Comparison of Depression Detection Algorithms And Analysing Differences In Stress Level During and Before The COVID Pandemic

Samarth P
Computer Science Engineering
JSS Science And Technology University
Mysuru, India
samarthp1@gmail.com

Tejus R S
Computer Science Engineering
JSS Science And Technology University
Mysuru, India
tejusrso@gmail.com

Nihal Pradhan
Computer Science Engineering
JSS Science And Technology University
Mysuru, India
nihalpradhan10@gmail.com

***Abstract*—The digital age has witnessed many people turning to social media to influence society with their views and express their feelings. This makes social media a huge reservoir of unstructured data. In this paper, we try to leverage a small chunk of the plethora of information social media hosts. We find the best possible algorithm that has the ability to extract accurate sentimental information from a large tweets dataset based on various parameters. The pandemic that has disrupted everything from the economy to our lifestyles has had a significant impact on our stress levels. We use the best algorithm to gain valuable insights on tweets during the COVID19 pandemic, before the pandemic, and during certain other disasters. We compare our results and give a possible explanation as to why it is so.**

***Index Terms*—COVID19, Natural Language Processing, Depression detection**

I. Introduction

The proliferation of internet and communication technologies, especially the online social networks have rejuvenated how people interact and communicate with each other electronically. The applications such as Twitter, Twitter, Instagram and alike not only host the written and multimedia contents but also offer their users to express their feelings, emotions and sentiments about a topic, subject or an issue online. On one hand, this is great for users of social networking site to openly and freely contribute and respond to any topic online; on the other hand, it creates opportunities for people working in the health sector to get insight of what might be happening at mental state of someone who reacted to a topic in a specific manner. [1]

In order to provide such insight, machine learning techniques could potentially offer some unique features that can assist in examining the unique patterns hidden in online communication and process them to reveal the mental state (such as 'happiness', 'sadness', 'anger', 'anxiety', depression) among social networks' users. As users express their feeling as a post or comments on the Twitter platform, sometimes their posts and comments refer to as emotional state such as 'joy', 'sadness', 'fear', 'anger', or 'surprise'. We analyze various features of Twitter comments by collecting data through an effective method of machine learning classification techniques and to make overall judgments regarding their various parts. [2]

Many researchers have demonstrated that utilizing user-created content accurately may help decide individuals' psychological wellness levels. For instance, Aldarwish and Ahmad [3] examined that the utilization of Social Network Sites (SNS) is expanding these days, particularly by the more youthful eras. Because the accessibility of SNS enables clients to express their interests, sentiments and offer day by day schedule [4] [5].

In this paper, we first extract a dataset of tweets with depressive keywords. We then use different algorithms like Logistic Regression, K-Nearest Neighbors, Decision Trees, Support Vector Machine, and CNN + LSTM neural network over this dataset to find out the best algorithm. We then use this model to run on more contextual datasets such as COVID related datasets, disaster-related datasets to gain valuable insights.

II. Methodology

A. Procuring the Dataset

The Dataset consists of labeled data i.e., an equal number of depression indicative tweets and random tweets (non-negative). There is no publicly available dataset for depressive tweets. So, we had to gather our own dataset. We used a tool called TWINT [6] which is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. We used TWINT with appropriate keywords that depict negative stressful emotions such as "depressed", "suicidal", "distraught", "pathetic", etc. Then random tweets were obtained from the Kaggle dataset twitter_sentiment. We mix these two datasets in equal proportions to get an evenly distributed dataset of both negative and non-negative tweets. This dataset forms the basis for training, validating and testing all our models. We use 60% of the dataset for training, 20% for validation and the rest 20% for testing the accuracy.

B. Data Processing

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Preprocessing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis. In this step, we remove links and images, hashtags, @ mentions, emojis with the help of a python module called ftfy is used fix weirdly worded texts. The goal of ftfy is to take in bad Unicode and output good Unicode, for use in your Unicode-aware code. The expandContractions function is used to expand words such as ”ain't” to ”am not”. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query. NLTK [7] (Natural Language Toolkit) in python has a list of stopwords stored in 16 different languages.

Stemming is the process of producing morphological variants of a root/base word. A stemming algorithm reduces the words “chocolates”, “chocolatey”, “choco” to the root word, “chocolate” and “retrieval”, “retrieved”, “retrieves” reduce to the stem “retrieve”. NLTK library in python has a function to stem the tweets.

Tokenization is essentially splitting a phrase, sentence, paragraph, or an entire text document into smaller units, such as individual words or terms. Each of these smaller units are called tokens.A Tokenizer is used to assign indices and filtering out unfrequent words. Tokenizer creates a map of every unique word and an assigned index to it. The parameter called num_words indicates that we only care about the top 20000 most frequent words.

Word embedding is the collective name for a set of language modeling and feature learning techniques in natural language processing (NLP) where words or phrases from the vocabulary are mapped to vectors of real numbers.Word2Vec [8] is one of the most popular technique to learn word embeddings using shallow neural network. Load the pretrained vectors for the Word2Vec model. Using a Keyed Vectors file, we can get the embedding of any word by calling .word_vec(word) and we can get all the words in the model's vocabulary through .vocab. After all the above preprocessing steps have bee completed we deem the data to be ready for use in classification algorithms. We now use this dataset to train the following algorithms.

C. Logistic Regression

In statistics, the logistic model is used to model the probability of a certain class. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one. In order to map predicted values to probabilities, we use the Sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities. We use this to predict whether the tweet is a depressed/negative tweet or not.

$$f(x) = \frac{1}{1 + \mathrm{e}^{-(x)}}$$

D. KNN classifier

K-nearest neighbors (KNN) algorithm is a type of supervised ML algorithm which can be used for both classification as well as regression predictive problems. However, it is mainly used for classification predictive problems in industry.K-nearest neighbors (KNN) algorithm uses ‘feature similarity’ to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. The number of neighbors we have used for our purposes is five and the metric we have defined it upon is Minkowski distance.

E. Decision Tree

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).

A tree is built by splitting the source set, constituting the root node of the tree, into subsets—which constitute the successor children. The splitting is based on a set of splitting rules based on classification features. This process is repeated on each derived subset in a recursive manner called recursive partitioning. The recursion is completed when the subset at a node has all the same values of the target variable, or when splitting no longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data. For our study we have used a decision tree with the criterion as entropy and the random state as zero.

F. Support Vector Machine (SVM)

Support vector machines (SVMs) are powerful yet flexible supervised machine learning algorithms which are used both for classification and regression. But generally, they are used in classification problems. Lately, they are extremely popular because of their ability to handle multiple continuous and categorical variables.

An SVM model is basically a representation of different classes in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized.The main goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH) and it can be done in the following two steps. First, SVM will generate hyperplanes iteratively that segregates the classes in best way. Then, it will choose the hyperplane that separates the classes correctly. For our study we use the SVM algorithm with the Radial basis function(RBF) kernel and the random state as zero.

G. CNN + LSTM Model

The CNN LSTM architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs (Long Short Term Memory) to support sequence prediction. CNN LSTM model is specifically designed for sequence prediction problems with spatial inputs, like images or videos or NLP. This architecture involves using Convolutional Neural Network (CNN) layers for feature extraction on input data combined with LSTMs to perform sequence prediction on the feature vectors. In short, CNN LSTMs are a class of models that are both spatially and temporally deep and sit at the boundary of Computer Vision and Natural Language Processing. These models have enormous potential and are being increasingly used for many sophisticated tasks such as text classification, video conversion, and so on. The model takes in an input and then outputs a single number representing the probability that the tweet indicates depression. The model takes in each input sentence, replace it with it's embeddings, then run the new embedding vector through a convolutional layer. CNNs are excellent at learning spatial structure from data, the convolutional layer takes advantage of that and learn some structure from the sequential data then pass into a standard LSTM layer. Last but not least, the output of the LSTM layer is fed into a standard Dense model for prediction. The model is trained EPOCHS (Five) time, and Early Stopping argument is used to end training if the loss or accuracy don't improve within 3 epochs. We use a CNN layer with 32 filters, kernel size as 3 and the activation as Relu. We then use an LSTM layer with 300 memory units with a dropout of 0.2. the dense layer has sigmoid activation for prediction.

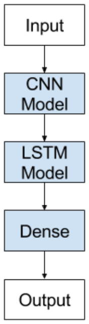
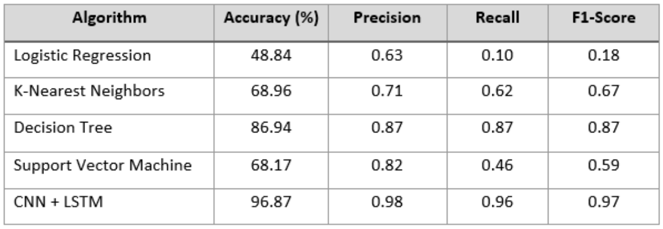


Fig. 1: CNN + LSTM Architecture

III. Results

We have used our dataset to train, validate and test on all the above mentioned algorithms. We will now be analysing the results on multiple parameters such as accuracy, precision, recall and F1-score. The accuracy is the percentage of the times the model has given the correct result during testing. Precision is the measure of how close the predicted value is to the actual value. Recall is the fraction of the total amount of relevant instances that were actually retrieved. F1-score conveys the balance between the precision and the recall. The results from different machine algorithms we have used is displayed below with the parameters. It is clear from the above depiction that the CNN+LSTM model is a clear winner across all judging parameters. This suggests that this model is the most ideal one to use for natural language processing. We will now further use this model to gain valuable insights and analyse other text based datasets.



Algorithm	Accuracy (%)	Precision	Recall	F1-Score
Logistic Regression	48.84	0.63	0.10	0.18
K-Nearest Neighbors	68.96	0.71	0.62	0.67
Decision Tree	86.94	0.87	0.87	0.87
Support Vector Machine	68.17	0.82	0.46	0.59
CNN + LSTM	96.87	0.98	0.96	0.97

Fig. 2: Comparison of results from different machine learning algorithms

IV. Further Work

The COVID19 pandemic has changed all of our lifestyles and has affected the economy significantly. When any of us undergo any changes, especially something as large as this, we undergo stress. In the intent to understand the magnitude of this stress, the differences between our lifestyle change and its correlation with mental stress, and to provide a platform for psychologists to understand avoid behavioral problems, we are conducting the following studies. We fist procure three datasets:

1) A large dataset of random tweets before the pandemic started. This represents the way the people interacted and how they did not have the pandemic to worry about. This was also when the global economy was doing well.
2) A dataset of COVID19 related tweets which represents the way the people interact on this topic. It largely represents how the stress levels have changed during the pandemic and how the people are coping wit it.
3) A dataset that contains tweets pertaining to disasters. These are tweets that represent natural or man-made disasters like the Australian bushfires and the way people empathised and stressed on them. This is taken before the pandemic, so it represents disasters that are not as widespread but hardh nonetheless.

The CNN + LSTM model is deployed on all the three of these datets to gain insights on how the stress levels have varied in the society.the outpouts on the three datasets were as follows:

1) Common Dataset Result:
   Total number of Tweets: 399701
   Number of Depressive Tweets: 8429
   Number of Non-Depressive Tweets: 391272
   Percentage of Depressive Tweets: 2.108826347694902
2) COVID19 Datset Result:
   Total number of Tweets: 244198
   Number of Depressive Tweets: 155024

Number of Non-Depressive Tweets: 89174
Percentage of Depressive Tweets: 63.48291140795585

3) Disaster Dataset Result:
Total number of Tweets: 11305
Number of Depressive Tweets: 6252
Number of Non-Depressive Tweets: 5053
Percentage of Depressive Tweets: 55.302963290579385

V. Conclusion

In this paper we checked multiple models to choose the best possible model for NLP and we found it to be the CNN + LSTM model. We used this model to find out the stressful tweets in three different datasets and we have shown the results we have found.

There is a clear difference between the the random dataset that we chose before the pandemic and the dataset during the pandemic. the fact that there were just 2.1% before the pandemic and this jumped to 63.48% for pandemic related tweets shows us how stressful the pandemic has been. this also suggests that the pandemic management teams across the globe need to take appropriate steps to alleviate these issues as much as possible.

The fact that the disaster related tweets are only 55.3% when you would expect it to be much higher than the pandemic because it contains all tweets related to only disasters shows us that not everyone empathises as much to stress over it. This is because only a few people are affected by such local disasters where as the pandemic has affected most people across the globe. Even though these disasters have caused more property and wildlife damage than the pandemic, the people tend tweet to spread awareness or support rather than empathise with them.

The above results also suggests people are most bothered when their lifestyle is changed rather than burning forests or other environmental disasters else where which ultimately affects the entire earth.If we want to effectively combat global warming and other important issues we need to fist influence and invoke sentiment and empathy regarding such issues on a global scale to make everyone contribute to the cause.

Our paper has highlighted the above analysis with substantial evidence. We intend this to be a talking point of many discussions at different levels of society and a base for prudent action to be taken on such important topics

References

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, “Predicting depression via social media,” in *Seventh international AAAI conference on weblogs and social media*, 2013.

[2] B. O’dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, “Detecting suicidality on twitter,” *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[3] M. M. Aldarwish and H. F. Ahmad, “Predicting depression levels using social media posts,” in *2017 IEEE 13th international Symposium on Autonomous decentralized system (ISADS)*. IEEE, 2017, pp. 277–280.

[4] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, “A depression detection model based on sentiment analysis in micro-blog social network,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2013, pp. 201–213.

[5] J. Zhou, Y. Zhao, H. Zhang, and T. Wang, “Measuring emotion bifurcation points for individuals in social media,” in *2016 49th Hawaii international conference on system sciences (HICSS)*. IEEE, 2016, pp. 1949–1958.

[6] F. Poldi, “Twint-twitter intelligence tool,” *URL: https://github.com/twintproject/twint*, 2019.

[7] E. Loper and S. Bird, “Nltk: the natural language toolkit,” *arXiv preprint cs/0205028*, 2002.

[8] X. Rong, “word2vec parameter learning explained,” *arXiv preprint arXiv:1411.2738*, 2014.