Need for Vision: A data-centric approach towards analysing impact of COVID-19.

Kaustav Das[1,*,†]

[1]*Amity School of Engineering and Technology (Amity University Kolkata)*

Abstract

From the beginning of 2020, we saw a rise of a new virus called the Coronavirus and ultimately a pandemic that anyone reading this paper must have been through. With the rise of COVID,many vaccines were found, the global vaccination drive as a result of this naturally fueled a possibility of Pro-Vaxxers and Anti-Vaxxers strongly expressing their support and concerns regarding the vaccines on social media platforms and along with this came up the need of quick identification of people who are experiencing COVID-19 symptoms. So in this paper, an effort has been made to facilitate the understanding of all these complications and help the concerned authorities. With the help of data in the form of Covid-19 tweets, a (machine-learning) classifier has been built which can classify users as per their vaccine related stance and also classify users who have reported their symptoms through tweets.

Keywords

Covid Tweets, Natural Language processing, Vaccine Stance, Covid Symptoms Report, Classification

1. Introduction

Globally, as of 6:28pm CEST, 7 October 2022, there have been 617,597,680 confirmed cases of COVID-19, including 6,532,705 deaths, reported to WHO. Fortunately, since December 2020 / January 2021, multiple pharmaceutical companies have put forward vaccines (e.g., AstraZenca, Pfizer, Moderna ,Covishield to name a few) that are claimed to reduce the chance of COVID infection and fatality. Naturally, governments across the world are procuring and administering these vaccines to their citizens. But with the rise in these vaccination drives there has been increase in vaccine hesitancy and anti-vaccination speeches all over the world, undermining the efforts to control the spread of the novel coronavirus. So, it is quite evident that there is a need for us and especially the authorities to analyse the societal angle i.e. the public sentiments towards the vaccines. A major motivation for this approach would be [1]

The hesitation towards the vaccine can come from political opinions, conspiracy theories against the government and just general skepticism. In this paper,an effort has been made to identify the user's stance based on Covid tweets crawled from social media platforms.The debate in vaccine stance can be traced long before the onset of Covid itself. The debate has such has been maintained through the active discourse of certain section of people primarily through social media labelled as the "Anti-Vaxxers" and the "Pro-Vaxxers",it is evident that Anti-Vaxxers are the ones who are against the administration of vaccines and the Pro-Vaxxers are the ones who

kaustav.das1@s.amity.edu (K. Das)

CEUR Workshop Proceedings (CEUR-WS.org)

support the administration of the vaccine to all of the population. There is another section of people who have maintained neutrality in their views regarding the vaccine and hence can be labelled as "Neutral".
Another matter of importance is the rapid identification of people who are experiencing COVID-19 symptoms, because it is extremely necessary for authorities to arrest the spread of the disease.So for such purposes,we specifically explore if tweets that report about someone experiencing COVID-19 symptoms (e.g., 'fever', 'cough') can be automatically identified.We call such tweets symptom-reporting tweets.
For both the purposes mentioned above,we built and train a machine learning classifier to classify the tweets into their respective classes to a certain precision.

- `For the 1st task of identifying vaccine stance:`
  We build a classifier for 3-class classification on tweets with respect to the stance reflected towards COVID-19 vaccines. The 3 classes are described below:
  - `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
- `For the 2nd task of detection of reporting symptoms:`
  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.

2. Data Collection

The data can be credited to [2]

1. **Data for vaccine stance analysis**: We crawled tweets between March-December 2020 with various vaccine-related keywords. We got tweets annotated with the three labels by three crowd-workers. For 1600 tweets, there was at least majority agreement among the crowd-workers. These 1600 tweets (tweet IDs, tweet texts, classes) has been used for training the machine learning classifier.

2. **Data for reporting symptoms identification**:We have crawled English tweets from February 2020 - June 2021 using keywords related to COVID-19 symptoms (e.g., ‘fever’, ‘cough’). We took a random sample from our collected set of tweets and got about 2K tweets annotated into the four classes by human workers.

3. Data Pre-processing

As the data collected is primarily tweets, in both the cases the data needs to be cleaned, transformed and in some cases even further sampling was needed to produce fair results. In regards to both the tasks, we followed same steps to clean the data.

3.1. Data-Cleaning

Following steps were taken to clean the data:

1. The column representing tweet ids were dropped as it was of no consequence after the data had been crawled.
2. The user handles were removed from the tweet using the ntx or the neattext library.
3. The urls or hyperlinks were removed as they did not provide any form of insight, from the tweet using the ntx or the neattext library.
4. The special characters for example: currency symbols,hashtags (#),percentages e.t.c were removed for further refinement of the data using the neattext library.

3.2. Natural language Processing

After the tweets were cleaned, some natural language processing was done on the tweets:

1. The emojis were also removed from all the individual tweets and replaced with their encoded meaning using demojize method of emoji library of python.
2. And finally contractions used in natural language like "I’ll" were fixed to become more meaningful like "I will" using the contractions library of python.
3. The tweets were all changed to lower case to maintain more uniformity during the classification task.
4. Lemmatization was performed on the tweets to convert each word to their base forms, further enhancing the classification job by preserving the context of the tweets, using the nltk or the natural language toolkit in python.
5. Using the nltk library, stopwords like "the","is" e.t.c were removed as they do not provide or preserve any context in the tweet.

Cleaned Data		
Task	Original Tweet	Cleaned Tweet
Vaccine Stance	@NikkiHaley @pfizer @realDonaldTrump Nothing to do with Trump. Delete tweet, you look foolish.	nothing trump delete tweet look foolish
Symptom Report	I used to go "no no, the cough is not because of my smoking, it's just a mild cold". Now I go "don't worry about my cough! It's because I'm a smoker!" #Covid_19 #smoking	i used go cough smoking mild cold now i go worry cough it i smoker covid19 smoking

Table 1: Original tweet vs Cleaned Tweet(with NLP)

3.3. Checking if the data is balanced or not

1. For vaccine stance task: The data provided was balanced for each of the 3 classes and needed no under-sampling or synthetic up sampling, to be used without a particular bias forming towards any of the classes.
2. For reporting symptoms classification:The data provided was quite imbalanced for each of the 4 classes. The 4 classes had such value distributions.
   a) non-reporting - 814 tweets
   b) primary - 437 tweets
   c) third-party - 196 tweets
   d) secondary - 127 tweets

   So to avoid any unnecessary bias in the classification task, we under-sampled the data to 127 tweets each class, which ensured better results and a 25 percent distribution for each class.

4. Building a tweet classifier

After the data has been processed, it's time we build a classifier catered to each task. To build a classifier we must first vectorize, i.e. convert the text data to a more machine learning applicable numerical format. The choice of model is the crucial step and depending on that hyper parameter tuning for the complete machine learning model can be done.

4.1. For vaccine stance classification:

4.1.1. Classifier models:

To classify tweets into the 3 aforementioned classes, we implemented various classification models like Naive Bayes, XGBoost along with TF-IDF vectorizers, Count Vectorizers, cosine similarity and word2vec models. Among all these models we specifically focused on XGBoost for it produced good results on training data and some other techniques to further supplement its performance. The models considered are listed below:

1. XGboost with count vectorizer and word2vec model
2. XGboost with count vectorizer and cosine similarity
3. XGboost with count vectorizer

Along with the machine learning models listed some other techniques including a word2vec model was trained on the given corpus of data. Cosine similarity was also used to further supplement the XGBoost classifier, which was referenced from this article [3].

4.1.2. Hyper parameter Optimization

: To fine tune the models and thus produce better performance, we used some hyper-parameter tuning techniques such as :

1. Random Search CV
2. Bayesian Optimisation with Hyper Opt

Bayesian Optimisation gave the most satisfactory results out of these 2 techniques and these were the optimized parameters obtained as per the task:

1. 'colsample_bytree': 0.7281704443843204
2. 'gamma': 0.24728061277513913
3. 'learning_rate': 0.3994722661863012
4. 'max_depth': 5.0
5. 'min_child_weight': 0.0
6. 'reg_alpha': 0.0
7. 'reg_lambda': 0.6148225282687034
8. 'objective': 'multi:softproba'

These were the exact parameters that were set to the model for the most fine-tuned classification.

4.1.3. Evaluation of tweet classification:

An 80-20 split was made in the data, 80% being training data and the rest 20% being the testing data. All evaluations were made based on Macro-F1 score. Finally after being tested, it was time to test the models on the actual test data collected. These were the results:

	Accuracy	macro-F1 score
XGBoost with Word2vec	0.46	0.461
XGBoost with Cosine similarity	0.497	0.490
XGBoost with Count Vectorizer	0.506	0.490

Table 2: Performance of XGBoost Classifiers on Vaccine Stance

So, from the above evaluation it is very clear that the best model was the one where XGBoost was used with Count Vectorizer (as per the macro f1 score). The reason for which the other two models couldn't perform well are yet to be found, although some outliers or some mislabelled tweet may have resulted to the decrease in their performance. In all cases Count Vectorizer, also proved to be much more efficient when it came to classifying ProVax or AntiVax tweets.

4.2. For symptom-reporting classifier:

4.2.1. Classifier models:

To classify the tweets into 4 different classes we have yet again focused on XGBoost as our primary classifier. Along with Xgboost we have used Count Vectorizer but this time we have focused on a range of words rather than individual words for better detection of symptoms.

4.2.2. Evaluation of models:

No hyper parameter tuning was performed on this model so we jump to the results part of the model. The model after being trained and tested on the training data, was tested on the actual test data which shows:

	Accuracy	macro-F1 score
XGBoost with Count vectorizer	0.507	0.428

Table 3: Performance of XGBoost Classifiers on Symptoms Reporting

5. Conclusion

The work done here is purely based on machine learning classifiers, through this we have attempted to make an efficient classifier using the XGBoost machine learning algorithm to classify tweets with a good macro - F1 score and also explore other similar text based classifiers which could be applicable to our context. **Future work:** In future I plan to explore problems like this with more sophisticated and state of the art deep-learning based classifiers like BERT, neural nets , etc.

Acknowledgments

Thanks to https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook with the hyper parameter tunning of the XGBoost model and https://machinelearningmastery.com/ for being a general guide throughout the process.

References

[1] S. Poddar, M. Mondal, J. Misra, N. Ganguly, S. Ghosh, Winds of change: Impact of covid-19 on vaccine-related opinions of twitter users, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 16, 2022, pp. 782–793.

[2] S. Whiting, I. A. Klampanos, J. M. Jose, Temporal pseudo-relevance feedback in microblog retrieval, in: European Conference on Information Retrieval, Springer, 2012, pp. 522–526.

[3] K. Park, J. S. Hong, W. Kim, A methodology combining cosine similarity with classifier for text classification, Applied Artificial Intelligence 34 (2020) 396–411.