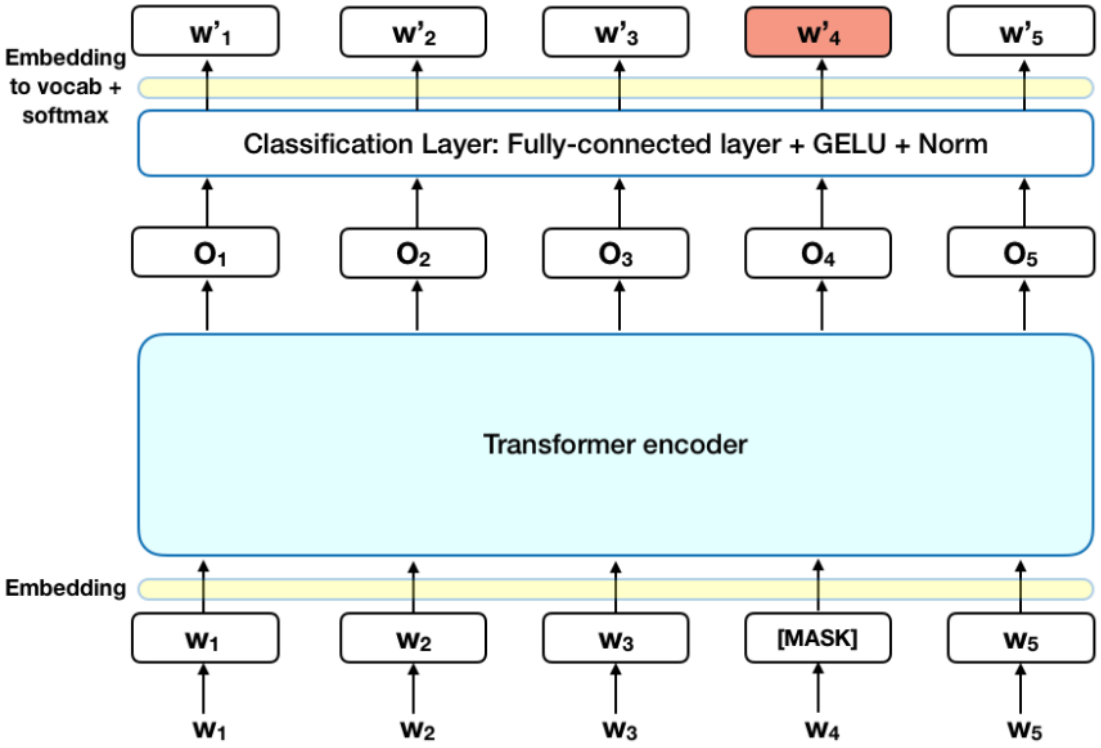
Figure 4.7

Bert Model

BERT (Bidirectional Encoder Representations from Transformers) is a recent published by researchers at Google AI Language. It has caused a stir in the Machine Learning

community by presenting state-of-the-art results in a wide variety of NLP tasks. (Jacob Devlin et al., 2019).

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. In its vanilla form, Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task. Since BERT's goal is to generate a language model, only the encoder mechanism is necessary.



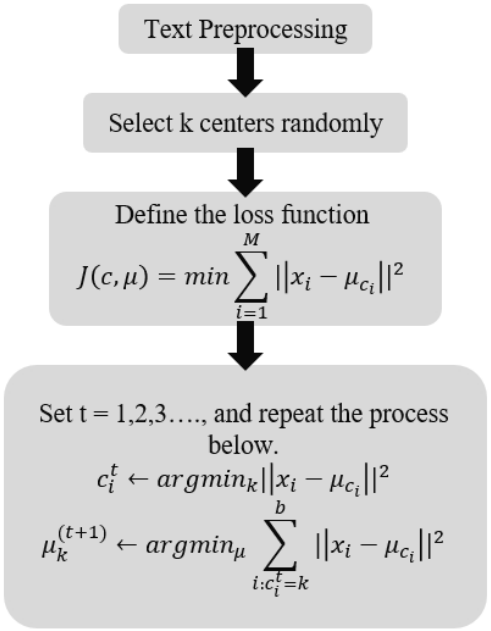
Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence.

Model Comparison (Training Dataset)

Firstly, we use 4 methods to generate the performance of each method and calculate precision, recall, f1score and accuracy of the training dataset.

K-means

K-means is a simple algorithm in clustering. It divides the sample data into K clusters according to the distance between the samples. It's aimed at making the points that in one cluster closer and the distance between different clusters as large as possible. The classification report of K-means is shown in Figure 4.8.



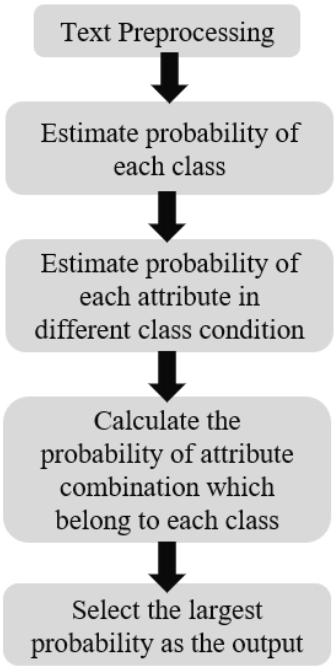
Cosine Distance:

	precision	recall	f1-score	support
0	0.94	0.78	0.85	1627
1	0.70	0.92	0.80	924
accuracy			0.83	2551
macro avg	0.82	0.85	0.82	2551
weighted avg	0.86	0.83	0.83	2551

Figure 4.8

Naïve Bayes

Naïve Bayes classifier is based on the training set T and estimate the prior probability P(c) of different classes and condition probability of each attribute P(xi|c). Here the maximum likelihood estimation (MLE) is used to estimate the corresponding probability. The classification report of Naïve Bayes is shown in Figure 4.9.



Naive Bayes:

	precision	recall	f1-score	support
0	0.96	0.96	0.96	1718
1	0.91	0.92	0.92	833
accuracy			0.94	2551
macro avg	0.94	0.94	0.94	2551
weighted avg	0.94	0.94	0.94	2551

Figure 4.10

Logistic Regression

Logistic Regression is a machine learning method, and it is used to solve the binary classification problem and estimate the likelihood of data samples.

$$sigmoid\ function: \sigma(x) = \frac{1}{1+e^{-x}}$$

And then combine the linear regression function with the sigmoid function and use the linear regression function's output as input of sigmoid function. We can get the logistic regression model.

$$y = \sigma\big(f(x)\big) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$$

If we have w^T, we can get y and it can be used to classify the data.

$Loss\ function: F(w)$

$$= \ln(P_{total}) = \ln\Big(\prod_{n=1}^{N} p^{y_n}(1-p)^{1-y_n}\Big) = \sum_{n=1}^{N} \ln(p^{y_n}(1-p)^{1-y_n})$$

$$= \sum_{n=1}^{N} (y_n \ln(p) + (1-y_n)\ln(1-p))$$

$$p = \frac{1}{1+e^{-w^T x}}$$

MLE (Maximum Likelihood Estimation) is a method of estimating parameters w.

The classification report of Logistic Regression is shown in Figure 4.10.

```
LogisticRegression()

              precision    recall  f1-score   support

           0       0.93      0.96      0.95      1691
           1       0.92      0.87      0.89       860

    accuracy                           0.93      2551
   macro avg       0.93      0.91      0.92      2551
weighted avg       0.93      0.93      0.93      2551
```

Figure 4.10

Xgboost

Xgboost (Extreme Gradient Boosting) uses gradient boosting as framework. It is developed from GBDT, and it optimize the learning process by using additive models and forward step-by-step algorithms.

We set the training dataset as $T = \{(x_1, y_1), (x_2, y_2) \ldots . (x_n, y_n)\}$, and loss function as $l(y_i, \hat{y}_i)$, and regularization term $\Omega(f_k)$

And the entire function equal to $L(\Phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k)$

Expression in linear space: $L(\Phi)$, i means the number i sample, k means the number k tree. $\hat{y}_i$ represents the prediction of number i sample x_i.

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i)$$

And then there are three steps can be taken to optimize the Xgboost function.

Step 1: Expand the two-rank Tailor, remove the constant terms and optimize the loss function.

Step 2: Expand the regularization term, remove the constant terms and optimize the regularization term.

Step 3: Combine the One-order term coefficient and quadratic term coefficient.

The classification report of Xgboost is shown in Figure 4.11.

xgboost:

	precision	recall	f1-score	support
0	0.94	0.94	0.94	1691
1	0.88	0.87	0.88	860
accuracy			0.92	2551
macro avg	0.91	0.91	0.91	2551
weighted avg	0.92	0.92	0.92	2551

Figure 4.11

The comparison of different method's f1 score is shown in 4.12.

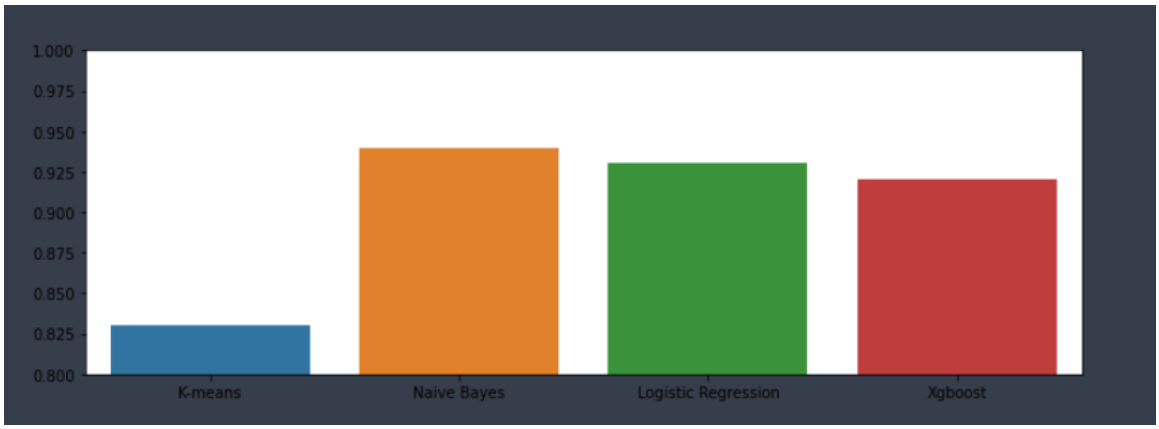


Figure 4.12

COVID Vaccine Misinformation Detection

As we mentioned in creating a TF-IDF & kNN model, we find that when k equal to 9, the model has the best performance, so we set k equal to 9 and train the model. After training the TF-IDF & kNN model, we use the testing dataset as input and predict the label of different text. The result is shown in Figure 4.13 and the distribution of normal COVID vaccine tweets and COVID vaccine misinformation tweets is shown in Table 4.2 and Figure 4.14.

	a	Label
0	those blood pressure medication have have deca...	0
1	kettering health doctor work build trust close...	0
2	now the time get boost household member covid ...	0
3	watch dudu sherarami director public health en...	1
4	partnership with the cayuga county health depa...	0
...	...	...
2998	why miralax laxative the covid vaccine have sp...	0
2999	icymi florida republican rep byron donalds joi...	0
3000	file this under shit ive be say since hence th...	1
3001	the news moderna now have billion sign deal fo...	0
3002	injection objection guy hatchard discuss polis...	0

Figure 4.13

Labels	Number
0	2183
1	817

Table 4.2

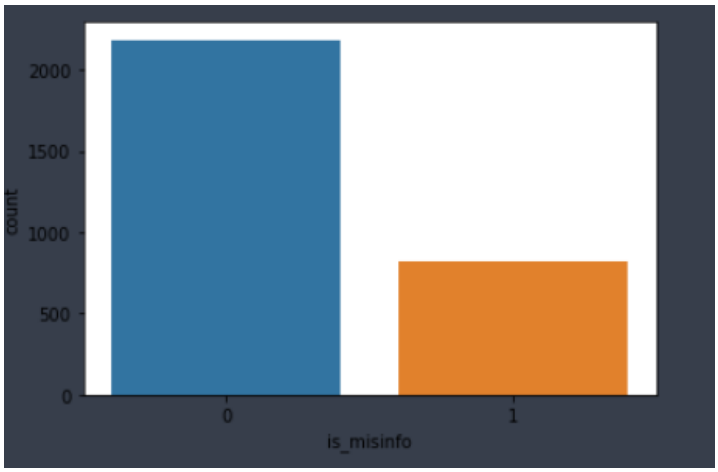


Figure 4.14

Model Comparison (Testing Dataset)

After the testing dataset is labeled, we use different methods to calculate the classification report of the testing dataset which is crawled from Twitter.

K-means

The classification report of K-means is shown in Figure 4.15.

Cosine Distance:

	precision	recall	f1-score	support
0	0.78	0.97	0.87	448
1	0.70	0.21	0.32	152
accuracy			0.78	600
macro avg	0.74	0.59	0.59	600
weighted avg	0.76	0.78	0.73	600

Figure 4.15

Naïve Bayes

The classification report of Naïve Bayes is shown in Figure 4.16.

Naive Bayes:

	precision	recall	f1-score	support
0	0.83	0.89	0.86	425
1	0.68	0.57	0.62	175
accuracy			0.80	600
macro avg	0.75	0.73	0.74	600
weighted avg	0.79	0.80	0.79	600

Figure 4.16

Logistic Regression

The classification report of Logistic Regression is shown in Figure 4.17.

LogisticRegression()

	precision	recall	f1-score	support
0	0.81	0.94	0.87	425
1	0.77	0.46	0.58	175
accuracy			0.80	600
macro avg	0.79	0.70	0.73	600
weighted avg	0.80	0.80	0.79	600

Figure 4.17

Xgboost

The classification report of Logistic Regression is shown in Figure 4.18.

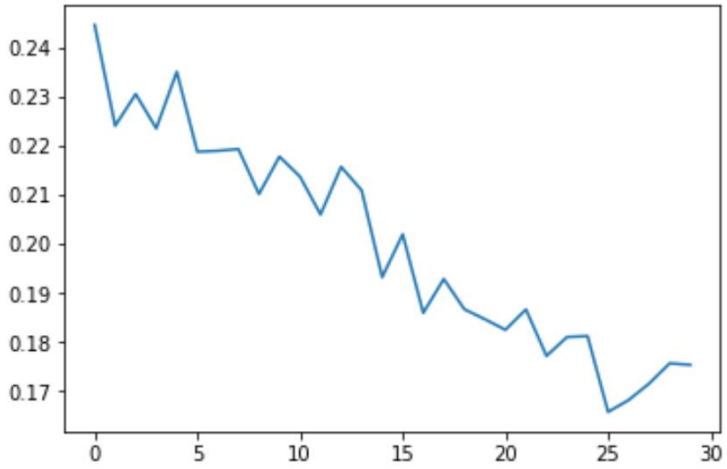
xgboost:

	precision	recall	f1-score	support
0	0.81	0.95	0.88	425
1	0.79	0.46	0.58	175
accuracy			0.81	600
macro avg	0.80	0.71	0.73	600
weighted avg	0.81	0.81	0.79	600

Figure 4.18

Bert

Accuracy on test data: 0.840062720501764



The comparison of different method's accuracy is shown in 4.19.

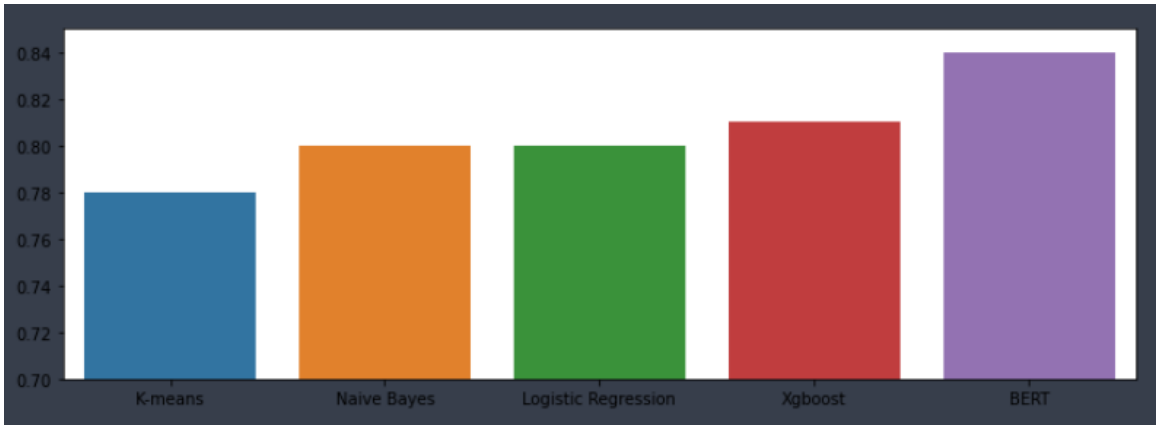


Figure 4.19

Discussion

In this project, we use two different model, TF-IDF & kNN model and Bert model to predict and classify the labels of tweets we crawl from Twitter. TF-IDF & kNN model calculates the TF-IDF of tweets in training dataset and testing dataset and use kNN algorithm to find k nearest neighbors and find the most common labels and labeled the tweets in testing dataset. This model is simple because it just compares the similarities between texts.

We can see that f1 score of misinformation tweets is low and f1 score of normal tweets is lower than before, too. And in Bert model, the accuracy of testing data is higher than other methods, but it's not high enough. One of the reasons may be the tweets we scrape from Twitter contain different kinds of contents which have no connection with COVID vaccine misinformation.

However, our models also perform not that bad in detecting COVID vaccine misinformation because all of the accuracies are about 80%.

One of the most essential factors that may affect the accuracy of our models is we can't judge statements in essence and classify whether statements are true or false. In order to judge statements in essence, we need to learn more about the misinformation sentence, the structure, the tone or the wording, these factors are useful to classify truth or false.

References

Lazer, David M., et al. “The Science of Fake News.” *Science*, vol. 359, no. 6380, 2018, pp. 1094–1096., https://doi.org/10.1126/science.aao2998.

World Health Organization. “How to Report Misinformation Online.” *World Health Organization*, World Health Organization, https://www.who.int/campaigns/connecting-the-world-to-combat-coronavirus/how-to-report-misinformation-online.

Marco-Franco, Julio Emilio, et al. “Covid-19, Fake News, and Vaccines: Should Regulation Be Implemented?” *International Journal of Environmental Research and Public Health*, vol. 18, no. 2, 2021, p. 744., https://doi.org/10.3390/ijerph18020744.

Shu, Kai, et al. “Fake News Detection on Social Media: A Data Mining Perspective.” *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 1, Sept. 2017, pp. 22–36., https://doi.org/10.1145/3137597.3137600.

Casero-Ripolles, Andreu. “Impact of COVID-19 on the Media System. Communicative and Democratic Consequences of News Consumption during the Outbreak.” *El Profesional De La información*, vol. 29, no. 2, Mar. 2020, https://doi.org/10.3145/epi.2020.mar.23.

Conte, Cristiano, et al. “Vaccines against Coronaviruses: The State of the Art.” *Vaccines*, vol. 8, no. 2, June 2020, p. 309., https://doi.org/10.3390/vaccines8020309.

Liu, Sixiao, and Haoran Chu. “Examining the Direct and Indirect Effects of Trust in Motivating COVID-19 Vaccine Uptake.” *Patient Education and Counseling*, Feb. 2022, https://doi.org/10.1016/j.pec.2022.02.009.

Wonodi, Chizoba, et al. "Conspiracy theories and misinformation about COVID-19 in Nigeria: Implications for vaccine demand generation communications." *Vaccine* 40.13 (2022): 2114-2121., https://doi.org/10.1016/j.vaccine.2022.02.005

Morales, Gabriela I, et al. “Exploring Vaccine Hesitancy Determinants during the COVID-19 Pandemic: An in-Depth Interview Study.” *SSM - Qualitative Research in Health*, vol. 2, Dec. 2022, https://doi.org/10.1016/j.ssmqr.2022.100045.

Laato, Samuli Laato, et al. “What Drives Unverified Information Sharing and Cyberchondria during the COVID-19 Pandemic?” *European Journal of Information Systems*, vol. 29, no. 3, June 2020, pp. 288–305., https://doi.org/10.1080/0960085X.2020.1770632.

Allington, Daniel, et al. "Media usage predicts intention to be vaccinated against SARS-CoV-2 in the US and the UK." *Vaccine* 39.18 (2021): 2595-2603., https://doi.org/10.1016/j.vaccine.2021.02.054

Zhang, Xichen, and Ali A. Ghorbani. “An Overview of Online Fake News: Characterization, Detection, and Discussion.” *Information Processing & Management*, vol. 57, no. 2, Mar. 2020, https://doi.org/10.1016/j.ipm.2019.03.004.

Bakir, Vian, and Andrew McStay. “Fake News and The Economy of Emotions.” *Digital Journalism*, vol. 6, no. 2, July 2017, pp. 154–175., https://doi.org/10.1080/21670811.2017.1345645.

Mosseri, Adam. “Addressing Hoaxes and Fake News.” *Meta*, 7 Nov. 2019, https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

Roth, Yoel, and Nick Pickles. “Updating Our Approach to Misleading Information.” *Twitter*, Twitter, 2020, https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.

Google. “Everyone, Everywhere, Benefits from a Healthy News Industry.” *Google*, Google, https://newsinitiative.withgoogle.com/.

Muric, Goran, et al. “Covid-19 Vaccine Hesitancy on Social Media: Building a Public Twitter Data Set of Antivaccine Content, Vaccine Misinformation, and Conspiracies.” *JMIR Public Health and Surveillance*, vol. 7, no. 11, 2021, https://doi.org/10.2196/30642.

Weinzierl, Maxwell A., and Sanda M. Harabagiu. “Automatic Detection of COVID-19 Vaccine Misinformation with Graph Link Prediction.” *Journal of Biomedical Informatics*, vol. 124, 2021, p. 103955., https://doi.org/10.1016/j.jbi.2021.103955.

Hu, Tao, et al. “Revealing Public Opinion towards Covid-19 Vaccines with Twitter Data in the United States: A Spatiotemporal Perspective.” 2021, https://doi.org/10.1101/2021.06.02.21258233.

Hayawi, K., et al. “ANTi-Vax: A Novel Twitter Dataset for COVID-19 Vaccine Misinformation Detection.” *Public Health*, vol. 203, 2022, pp. 23–30., https://doi.org/10.1016/j.puhe.2021.11.022.

Wang, Xin, et al. "Factors influencing fake news rebuttal acceptance during the COVID-19 pandemic and the moderating effect of cognitive ability." *Computers in human behavior* 130 (2022)., https://doi.org/10.1016/j.chb.2021.107174

Gupta, Ashish, et al. "Understanding patterns of COVID infodemic: A systematic and pragmatic approach to curb fake news." *Journal of business research* 140 (2022): 670-683., https://doi.org/10.1016/j.jbusres.2021.11.032

Jacob Devlin, et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". NAACL-HLT 2019. 4171–4186. https://arxiv.org/pdf/1810.04805.pdf

Sang-Woon Kim, Joon-Min Gil, Human-centric Computing and Information Sciences 9, Article number:30(2019):Research paper classification systems based on TF-IDF and LDA schemes

Yun-lei Cai, Duo Ji ,Dong-feng Cai: Proceedings of NTCIR-8 Workshop Meeting, June 15–18, 2010, Tokyo, Japan: A KNN Research Paper Classification Method Based on Shared Nearest Neighbor

Bin Yao, Feifei Li, Piyush Kumar, Computer Science Department, Florida State University, Tallahassee, FL, U.S.A: K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free

Rajnish Kumar, International Research Journal of Engineering and Technology (IRJET) Volume: 07 Issue: 12 | Dec 2020: Fake News Detection using Passive Aggressive and TF-IDF Vectorizer

Andres Corrada-Emmanuel, W. Bruce Croft, Vanessa Murdock, Center for Intelligent Information Retrieval Department of Computer Science University of Massachusetts Amherst Amherst, MA 01003-9264 :Answer Passage Retrieval for Question Answering