Classification of COVID19 tweets using Machine Learning Approaches

Anupam Mondal[1], Sainik Kumar Mahata[2], Monalisa Dey[3], Dipankar Das[4]
[1,2,3] Institute of Engineering and Management, Kolkata, India
[4] Jadavpur University, Kolkata, India
[1]link.anupam@gmail.com, [2]sainik.mahata@gmail.com
[3]monalisa.dey.21@gmail.com, [4]dipankar.dipnil2005@gmail.com

Abstract

The reported work is a description of our participation in the "Classification of COVID19 tweets containing symptoms" shared task, organized by the "Social Media Mining for Health Applications (SMM4H)" workshop. The literature describes two machine learning approaches that were used to build a three-class classification system, that categorizes tweets related to COVID19, into three classes, viz., self-reports, non-personal reports, and literature/news mentions. The steps for pre-processing tweets, feature extraction, and the development of the machine learning models, are described extensively in the documentation. Both the developed learning models, when evaluated by the organizers, garnered F1 scores of 0.93 and 0.92 respectively.

1 Introduction

In order to identify personal tweets related to COVID-19, it becomes necessary to distinguish them from tweets made by others related to this issue. Classification of medical symptoms from posts related to COVID-19 poses two major challenges: Firstly, the amount of information available as news articles, scientific papers etc that describe various medical symptoms is huge (Mondal et al., 2017; Kushwaha et al., 2020). All this information makes it extremely difficult to spot significant user reported information. Secondly, there are multiple users who report information which is not experienced by themselves but by other people they know or come across (Mondal et al., 2018; Li et al., 2020). This makes the task of identifying self reported information from the huge amount of discourse available very complex.

The current shared task (Task No. 6)[1], namely "Classification of COVID19 tweets containing symptoms" provided participants with three classes viz. **i.** self-reports **ii.** non-personal reports, and **iii.** literature/news mentions. This task is a three way classification task.

For developing the learning models, we used traditional Machine Learning (ML) and state-of-the-art Deep Learning (DL) approaches (Imran et al., 2020; Chakraborty et al., 2020; Gencoglu, 2020). Besides, extra features, like Parts-of-Speech (POS) tags as well as Term Frequency-Inverse Document Frequency (TF-IDF) was used, which enabled the developed models to learn the hidden classes better.

Upon evaluation, our developed models performed well and this was ratified by the fact that they garnered F1 scores of 0.93 (ML model) and 0.92 (DL model) respectively.

The rest of the paper is organized as follows. Section 2 describes the data and methodology that was used to develop both the models. This section describes the pre-processing steps, will talk about the extra features that were used, and will also narrate the learning models that were used to build our systems. Following this, Section 3 and 4 will chronicle the results and the concluding remarks respectively.

2 Methodology

Initially, the organizers provided us with 9,567 training data and 500 validation data. This labeled dataset consisted of three fields; tweet id, the actual tweet, and the respective label (self-reports, non-personal reports, and literature/news mentions). The training and validation data were later combined and it was pre-processed for further development. Steps of pre-processing the tweets included the removal of extra characters to clean the data. The extra characters that were removed/cleaned included mentions, punctuation's and URLs. Additionally, words from hashtags were extracted and extra spaces were contracted. After the pre-processing steps, POS tags of individual words of every tweet were found out using the python pack-

[1]https://healthlanguageprocessing.org/smm4h-2021/task-6/

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 135–137

ages Natural Language Toolkit[2] (NLTK). POS tag features were used as they can help in determining authorship as people's use of words varies. On the other hand, it can easily differentiate between the same words, applied in different settings. E.g., "like" is a verb semantically charged with positive weight, as in "I like you", but it becomes neutral conjunction, as in "I am like you".

For feeding the extra POS tag feature along with individual words, to the ML model, we concatenated them to form an extended input of the structure

$$W_1\_P_1 \quad W_2\_P_2 \quad .... \quad W_n\_P_n$$

where W are the word and P are the POS tag of the word. After the concatenation was done, the input was fed to a TF-IDF vectorizer, which converts a collection of raw documents to a matrix of TF-IDF features. In order to extract the most descriptive terms in a document, we have used TF-IDF features. Besides, this feature assists in computing the similarity between two words which enhances the feature quality to allow even simple models to outperform more advanced ones.

Additionally, the corresponding labels of the tweets were fed to a Label Encoder, which encodes target labels with a value between 0 and n_classes - 1, where n_classes in our case was 3.

Both these vectorized inputs and encoded outputs were fed to a Multi-layer Perceptron classifier (MLP), where alpha was kept at 1 and maximum iterations were kept at 1,000.

For training the DL model, we took the words, POS tags, and TF-IDF values as separate inputs passed them through their respective default Tensorflow embedding layer, and concatenated the outputs. The output, from the concatenation layer, was then passed through two layers of bi-directional Long-Short Term Memory (Hochreiter and Schmidhuber, 1997) (LSTMs) and finally fed to a dense layer which mapped the tensors to the respective labels.

Other parameters of the model were as follows. Optimizer was kept as "adam" and loss was kept as "SparseCategoricalCrossentropy". The batch size was kept as 128 and the number of epochs was fixed at 50. Also, the early stopping mechanism, where the metric was fixed to validation loss, was applied to stop over-fitting. A depiction of the developed model is shown in Figure 1.

[2]https://www.nltk.org/

Both the ML and DL models were then deployed on the 500 validation data provided by the organizers and upon submission, garnered F1 scores of 0.98 and 0.97 respectively. Other validation metrics are shown in Table 1.

Model	Precision	Recall	F1 Score
ML model	0.9720	0.9881	0.98
DL model	0.9660	0.9660	0.97

Table 1: Evaluation scores of the developed models.

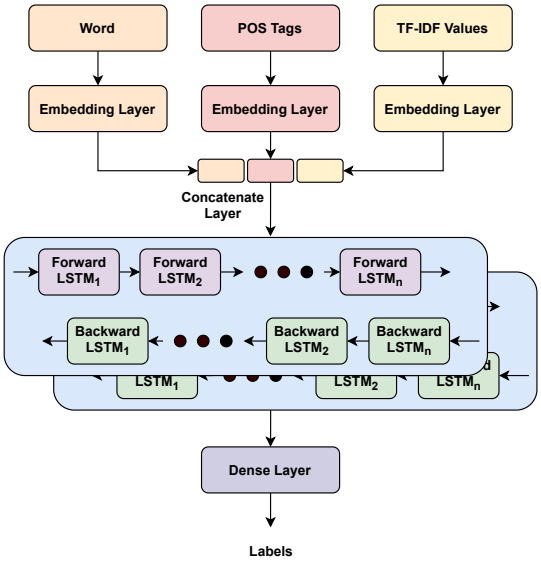

Figure 1: Developed deep learning model for the classification.

3 Evaluation

6,500 tweets were provided by the organizers of the shared task, as test data. Both the developed models were deployed on the same and the results were submitted for evaluation. Upon evaluation, our models garnered micro f1 scores of 0.93 and 0.92, for the ML and DL models respectively. Other scores are shown in Table 2.

Model	Precision	Recall	F1 Score
ML model	0.9337	0.9337	0.93
DL model	0.9248	0.9248	0.92

Table 2: Evaluation scores of the developed models.

Additionally, the arithmetic median of all submissions made by other participating teams is shown in Table 3.

4 Conclusion

The reported system paper presents two models developed using ML and DL approaches, that were trained to classify tweets related to COVID19, into personal/non-personal mentions or standard literature. From the results, we can see that since the amount of training data was low, traditional ML

Precision	Recall	F1 Score
0.93235	0.9337	0.93

Table 3: Median scores of all the participating teams.

methods performed very well. On the contrary, the proposed DL model performed as well, if not better, on the same less amount of data. This is an interesting observation as, more often than not, DL methods rely on huge amounts of data for learning patterns. As future work, we plan to expand this work, by increasing the data and applying state-of-the-art embedding methods like BERT, RoBERTa, etc., on the same.

References

Koyel Chakraborty, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment analysis of covid-19 tweets by deep learning classifiers—a study to show how popularity is affecting accuracy in social media. *Applied Soft Computing*, 97:106754.

Oguzhan Gencoglu. 2020. Large-scale, language-agnostic discourse classification of tweets during covid-19. *Machine Learning and Knowledge Extraction*, 2(4):603–616.

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

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on covid-19 related tweets. *IEEE Access*, 8:181074–181090.

Shashi Kushwaha, Shashi Bahl, Ashok Kumar Bagha, Kulwinder Singh Parmar, Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. 2020. Significant applications of machine learning for covid-19 pandemic. *Journal of Industrial Integration and Management*, 5(4).

Irene Li, Yixin Li, Tianxiao Li, Sergio Alvarez-Napagao, Dario Garcia-Gasulla, and Toyotaro Suzumura. 2020. What are we depressed about when we talk about covid-19: Mental health analysis on tweets using natural language processing. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*, pages 358–370. Springer.

Anupam Mondal, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. 2017. Employing sentiment-based affinity and gravity scores to identify relations of medical concepts. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7. IEEE.

Anupam Mondal, Dipankar Das, and Sivaji Bandyopadhyay. 2018. A content-based recommendation system for medical concepts: Disease and symptom. In *15th International Conference on Natural Language Processing*, page 120.