A Machine Learning Approach for COVID-19 Tweet Classification

Subinay Adhikary[1]

[1]Indian Institute of Science Education and Research Kolkata,Campus Road, Mohanpur, West Bengal 741246

Abstract

Vaccine-related information is awash on social media platforms like Twitter and Facebook. One party supports vaccination, while the other opposes vaccination and promotes misconceptions and misleading information about the risks of vaccination. The analysis of social media posts can give significant information into public opinion on vaccines, which can help government authorities in decision-making.This paper describes the dataset used in the shared task, and compares the performance of different classification that are provax, antivax and last neutral for identifying effective tweets related to Covid vaccines.We experimented with a classification-based approach. Our experiment shows that SVM classification performs well in order to effiective post.We're going to do this because vaccination is an important step for Covid19 so people can easily fix the news about the vaccine and grab their own slot and symptom detection is also playing a important part to arrest the spread of disease.

Keywords

Covid-19, Twitter, Vaccination, Classification

1. Introduction

In the face of a COVID-19 outbreak that shows no signs of slowing down, vaccination seems to be the only possible solution. Around the globe, people began to share their thoughts about vaccination. In spite of this, many people are sceptical about vaccinations for a variety of reasons. Social media sites like Twitter and Facebook are inundated with vaccine-related information[1]. One group of individuals is in favour of vaccination, while another opposes vaccination and spreads myths and false information about the dangers of vaccination. Using social media posts to study can provide useful insights.[2][3]

It has been identified that Covid-19 spreads very fast. Therefore, it is reacquired to identify the symptoms of Covid-19 as soon as possible. By identifying Covid symptoms based on their classification, we can identify them more effectively.

Here, we have used SVM, which is a very simple and efficient classifier algorithm that is widely used for pattern recognition and has a very good classification performance compared to other classifiers.The proposed model is validated with the tasks of IRMiDis FIRE-2022.

Forum for Information Retrieval Evaluation 2022 (FIRE 2022)

subinayadhikary612@gmail.com (S. Adhikary)

2. Tasks

- **Task 1: COVID-19 vaccine stance classification from tweets**
  The Covid-19 vaccination is of utmost importance to prevent the disease.Here the task is to build a effective classifier to understand the user's stance in social media.The three classes are as follows:

  1. **AntiVax** - the tweet indicates hesitancy (of the user who posted the tweet) towards the use of vaccines.
  2. **ProVax** - the tweet supports / promotes the use of vaccines.
  3. **Neutral** - the tweet does not have any discernible sentiment expressed towards vaccines or is not related to vaccines

- **Task 2: Detection of COVID-19 symptom-reporting in tweets**
  It is important for authorities to identify people who are experiencing COVID-19 symptoms as quickly as possible in order to prevent the spread of the disease. The purpose of this task is to detect tweets that relate to COVID-19 symptoms (e.g., 'fever', 'cough'). Such tweets are referred to as symptom-reporting tweets.Build an effective classifier for 4-class classification on tweets that can detect tweets that report someone experiencing COVID-19 symptoms. The 4 classes are described below:

  1. **Primary Reporting** - The user (who posted the tweet) is reporting symptoms of himself/herself.
  2. **Secondary Reporting** - The user is reporting symptoms of some friend / relative / neighbour / someone they met.
  3. **Third-party Reporting** - The user is reporting symptoms of some celebrity / third-party person.
  4. **Non-Reporting** - The user is not reporting anyone experiencing COVID-19 symptoms, but talking about symptom-words in some other context. This class includes tweets that only give general information about COVID-19 symptoms, without specifically reporting about a person experiencing such symptoms.

3. Dataset

- For the Task1 we got 4392 tweets that were labelled with 3 classes.
- For the Task2, there was 1574 tweets that were labelled with 4 classes.

.

4. Our Approach

We discuss the overall design and implementation of our approach in this section. We borrow the advantages and capabilities of machine learning algorithms to implement the two abovementioned tasks which are properly discussed in the consequent subsections.

4.1. Task1

COVID-19 vaccine stance classification from tweets :
The purpose of this task is to identify the public sentiments toward vaccines based on Twitter data. Model selection, vectorization, and preprocessing are the three steps of the overall process.

4.1.1. Preprocessing

This phase involves cleaning up of the provided tweets labeled as AntiVax, Provax or Neutral in the training dataset.All the words starting with hashtags or containing multiple spaces, special characters,urls,punctuations or stopwords are firstly trimmed from every tweet.

4.1.2. Vectorization

After streaming and lemmatization are applied,the pre processed level tweets are then transform to numeric feature vector using term frequency inverse document frequency. After transforming the unstructured tweet data into numeric structured data.

4.1.3. Model Selection

This numeric structured data was fed into two classifiers: Naive Bayes and SVM(Support Vector Machine). In comparison with these two classifiers, SVM performed better. In this training set, 1676 provax tweets and 1081 antivax tweets are used as training data, and 1635 neutral tweets are used for training and get traing accuracy 0.65.

	precision	recall	f1-score	support
0	0.67	0.72	0.69	432
1	0.62	0.61	0.61	407
2	0.66	0.59	0.62	259
accuracy			0.65	1098
macro avg	0.65	0.64	0.64	1098
weighted avg	0.65	0.65	0.65	1098

Table 1: Task1 classification result

4.2. Task2

Detection of COVID-19 symptom-reporting in tweets
IThe purpose of this task is to identify the symptoms toward Covid-19 based on Twitter data. Model selection, vectorization, and preprocessing are the three steps of the overall process.

4.2.1. Preprocessing

This phase involves cleaning up of the provided tweets labeled as AntiVax, Provax or Neutral in the training dataset.All the words starting with hashtags or containing multiple spaces, special characters,urls,punctuations or stopwords are firstly trimmed from every tweet.

4.2.2. Vectorization

After streaming and lemmatization are applied,the pre processed level tweets are then transform to numeric feature vector using term frequency inverse document frequency. After transforming the unstructured tweet data into numeric structured data.

4.2.3. Model Selection

This numeric structured data was fed into two classifiers: Naive Bayes and SVM(Support Vector Machine). In comparison with these two classifiers, SVM performed better.In this training set containing 437 primary reporting tweets and 127 Secondary Reporting tweets and 196 Third-Party reporting tweets and 814 tweets are Non-Reporting and get training accuracy 0.69.

	precision	recall	f1-score	support
0	0.70	0.91	0.79	194
1	0.74	0.27	0.39	64
2	0.67	0.67	0.67	107
3	0.70	0.24	0.36	29
accuracy			0.69	394
macro avg	0.70	0.52	0.55	394
weighted avg	0.70	0.69	0.66	394

Table 2: Task2 classification result

5. Evaluation

Task 1 - IRMiDis Track results are evaluated using overall accuracy and the macro-F1 score on the three classes as metrics. The result of our submitted automated run for Task 1 is shown in Table 1.

Team ID	File name	Accuracy	macro F1-Score
Subinay-IISERK	Vax_labels.csv	0.275	0.281

Task 2 - IRMiDis Track results are evaluated using overall accuracy and the macro-F1 score on the four classes as metrics. The result of our submitted automated run for Task 2 is shown in Table 2.

Team ID	File name	Accuracy	macro F1-Score
Subinay-IISERK	Symptoms_labels.csv	0.295	0.324

6. Conclusion

Natural Language Processing was used for this work submitted to the IRMiDis Track processing of the tweets. Machine Learning models were used to identify tweets in the training data . We used the NeatText package of NLP for cleaning our tweets. Then techniques like Tokenization and Lemmatization were used to pre-process the tweets. Later, we used a few Machine Learning models to train our dataset among which the SVM(Support Vector Machine) model gave the highest F-score.

References

[1] E. D’Andrea, P. Ducange, A. Bechini, A. Renda, F. Marcelloni, Monitoring the public opinion about the vaccination topic from tweets analysis, Expert Systems with Applications 116 (2019) 209–226.

[2] L.-A. Cotfas, C. Delcea, I. Roxin, C. Ioanăş, D. S. Gherai, F. Tajariol, The longest month: analyzing covid-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement, Ieee Access 9 (2021) 33203–33223.

[3] M. M. Müller, M. Salathé, Crowdbreaks: tracking health trends using public social media data and crowdsourcing, Frontiers in public health 7 (2019) 81.