Stance Detection on COVID-19 vaccination tweets

Drumil Shah 202018009@daiict.ac.in
Sambhav Gulla 202018018@daiict.ac.in
Himani Desai 202018030@daiict.ac.in
Namra Jain 202018059@daiict.ac.in

Department - MSc. Data Science
College - DAIICT, Gandhinagar, Gujarat.

Abstract **- The coronavirus outbreak has brought unprecedented measures, which forced the authorities to make decisions related to the instauration of lockdowns in the areas most hit by the pandemic. Social media has been an important support for people while passing through this difficult period. Stance detection is a subproblem of sentiment analysis where the stance of the author of a piece of natural language text for a particular target (either explicitly stated in the text or not) is explored. The stance output is usually given as Favor, Against, or Neither. The dataset that we have referred to was collected tweets between November 9, 2020 and December 8, 2020. It contains a balanced dataset with 3249 tweets annotated in the categories "against" (0), "neutral" (1) and "in favor" (2). In this paper, we target stance detection on corona-related tweets and present the performance results of our SVM-based stance classifiers on such tweets. Next, we evaluate SVM classifiers using different feature sets for stance detection on this data set. The employed features are based on n-grams. We have also used the LSTM model for the same purpose which gives better results than the SVM classifier.**

Keywords **- stance, stance detection, tweets, COVID-19, SVM, n-grams, Support Vector Machine, tokenizer, LSTM, RNN.**

INTRODUCTION

SARS-CoV-2 and the resulting COVID-19 disease is one of the biggest challenges of the 21st century. At the time of this publication, about 53 million people have tested positive and 2.8 million people have died as a result [1]. W.H.O declared this outbreak as a public health emergency [2] and mentioned the following; the virus is being transmitted via the respiratory tract when a healthy person comes in contact with the infected person. The virus may transmit between persons through other roots which are currently unclear.During these difficult times, people have taken to social media to discuss their fears, opinions, and insights on the global pandemic. In this context, the present paper analyzes the public opinion related to the vaccination process in the case of COVID-19, by considering the messages posted on Twitter. The period between November 9, 2020 – when Pfizer and BioNTech announced the development of a vaccine that is more than 90% effective, to December 8, 2020 – when the vaccination process has started in the UK, has been considered. A number of 2770 tweets have been collected and cleaned[3]. The performance for stance detection of two machine learning algorithms (both classical machine learning and deep learning algorithms) has been compared on an annotated dataset. The chosen approach can be easily integrated in a system which can allow interested organizations a proper monitoring of

the public opinion regarding the vaccination process in the case of the new coronavirus

DATASET

Dataset used here is tweets made by different people on twitter which was labelled data and it is available at https://github.com/liviucotfas/covid-19-vaccination-stance-detection. The data available was in an unhydrated form (i.e. Tweet ID with their corresponding labels) so, we hydrated the data and obtained several information like: text, url, created_at, hashtags, etc in a csv file. Then we modified the csv file for only required data like id, text and category.

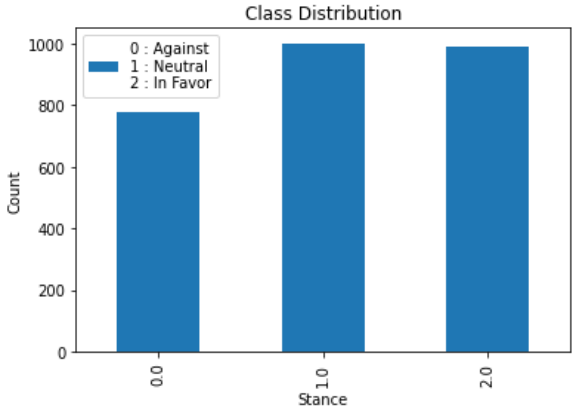
Then we cleaned the text (tweets) from the dataset by:

1. Removing punctuation marks, special characters, hashtags, urls and numerical values.
2. Removing stopwords
3. Performing Lemmatization on each text.

An ***n*-gram** is a contiguous sequence of *n* items from a given sample of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The *n*-grams typically are collected from a text or speech corpus.

n-grams are used for efficient approximate matching. By converting a sequence of items to a set of *n*-grams, it can be embedded in a vector space, thus allowing the sequence to be compared to other sequences in an efficient manner. Here we have used Bi-gram and words for analysis.

The data is then splitted into training data and testing data with a ratio of 8:2 using train_test_split module of sklearn.models. Here the dataset is labelled for covid-vaccination as a stance target and so labels 0, 1 and 2 depict Against, neutral and in favour respectively.

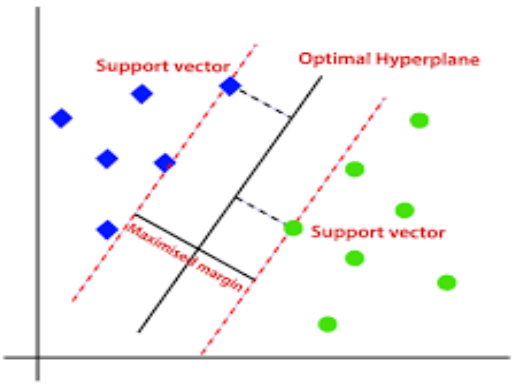


MODELS

We have implemented two different models with the same dataset to identify a better model among them for stance detection.

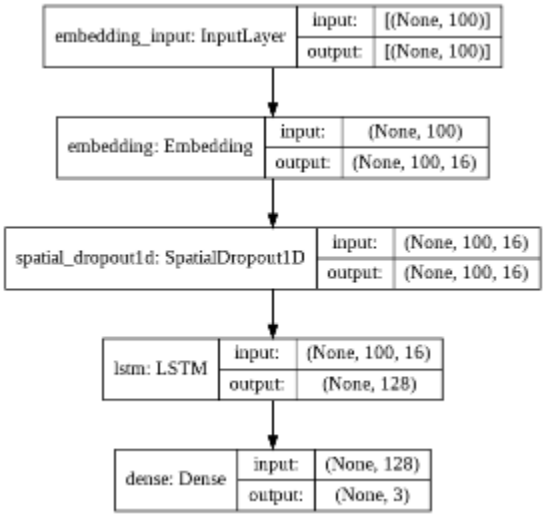
1. SVM

In machine learning, **support-vector machines (SVM)** are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis[5]. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N is the number of features) that distinctly classifies the data points. We have used n-gram feature extraction technique and provided the data for training the SVM model.



2. LSTM

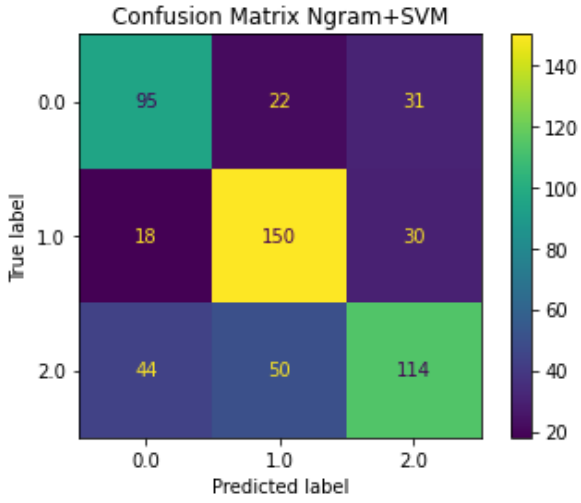
Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. LSTMs are explicitly designed to avoid the long-term dependency problem[6]. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn. We have performed tokenization on the textual tweets for providing them to train the model using tokenization from keras.preprocessing.text.



RESULTS

❖ SVM

- Accuracy - 67.87%
- **Confusion matrix:**

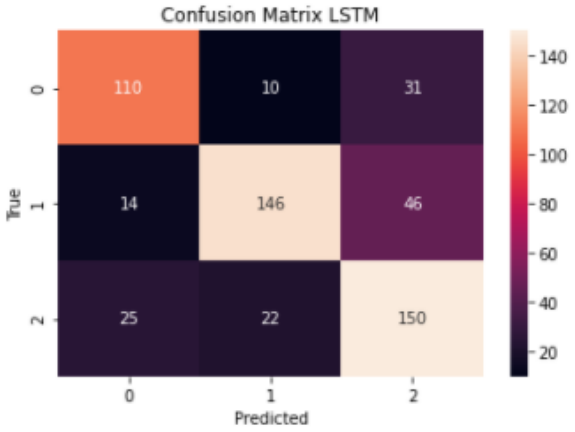


❖ **Performance matrix**

	precision	recall	f1-score
0.0	0.61	0.64	0.62
1.0	0.68	0.76	0.71
2.0	0.65	0.55	0.60

❖ LSTM

- Accuracy - 73%
- **Confusion matrix:**



- **Performance matrix**

	precision	recall	f1-score
0.0	0.74	0.73	0.73
1.0	0.82	0.71	0.76
2.0	0.66	0.76	0.71

CONCLUSION

Social media(Twitter) allows people to not only share their opinions but also express their feelings. A widely used social media platform is Twitter.

Here we have selected 2770 tweets which were processed and analyzed. The implementation was done through several steps starting from pre-processing to getting the optimal classification model. This work has performed lemmatization steps to get the best result and overcome the problems of converting words to its right root. So, this approach is the best and delivers the best result.This model works and will train with 80% of the dataset and test with 20%.This paper presented the stance detection of the collected tweets using various approaches to classify people's opinions based on most trending topics in the world. Out of both models LSTM model is better as it's accuracy is higher than SVM. But this thing keeps changing according to the training data as model training is a data driven task.

CONTRIBUTION

- Drumil Shah - Preprocessing of dataset
- Sambhav Gulla - LSTM model implementation
- Himani Desai - Hydrated tweets
- Namra Jain - SVM model implementation

REFERENCES

1. https://www.bing.com/covid
2. Medscape Medical News, The WHO declares public health emergency for novel coronavirus (2020) https://www.medscape.com/viewarticle/924596
3. https://ieeexplore.ieee.org/document/9354776
4. https://dl.acm.org/doi/abs/10.1145/3369026
5. https://en.wikipedia.org/wiki/Support-vector_machine
6. https://en.wikipedia.org/wiki/Long_short-term_memory