Classification of COVID-19 Tweets

Sumana Sree Madasu

Indian Institute of Science Education and Research Tirupati,Andhra Pradesh,India

Abstract

Classification is a vital work to human beings in day today life as it breaks down complex subjects. In the same way, text classification is very important to understand and realize the subject of the text.

Keywords

data preprocessing, Support Vector Machine, Doc2Vec.

1. Introduction

A classification pipeline is constructed to classify the given set of COVID -19 tweets. Here, Data set containing COVID-19 tweets are classified into three different classes using a specific classification pipeline and this classification solves a real world problem. Data is preprocessed, Vectorized using a Doc2Vec model and then Support Vector Machine classifier is used to train the model. The model is then saved and can be used to predict new tweets.

2. Data Preprocessing

Given data have different characteristics, numbers, symbols etc. and this makes the text preprocessing one of the most critical steps in the classification pipeline. Firstly, necessary libraries are imported and then the data is preprocessed by using a few methods depending on the classification task. Here, two functions called clean_text and clean_numbers are used, they return modified text [1](after HTML decoding, lowercase text, replaces symbol by space, deleting few symbols, stopwords and numbers).

3. Implementing Doc2Vec

To implement Gensim's Doc2Vec, every document has to be labeled [2]. Here ,this is done using the TaggedDocument method. Then data is split into train and test sets using the train_test _split from Scikit-Learn library and stratify parameter is used on the data column 'label'. For representing each tweet , a Doc2Vec model is built with each vector as 300 dimension, which iterates over the training corpus 30 times, minimum word count is set to 2 to discard rarely occurring words and alpha is set to 0.065. Vector representation is obtained using the above trained Doc2Vec model for the vocabulary of the data.

madasu@students.iisertirupati.ac.in (S. S. Madasu)

4. Classification Algorithm

Support Vector Machine :

Now, the training data vector list is fitted on the Support Vector Machine (SVM) classifier,which is a traditional machine learning technique [3] and this model is used to predict the labels on the testing data vector list(i.e., validation data list).

5. Classification Report

Classification metrics like f1 score, precision and accuracy are calculated to evaluate the model.

Accuracy - 0.5563

	precision	recall	f1-score	support
Antivax	0.53	0.54	0.53	216
Neutral	0.59	0.62	0.60	327
ProVax	0.54	0.51	0.52	336
accuracy			0.56	879
macro avg	0.55	0.55	0.55	879
weighted avg	0.56	0.56	0.56	879

6. Conclusion

Classification performances of the classifiers depends on quality of training text corpuses to some extent. Most text classification problems are linearly separable [4] and SVM helps in finding those linear separators and this says that SVM can perform well in text classification [5] when over fitting issue is taken care of. For large data samples, Deep Learning techniques show better performance than Traditional Machine Learning techniques.

References

[1] A. Kadhim, An evaluation of preprocessing techniques for text classification, International Journal of Computer Science and Information Security, 16 (2018) 22–32.

[2] Q. Le, T. Mikolov, Distributed representations of sentences and documents, in: International conference on machine learning, PMLR, 2014, pp. 1188–1196.

[3] E. Ikonomakis, S. Kotsiantis, V. Tampakas, Text classification using machine learning techniques, WSEAS transactions on computers 4 (2005) 966–974.
[4] T. Joachims, Text categorization with support vector machines: Learning with many relevant features, in: European conference on machine learning, Springer, 1998, pp. 137–142.
[5] A. Basu, C. Walters, M. Shepherd, Support vector machines for text categorization, in: 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the, 2003, pp. 7 pp.–. doi:10.1109/HICSS.2003.1174243.
References: