Covid-19 Vaccination Stance Detection Using Natural Language Processing and Machine-Learning Algorithms.

Harsh **Tita**[1], Rashi **Sharma**[1]

[1] *Amity University, Kolkata, India*

Abstract

The coronavirus outbreak has resulted in unprecedented measures, forcing authorities to make decisions related to establishing lockdowns in areas most affected by the pandemic. Social Media have supported people during this difficult time. On November 9, 2020, when the first vaccine with an efficacy rate over 90% was announced, social media reacted and people around the world began to express their feelings about this vaccination.

This paper aims to analyze the dynamics of opinion on COVID-19 vaccination, in which the civil society is highly manifested in the vaccination process. We compared classical machine learning algorithms to select the best performing classifier. 4,392 tweets were collected and analyzed. The proposed approach can help governments create and evaluate appropriate communication tools to provide clear and relevant information to the general public, increasing public confidence in vaccination campaigns.

Keywords

Twitter, COVID-19, stance classification, vaccine.

1. Introduction

The coronavirus outbreak caused by the novel coronavirus SARS-CoV-2 has brought a series of changes in many aspects of people's economic and social life. Since its occurrence, the coronavirus pandemic has continued to monopolize the different parts of the world, reaching 220 countries and territories by December 9, 2020 [1]. Governments have tried to address the outbreak by considering a series of measures, not all of them in accordance with the general public opinion. In all this time, the rapid growth of the number of cases globally has produced panic, fear and anxiety among people [2]. Due to the current situation generated by the lockdown in some parts of the world and social distancing in others, the use of social media globally has intensified [2], as it succeeds in connecting people from geographically different places and allows them to exchange ideas and information related to a series of aspects that have occurred in this period. Even more, people seem to rely on the information posted on social media. As a result, social media platforms have become mediator channels between each individual and the rest of the world and have gained more and more attention, being one of the fastest growing information systems for social applications [3], [4]. On this channel, individuals show their different views, opinions and emotions during the various events that occur due to the coronavirus pandemic [3].

Among some of the popular social media platforms, Twitter receives special attention. This is because users can easily disseminate information about their opinions on a particular topic through public messages called tweets [5]. In addition to the information voluntarily provided by the user, Tweets may also contain information about the user's location and may include links, emoticons, and hashtags that allow the user better express the emotions, making it a source of valuable information [5], [6]. Additionally, Twitter is used by government officials and politicians to inform the public about their activities and major events.[7].

Forum for Information Retrieval Evaluation, December 9-13, 2022, India

EMAIL: harshtita01@gmail.com (Harsh Tita), sharma.rashi2408@gmail.com

The issue of vaccination is one of the many issues that have raised many questions on social media, most of which relate to the safety of the overall process. Therefore, many studies have analysed the impact of various social media campaigns on resistance to vaccination [8], [9] or public sentiment related to the vaccination process [5], [10]. Furthermore, compared to other vaccination situations reviewed in the scientific literature, COVID-19 vaccination raises new questions related to the relatively short time span of vaccine development. It is well known that the process of developing a vaccine usually takes 10 years [11]. Note, however, that for the mumps vaccine, the fastest vaccine development before was 4 years [12], and nearly 40 years after the discovery of HIV, no effective vaccine has yet been developed. However, due to the state of emergency, the COVID-19 vaccination deadline has been shortened [11].
In this context, this paper analyses public opinion regarding the vaccination process in case of COVID 19, considering news posted on Twitter. Clean dataset was extracted, including 4392 tweets. The performance of multiple machine learning algorithms (both traditional and deep learning algorithms) was compared using annotated data sets. Best performing algorithms were selected and used to analyse the dataset. We collected and annotated the COVID-19 vaccination dataset, determined the best classifier for stance detection of COVID-19 vaccinations, and associated the number of tweets with stance (e.g., ProVax, Neutral and AntiVax).
Selected approaches can be easily integrated into systems that allow interested organizations to adequately monitor public opinion regarding the vaccination process in case of the novel coronavirus.

2. Methodology

The steps taken to analyze public opinion on COVID-19 vaccination from social media messages are stated below. The first step is to collect the COVID-19 vaccination stance dataset, which contains tweets in English. A randomly selected subset of this dataset was manually annotated as Neutral, ProVax or AntiVax to be used in the training phase of the pose classification algorithm. Due to its unstructured nature and informal writing style, tweets from the collected dataset were preprocessed in the next step to improve the performance of the pose classification algorithm.

In the current work, the performance of several classic machine learning algorithms was evaluated based on the following widely used metrics: accuracy, precision, recall and f-score. Accuracy is the ratio of correctly predicted observations to all observations and is defined as shown in (1). where TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives.

Thus, TP represents the number of real positive tweets classified as positive, FP is the number of real negative tweets classified incorrectly classified as positives, TN represents the number of negative tweets correctly classified as negative and FN is the number of real positive tweets incorrectly classified as negative.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}} \quad (1)$$

Precision, which represents the ratio of correctly predicted positive observations to the total predicted positive observations, is computed as shown in (2).

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FP}} \quad (2)$$

Recall, representing the ratio of correctly predicted positive observations to all the observations in the actual class, is computed as shown in (3).

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP+FN}} \quad (3)$$

Starting from Precision and Recall, the F-Score can be computed as a weighted average, as shown in (4).

$$\boldsymbol{F-Score} = \boldsymbol{2} * \frac{\boldsymbol{Precision*Recall}}{\boldsymbol{Precision+Recall}} \quad (4)$$

Finally, the best performing algorithm has been used to analyze the evolution of the public stance towards vaccination in the considered period.

2.1. Data Pre-Processing:

The main components of the stance detection process are the pre-processing, the feature extraction and the machine learning classification. The preprocessing step cleanses the text and the feature extraction transforms the raw text data into feature vectors.

We have performed various pre-processing steps on the dataset that mainly dealt with removing stop words. The text document is then converted into the lowercase for better generalization. Subsequently, the punctuations were cleaned and removed thereby reducing the unnecessary noise from the dataset. After that, we have also removed the stop words from the words along with removing the URLs as they do not have any significant importance.

At last, Lemmatization (reducing the derived words to their root form known as lemma) was performed for better results. Stop words are the words in any language which does not add much meaning to a sentence. They can safely be ignored without sacrificing the meaning of the sentence.

Countvectorizer: Machines cannot understand characters and words. So, when dealing with text data we need to represent it in numbers to be understood by the machine. Countvectorizer is a method to convert text to numerical data. The text is transformed to a sparse matrix. Countvectorizer makes it easy for text data to be used directly in machine learning and deep learning models such as text classification. We have used CountVectorizer for tuning the learning process and set its hyperparameters as the following:

- max_features:5000, which implies that top 5000 frequent words from the data is selected
- stop_words: an array of redundant words has been passed.

We used inbuilt functions mentioned below to train our model:

- train_test_split(): This function splits the dataset into a train and test set with a specified criteria of split, we started with a fraction setting of 0.2. This means we used 80% of our dataset for training our model and 20% for testing and evaluating our model.

- TfidfVectorizer() : Tf-idf is used to handle text data for machine learning purposes, it stands for term frequency — inverse document frequency and is represented by the formula below, using this function we convert all words into tf-idf scores.
  TF — IDF = TF (Term frequency) * IDF (Inverse document frequency)
  Term frequency — The number of times the term occurs in a given document.
  IDF — The number of documents in which the given term is found.

- make_pipeline() : This function is used for defining our data pipeline. In this we can apply a list of transforms, followed by a final estimator. We used Bernoulli for our case.

2.2. Learning Algorithms:

A machine learning approach has been used in order to accurately determine the stance towards vaccination in the collected tweets. Starting from the annotated dataset, the performance of several popular classification algorithms has been investigated: Bernoulli Naïve Bayes, Support vector machine (SVM), Multinomial logistic regression, Logistic Regression Machine Learning, The KNN classifier, Gradient Boosting.

1) Multinomial Naïve Bayes

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

2) Support Vector Machine (SVM)

Support Vector Machines (SVM) [99] are a family of supervised learning algorithms used for classification, regression and other tasks such as outlier detection. While other classification algorithms suffer from overfitting, one of the advantages of SVM is that they are less prone to this situation [100]. Another advantage resides in the fact that besides binary classification, multiclass classification can be performed by combining several binary classification functions. For this, each class is considered individually at a time, and for each class a classifier is searched that separates it from the other classes [101].

3) Bernoulli Naïve Bayes

This is used for discrete data and it works on Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So, when the feature values are binary, we know that we have to use Bernoulli Naive Bayes classifier.

4) K-Nearest Neighbor

The K-Nearest Neighbor or the KNN algorithm is a machine learning algorithm based on the supervised learning model. The K-NN algorithm works by assuming that similar things exist close to each other. Hence, the K-NN algorithm utilizes feature similarity between the new data points and the points in the training set (available cases) to predict the values of the new data points. In essence, the K-NN algorithm assigns a value to the latest data point based on how closely it resembles the points in the training set. K-NN algorithm finds application in both classification and regression problems but is mainly used for classification problems.

5) Logistic Regression

Logistic Regression Machine Learning is basically a classification algorithm that comes under the Supervised category (a type of machine learning in which machines are trained using "labelled" data, and on the basis of that trained data, the output is predicted) of Machine Learning algorithms. This simply means it fetches its roots to the field of Statistics. The main role of Logistic Regression in Machine Learning is predicting the output of a categorical dependent variable from a set of independent variables. In simple words, categorical dependent variable means a variable that is dichotomous or binary in nature having its data coded in the form of either 1 (stands for success/yes) or 0 (stands for failure/no).

6) Gradient Boosting Algorithm

It is a popular boosting algorithm. In gradient boosting, each predictor corrects its predecessor's error. The weights of the training instances are not tweaked, instead, each predictor is trained using the residual errors of predecessor as labels. Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

2.3. Approach Used:

Start Program

```
mn= [KNN( ), SVM(), LogisticRegression(), GradientBoostClassifier(), BernoulliNB(),
MultinomialNB()] ;
for (i=0; i<6; i++)
  Model= mn[i];
  Model.fit();
  Model.predict();
  print(Accuracy(i), confusion_matrix, classification_report);
end loop
```

End Program

3. Results

Let's view the performance of each of the machine learning algorithms/classifications by representing evaluation metrics such as recall, precision and f1-score.

Class	Precision	Recall	F-Score
AntiVax	0.43	0.77	0.55
Neutral	0.59	0.58	0.58
ProVax	0.68	0.34	0.46

Table 1

Table-1 includes the results achieved using K- Nearest Neighbor classifier.

	AntiVax	Neutral	ProVax
AntiVax	173	30	21
Neutral	95	177	35
ProVax	135	94	119

Confusion matrix for the KNN classifier is as shown in the above figure.

Class	Precision	Recall	F-Score
AntiVax	0.70	0.55	0.61
Neutral	0.58	0.73	0.65
ProVax	0.61	0.55	0.58

Table 2

Table-2 includes the results achieved using Gradient Boosting Classifier.

	AntiVax	Neutral	ProVax
AntiVax	123	45	56
Neutral	15	224	68
ProVax	38	117	193

Confusion matrix for the Gradient Boosting Classifier is as shown in the above figure.

Class	Precision	Recall	F-Score
AntiVax	0.70	0.57	0.63
Neutral	0.62	0.72	0.66
ProVax	0.64	0.62	0.63

Table 3

Table-3 includes the results achieved using Logistic Regression.

	AntiVax	Neutral	ProVax
AntiVax	127	44	53
Neutral	16	221	70
ProVax	39	93	216

Confusion matrix for the Logistic Regression is as shown in the above figure.

Class	Precision	Recall	F-Score
AntiVax	0.73	0.54	0.62
Neutral	0.62	0.73	0.67
ProVax	0.63	0.64	0.63

Table 4

Table-4 includes the results achieved using Support Vector Machine.

	AntiVax	Neutral	ProVax
AntiVax	122	40	62
Neutral	14	225	68
ProVax	31	96	221

Confusion matrix for the Support Vector Machine Classifier is as shown in the above figure.

Class	Precision	Recall	F-Score
AntiVax	0.70	0.49	0.58
Neutral	0.66	0.68	0.67
ProVax	0.59	0.68	0.63

Table 5

Table-5 includes the results achieved using Multinomial Naïve Bayes.

	AntiVax	Neutral	ProVax
AntiVax	110	33	81
Neutral	13	209	85
ProVax	35	76	237

Confusion matrix for the Multinomial Naïve Bayes is as shown in the above figure.

Class	Precision	Recall	F-Score
AntiVax	0.59	0.58	0.58
Neutral	0.58	0.69	0.63
ProVax	0.66	0.54	0.59

Table 6

Table-6 includes the results achieved using Bernoulli Naïve Bayes.

	AntiVax	Neutral	ProVax
AntiVax	130	53	41
Neutral	36	213	58
ProVax	56	103	189

Confusion matrix for the Logistic Regression is as shown in the above figure.

After applying various Machine Learning Algorithms on the Training data-set we got accuracies as mentioned below in table-7.

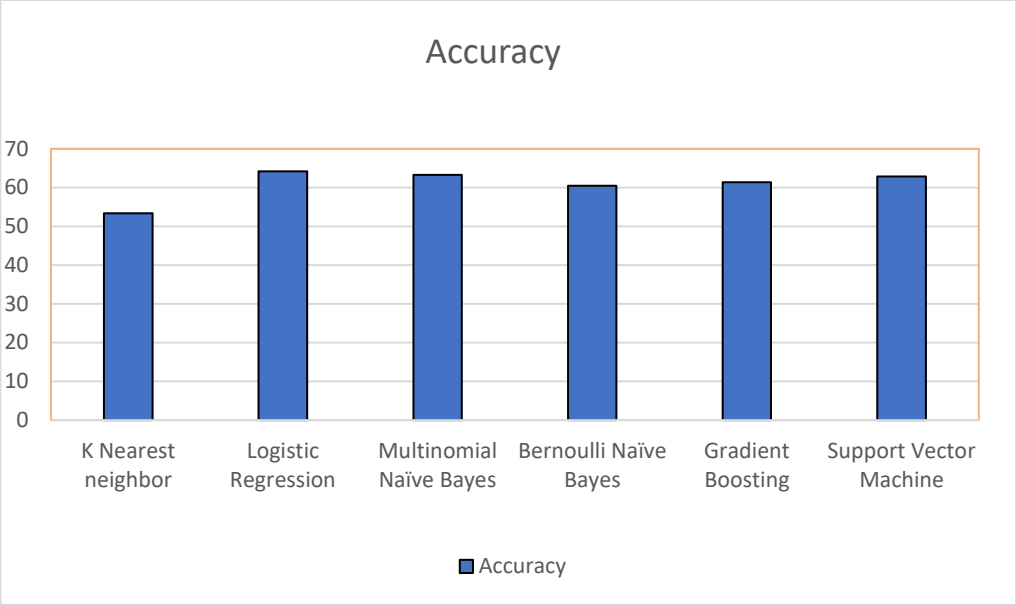
Algorithm	Accuracy
K-Nearest Neighbor	53.4%
Gradient Boosting Classifier	61.4%
Logistic Regression	64.2%
Support Vector Machine	62.9%
Multinomial Naïve Bayes	63.3%
Bernoulli Naïve Bayes	60.5%

Table 7

Algorithm	Accuracy	F1 - Score
Bernoulli Naïve Bayes	49%	0.473
Support Vector Machine	48.7%	0.471
Logistic Regression	47.7%	0.469
Multinomial Naïve Bayes	46.6%	0.458
K-Nearest Neighbor	44.4%	0.432
Gradient Boosting Classifier	42.5%	0.392

Table 8

Below is the pictorial representation of the accuracies obtained by each machine learning classifier:



4. Conclusion

In the current study, the initial announcement of a coronavirus vaccine and the first real vaccination process initiated outside of limited clinical trials were analyzed using machine learning-based stance detection. Several classical machine learning and deep learning algorithms were compared, and the best performing classifier was selected based on the performance metrics. The proposed approach used Bernoulli Naïve Bayes with an accuracy of 49% to classify tweets into three main classes: ProVax, AntiVax, and Neutral regarding the COVID-19 vaccination. The purpose of this paper was to monitor changes in the stance towards COVID-19 vaccination through tweets.

With many countries around the world planning to initiate vaccination processes for COVID-19, early detection of changes in opinion can be very useful and help government decision makers to take steps to curb infections. This can be very helpful as it allows us to drive targeted actions. Possible future research directions include the development of better performing stance classification algorithms, as well as extending the analyzed period, especially given the fact that the vaccination process is expected to take a relatively long period of time.

5. References

[1] Worldometer. (Dec. 9, 2020). Coronavirus Update (Live): 63,777,845 Cases and 1,477,777 Deaths From COVID-19 Virus Pandemic. Accessed: Dec. 9, 2020. [Online]. Available: https://www.worldometers.info/ coronavirus/

[2] K. Chakraborty, S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, ‘‘Sentiment analysis of COVID-19 tweets by deep learning classifiers—A study to show how popularity is affecting accuracy in social media,’’ Appl. Soft Comput., vol. 97, Dec. 2020, Art. no. 106754, doi: 10.1016/j.asoc.2020.106754.

[3] A. H. Alamoodi, B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, ‘‘Sentiment analysis and its

applications in fighting COVID-19 and infectious diseases: A systematic review,'' Expert Syst. Appl., vol. 167, Apr. 2021, Art. no. 114155, doi: 10.1016/j.eswa.2020.114155.

[4] G. Appel, L. Grewal, R. Hadi, and A. T. Stephen, ''The future of social media in marketing,'' J. Acad. Marketing Sci., vol. 48, no. 1, pp. 79–95, Jan. 2020, doi: 10.1007/s11747-019-00695-1.

[5] E. D'Andrea, P. Ducange, A. Bechini, A. Renda, and F. Marcelloni, ''Monitoring the public opinion about the vaccination topic from tweets analysis,'' Expert Syst. Appl., vol. 116, pp. 209–226, Feb. 2019, doi: 10.1016/j.eswa.2018.09.009.

[6] A. Giachanou and F. Crestani, ''Like it or not: A survey of Twitter sentiment analysis methods,'' ACM Comput. Surv., vol. 49, no. 2, Nov. 2016, Art. no. 28, doi: 10.1145/2938640.

[7] J. Golbeck, J. M. Grimes, and A. Rogers, ''Twitter use by the U.S. Congress,'' J. Amer. Soc. Inf. Sci. Technol., vol. 61, no. 8, pp. 1612–1621, May 2010, doi: 10.1002/asi.21344.

[8] E. A. Pedersen, L. H. Loft, S. U. Jacobsen, B. Søborg, and J. Bigaard, ''Strategic health communication on social media: Insights from a Danish social media campaign to address HPV vaccination hesitancy,'' Vaccine, vol. 38, no. 31, pp. 4909–4915, Jun. 2020, doi: 10.1016/j.vaccine.2020.05.061.

[9] K. Dedominicis, A. M. Buttenheim, A. C. Howa, P. L. Delamater, D. Salmon, S. B. Omer, and N. P. Klein, ''Shouting at each other into the void: A linguistic network analysis of vaccine hesitance and support in online discourse regarding California law SB277,'' Social Sci. Med., vol. 266, Dec. 2020, Art. no. 113216, doi: 10.1016/j.socscimed.2020.113216.

[10] S. Martin, E. Kilich, S. Dada, P. E. Kummervold, C. Denny, P. Paterson, and H. J. Larson, ''Vaccines for pregnant women· · · ?! Absurd'— Mapping maternal vaccination discourse and stance on social media over six months,'' Vaccine, vol. 38, no. 42, pp. 6627–6637, Sep. 2020, doi: 10.1016/j.vaccine.2020.07.072.

[11] T. T. Le, Z. Andreadakis, A. Kumar, R. G. Román, S. Tollefsen, M. Saville, and S. Mayhew, ''The COVID 19 vaccine development landscape,'' Nature Rev. Drug Discovery, vol. 19, no. 5, pp. 305–306, Apr. 2020, doi: 10.1038/d41573-020-00073-5.

[12] J. F. Modlin, W. A. Orenstein, and A. D. Brandling-Bennett, ''Current status of mumps in the united-states,'' J. Infectious Diseases, vol. 132, no. 1, pp. 106–109, Jul. 1975.