Short Text Classification Using TF-IDF Features and Fast Text Learner

Zeshan Khan, Umar Naseer, Muhammad Atif Tahir
{zeshan.khan,umar.naseer,atif.tahir}@nu.edu.pk
FAST School of Computing, National University of Computer and Emerging Sciences, Pakistan

ABSTRACT

The spread of the COVID-19 is a challenge for the health sector. This pandemic created health and financial issues for the whole world. The medical experts are working for the diagnostics and reasons behind the COVID-19 disease and its spread. Some conspiracies are being spread related to the COVID-19 disease and its spread. Such conspiracies can be seen on social media including Twitter. In this research, the conspiracies of the COVID-19 have been analyzed from the public tweets. The tweets of the conspiracies have been filtered from the tweets of the COVID-19 disease, symptoms, and other discussions related to the disease. The analysis of the COVID-19 related tweets resulted into three conspiracy classes, the COVID-19 tweets without any conspiracy and the conspiracies. A model is presented for the classification of tweets into three conspiracy classes with the Matthews Correlation Coefficient (MCC) of 0.294.

1 INTRODUCTION

Social media became a source of information sharing from just closed group chats. The information-sharing generated several trust issues in the information shared on the social media platforms. Currently, Twitter is one of the most used public post-sharing platforms. There are a huge number of tweets being shared daily. There may have several tweets containing misinformation.

In the year 2019 a disease, COVID-19 badly damaged human lives and the economy. There are several solutions proposed for the treatment and spread control of the disease. The guidelines of the health organizations are affected by the false information shared by various individuals. Some of this false information is relating COVID-19 spread with some technological inventions including 5G. A log of people is making a relationship between COVID-19 disease with the 5G technology towers. The time era of COVID-19 and the 5G technology are the same but that doesn't show the one as a cause of other or vise versa.

2 RELATED WORK

The NLP domain is effective from the last decade for various analyses of textual data. One of the domains of textual analysis is text classification. The text classification becomes more challenging when the provided text consists of very short documents. It's very difficult to build a context with the short textual document and the benefit of the short document is the ease of processing.

There is significant work available on the domain of text classification and especially on short text classification. Some of the researchers used various textual feature extraction techniques and then applied classifiers to the textual features. The classifiers of the SVM [14] for the classification using TF-IDF as feature vector [9, 12]. The SVM-based approaches are good in timely detection or classification of the text with lower detection accuracy. There is another group of researches done using neural network-based approaches. The researchers used some pre-trained neural networks like BERT [1] then fine-tuned with the classification dataset [8].

Another type of research for this task is based on graph neural networks (GNN). The GNNs are neural networks that can capture the dependence of the graphs architecture by message passing between perceptrons of the network. There are various variants of GNN for the priority of usage in the domain including graph convolutional network (GCN), graph attention network (GAT), graph recurrent network (GRN), etc. The GNNs are good in detection accuracy with a high time and computational cost [3, 7, 13].

3 APPROACH

The research is based on three different methodologies for a diverse detection of the tweet-class. The Neural Network approaches are performing well in the current era with a limitation of the high availability of the data.

The first approach that has been explored in this research is the usage of cosine similarity between text vectors for the detection of conspiracies in the tweet texts [2, 5]. The idea used in this approach is to split the tweet texts into sentences and apply the learner for classification. A similar learner is used to train on the whole tweet as a single unit. The learner fasttext evaluated both the split and the combined tweet for the MCC. The architecture is visually presented in Figure 1.

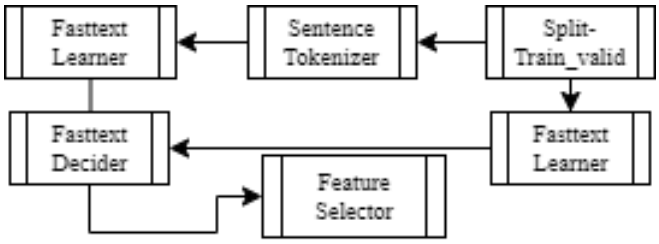

Figure 1: Architecture for FAST Text.

The second approach used in the research is the classification of the TF-IDF vector [6]. This methodology converted tweets into sentences to make the instances smaller. The TF-IDF features are extracted from the sentences. The TF-IDF feature returned in a feature vector of 1045 with most of the zero values. A feature reduction technique of the principal component analysis was performed to reduce the number of features for computation and accuracy advantages [15]. The reduced features vectors were classified using majority voting of some diverse learners including Decision Tree

MediaEval'21, December 13-15 2021, Online

Classifier, Linear Discriminant Analysis, and Logistic Regressions. The architecture is summarised in Figure 2.

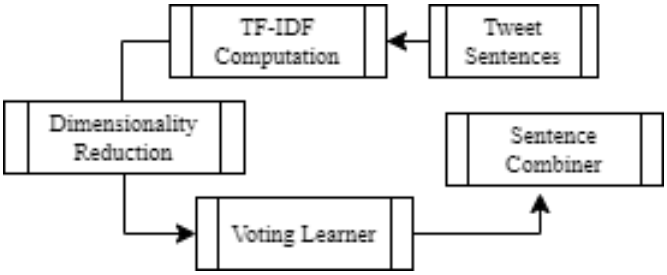

Figure 2: Methodology of Majority Voting of Classifiers using TF-IDF features.

The third algorithm for the detection of conspiracies in the tweet is based on a fully connected neural network with the TF-IDF features of the tweets [6]. The selection of the importance of words is done using two phases of the removal of the word from tweets text. In the first phase the categories of the words that are higher important for the decision between conspiracy have been selected which includes the Nouns, verbs, etc. The second phase of the selection of the important terms is based on the Principal component analysis. The PCA-based top 500 features have been selected to provide to the neural network for ternary decision between conspiracy classes. The detailed architecture of the algorithm can be seen in Figure 3.

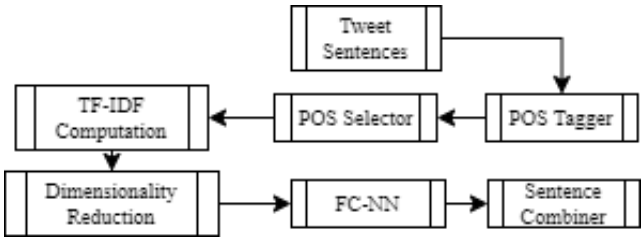

Figure 3: Architecture of Neural Network based Classification

4 DATASET

The research is conducted using the dataset of the MediaEval 2021 under the task of FakeNews: Corona Virus and Conspiracies Multimedia Analysis [10, 11]. The dataset is a set of tweets by various Twitter accounts. The tweets consist of text of the tweet for the detection task. There is some other information available with the tweet for various objectives. The task of the conspiracy classification needs only tweets text and conspiracy class for the training data. The training data provided was of 1511 tweets with various lengths from a few words to several sentences. The class distribution of the tweets, class A, B and, C, was 754, 262, and 495 respectively. The test set was comprised of 266 tweets for the detection of the classes from three classes. The dataset shows there is a class imbalance between the three provided classes. Another finding in the dataset is of class decidability, the class B and C are much closer to each other than the class A.

5 RESULTS AND ANALYSIS

Three approaches were designed to solve the challenge of the conspiracy detection in the tweets. The first approach based on fasttext classification was evaluated on the training data with 30% as validation data. It was evaluated by training with various wordNgrams, learning rates, dimensions and epochs. The best hyperparameters for the fasttext resulted in 1-word gram with a learning rate of 0.7 and 800 dimensions. The model is trained for the 50 epochs due to the limitation of the availability of resources. This approach resulted in 0.89 accuracies on the validation dataset of 30% extracted from the training dataset. The same model when applied to the test dataset resulted in an MCC of 0.294. The second approach for the computation of conspiracy was based on the TF-IDF vector classification using a majority voting classifier. This methodology used the dimensionality reduction technique of PCA with the selection of the top 500 features out of 7147 features. The methodology resulted in an accuracy of 0.56 with an MCC of 0.20 when executed on the training dataset with 30% as validation dataset. The same approach when applied to the test dataset it resulted in an MCC of 0.03. The third approach that is executed in the research is based on a fully connected neural network on reduced TF-IDF features. In this approach, the words of the tweet were selected based on their categorical/ part of speech (POS) importance in decision making. We applied various learners to several categories of the words in a tweet e.g. the verb, nouns, adverbs, adjectives, etc. These learners were guided about the decidability of the various POS. The selected set of words is then used for the computation of the TF-IDF and then PCA is used to reduce the feature vector length from 6898 to 500. The approach is applied to the validation data and the test data and resulted in 0.17 and 0.07 MCC scores respectively.

6 CONCLUSION AND FUTURE WORK

Short text classification is a challenging topic in the domain of natural language processing. There are several challenges due to the unavailability of the context of the sentence due to lesser sentences. Various learners applied on the short text classification and fasttext [2] resulted in the best learner for tweet classification. The results of fasttext guided the use of neural-network (NN) based approaches or LSTM [4] can give better results for the classification of the short text.

The results of the approaches above show that the neural network (NN) based approaches can result in better detection accuracy. The deep learning (DL) based approaches will be explored further to improve detection accuracy. The data is limited and the nature is very close to the various NLP datasets, So, the transfer learning approaches can also be beneficial e.g. BERT can be used with the pre-trained weights for a better understanding of the words and relationships then the tweet data will fine-tune it to decide between conspiracy classes.

REFERENCES

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[2] Stanislav Glebik. 2021. FAST TEXT. https://github.com/facebookresearch/fastText/. [Online; accessed 25-November-2021].
[3] Abdullah Hamid, Nasrullah Shiekh, Naina Said, Kashif Ahmad, Asma Gul, Laiq Hassan, and Ala Al-Fuqaha. 2020. Fake news detection in social media using

graph neural networks and NLP Techniques: A COVID-19 use-case. *arXiv preprint arXiv:2012.07517* (2020).

[4] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[5] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[6] Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of massive data sets.* Cambridge university press.

[7] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. 2019. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 4821–4830.

[8] A Malakhov, A Patruno, and S Bocconi. 2020. Fake news classification with BERT. In *Multimedia Evaluation Benchmark Workshop 2020, MediaEval 2020.*

[9] Manfred Moosleitner, Benjamin Murauer, and Günther Specht. 2020. Detecting Conspiracy Tweets Using Support Vector Machines. (2020).

[10] Konstantin Pogorelov, Daniel Thilo Schroeder, Luk Burchard, Johannes Moe, Stefan Brenner, Petra Filkukova, and Johannes Langguth. 2020. Fakenews: Corona virus and 5g conspiracy task at mediaeval 2020. In *MediaEval 2020 Workshop.*

[11] Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. WICO Text: A Labeled Dataset of Conspiracy Theory and 5G-Corona Misinformation Tweets. In *Proc. of the 2021 Workshop on Open Challenges in Online Social Networks.* 21–25.

[12] Daniel Thilo Schroeder23, Konstantin Pogorelov, and Johannes Langguth. 2020. Evaluating Standard Classifiers for Detecting COVID-19 Related Misinformation. (2020).

[13] Nguyen Manh Duc Tuan and Pham Quang Nhat Minh. 2020. FakeNews Detection Using Pre-trained Language Models and Graph Convolutional Networks. (2020).

[14] Lipo Wang. 2005. *Support vector machines: theory and applications.* Vol. 177. Springer Science & Business Media.

[15] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2, 1-3 (1987), 37–52.