Multi-label Sentiment Classification and Analysis of COVID-19 Tweets

Fizza Tauqeer
Information Technology University
msds19034@itu.edu.pk

Muhammad-ur-Rehman
Information Technology University
msds19058@itu.edu.pk

Soban Mahmood
Information Technology University
mscs16012@itu.edu.pk

Abstract

Though Twitter sentiment analysis modulations are rife in the current age of Deep Learning, there is little focus on the existence of multiple emotions in tweets as it presents itself as a taxing task due to a prevalent inability in identifying inter-label dependencies and correlations. The presence of one emotion entails a highly likely occurrence of another in the same subset, yet is usually disregarded when attempting sentiment analysis of tweets in order to focus primarily on singular labels. If explored concretely, it is liable to open up exploitation of why the existence of one emotion implies the existence of another, and how they can be utilized together to better approach emotion-based openings. The aim of this project is to utilize sentiment-relevant frameworks such as Bag-of-Words (BOW) and Bidirectional Encoder Representation from Transformers (BERT), alongside sequential networks such as Long Short-Term Memory (LSTM) to identify these multiple labels at higher rates of efficiency than those provided by usual attempts of most modular methods. Through our attempt, we will also analyze the identifiers that trigger each emotion, such as month frequency and location-based clusters. The aim of this research is to recognize and analyze the emotions of tweets relevant to COVID-19, in order to provide beneficial value to prospective stake holders in making decisions with regard to the virus, general health and well-being. We were able to achieve an accuracy rate of 82% and 87% respectively with optimal individual loss functionalities for the task using a BERT-base uncased model tuned to certain parameter settings.

1. Introduction

Social media platforms such as Twitter and Facebook are vast and vital sources of information when it comes to analyzing, observing and making educated deductions in relation to everyday events happening in our periphery. Unfortunately, one such event currently taking over these applications happens to be COVID-19 - a novel type of contagious coronavirus that attacks the living organism's respiratory system leading to its breakdown - which began to illustriously appear in late November of 2019 and has since become a global pandemic in a mere matter of 6 months. Many countries have gone into severe lockdowns at the disease's onslaught to contain transmission for the safety of their nationals. This lockdown has – naturally - appeared to affect numerous individuals mentally and physically, with primary hypotheses suggesting social confinement, onslaught of unemployment, lack of consumer resources and the impeding expectation of financial crisis. Consequently, a huge number of social media users are intentionally and unintentionally actively expressing their experiences, thoughts and feelings on these social media platforms, particularly Twitter.

With this massive flow of continuous information, it is imperative to gauge through the help of these outlets of how people are being affected by, and reacting to, COVID-19 and what type of emotions they are expressing in their tweets. Such analysis can aid future creators of helper systems, organizations, governments and mental health workers of how severe a solution or treatment need be based on the context of these emotions and situations. For example, areas where negative emotions such as fear and pessimism are at an all-time high will need to concretely implement therapy organizations to combat the aftermath of the virus such that psychotic episodes can be avoided at best given the expected mentally fragile nature of most individuals having to deal with this crisis. Similarly, it can be expected that virtual shopping platforms will improvise their platforms given the exponential rise in online shopping in the current times. Such significant mappings were heavily considered as we approached the intended problem.

Attempting sentiment analysis, it is evident that any opinion of any individual is liable to sustain many emotions rather than only one. Approaching the sentiment identification problem in such a way ensures that relative emotions can also be recognized in similar fashion to individual ones. Hence, if one emotion is able to be detected then it is equally likely that related emotions can also be detected congruently. With so many nuances present in the emotions spectrum, this helps in gauging the effect of sentiments and their trends with respect to COVID-19.

Our study aims to provide a multi-label reliant sentiment model system (based on an analysis of emotions such as happiness, sadness, fear, optimism, and many more emotions that appear on the encompassing Plutchik's Wheel of Emotions). Our final solution is based on a thorough study of what optimized algorithms we utilized and the motivation behind each (such as the currently best BERT model) and what interventions must be made to improve existing results (such as our experimentation on the BERT and LSTM model combined), while bringing ingenuity. The system is also focused with ensured data from countries such as Pakistan, USA, Italy and Iran as they have been proven to be hit the hardest from COVID-19 and thus can naturally provide more realistic data and sentiments in tweets. For this purpose, model testing for sentimental analysis of multiple labels has been done initially in order to derive prized graphical conclusions based on numerous emotions. Upon obtaining these results, we moved towards taking advantage of locational and time-series data identifiers for analytical purposes, which we have included in our 'Experiments' and 'Conclusion' sections. We have also provided a comparison between our implementations and previously carried out research alongside our work in order to certify our attempt's standing in the overall approach of sentiment analysis and classification.

Using our methodology outlined in the 'Methods' and 'Experiments' sections, we were able to achieve a sustainable accuracy of 82% on Testing Data consistent to emotion-tagged Training Data. Earlier, we had achieved 87% accuracy, but upon the replacement of the Binary Cross-Entropy with Logits Loss functionality with the instructor-suggested Focal Loss as our primary Loss function it decreased a further 5%. We have touched upon the reasoning for such jumps in the 'Experiments' section extensively.

One can find the implementation details of our project attempt at the following GitHub Links:

https://github.com/MuhammadurRehman19058/MSDS19058_Project_DLSpring2020

2. Related Work

In order to acquaint ourselves with how the problem had been attempted at in previous literature and research articles, we analyzed multiple related works which focused on predicting sentiments of Twitter users. It was also noticed that the problem of sentiment analysis being addressed specific to COVID-19 had by this time already begun to take momentum. The works mentioned below inspired our own problem methodology quite significantly as well.

Manguri et al. [1] used a Naïve-Bayes (NB) based classification technique through predefined sentiments of lexicons on Twitter data which was extracted and analyzed through the help of beginner-friendly TextBlob and Tweepy Python-based libraries. The goal of this research appeared to be more oriented towards analytical tasks in contrast to classifying tweets on emotions. The study is a stark representation of the fact that using basic frameworks can also yield fruitful results and insights The procedure of their research is defined generally in Fig. 1 below:

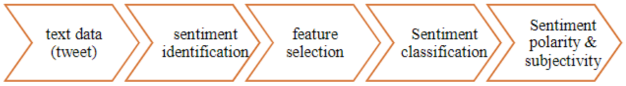


Figure 1: Sentiment Analysis Procedure with a core NB framework

This study guided us in our strategy of scrapping tweets data from Twitter, as otherwise it focused solely on positive, negative and neutral sentiments, which did not correlate to our method of sentiment identifiers. They were able to gauge meaningful statistics, such as that of 36% of people holding positive sentiments, while 14% held negative views. However, it is to be noted that the study was conducted at a time when COVID-19 was not being considered as a major threat to health world-wide, thus it may be the reason why positive feelings are in majority in comparison to negative feelings.

Nawaz et al. [2] categorized tweets into long (number of text characters < 120) and short (number of text characters < 77) tweets using Naïve-Bayes, K-Nearest Neighbor (KNN) and Linear, Logistic Regression models with a maximal range of 91% accuracy on short tweets and a minimal range of 74% accuracy on long tweets. Similar to previously mentioned literature, this outlook also focused primarily on positive and negative sentiments as a binary class problem through machine learning techniques rather than deep learning methodologies, hence we could not replicate a majority of their work.

However, their usage of extracting and pre-processing Tweets using the rTweet package in R was a welcome solution to our troubles with data acquisition in comparison to other programming language frameworks such as that in Python, which we did implement successfully for our problem statement.

Irene et al. [3] used a pre-trained BERT model for single-label classification and a fine-tuned BERT model for multi-label classification on the EmoCT (Emotion-Covid19-Tweet) dataset curated and emotion-tagged by the authors themselves for English, Spanish, Portuguese, Japanese, German and Chinese language tweets. The single-label classifier amounted to 95% accuracy, while the multi-label classifier amounted to an average precision of 64% with a minimal coverage error of 3.2% respectively.

This study served as our prime motivation in including the BERT model for our own classification task, as prior to it we were solely focusing on implementing a baseline bidirectional LSTM network with pre-trained global vectors (GloVe) of Twitter embeddings due to the influence of our paper presentation related to our problem, namely the research “A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets” [4]. However, due to the idea being relatively new and unexplored, the credibility of this research was questioned by our course instructor, thus providing us with the opportunity to explore other model ventures as well. We decided to implement both of these mentioned techniques for comparative purposes (as explained in our ‘Methods’ and ‘Experiments’ sections), but luckily our peaked interest in this specific literature guided us towards implementing the better approach as is certified later on in this paper. It was also an eye-opener towards what attributes were a gold-mine in being exploited in terms of analytics and visualizations to gather insights into what trends were emerging with the continuous growth of COVID-19. A prime example of such insights can be seen in Fig. 2 below.

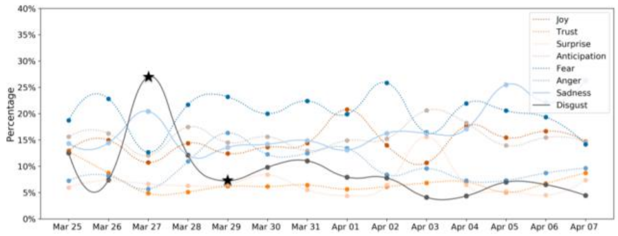


Figure 2: Emotion Trend on the word ‘lockdown’ from March 25 to April 7, 2020

Shifting away from solely analyzing Twitter sentiments, Jelodar et al. [5] modelled an LSTM-based approach on Reddit threads that revolved around discussions on COVID-19 in order to gather sentiments ranging from very positive to very negative, along with further sub-divisions. This study was one of the first known implementations of the sentiment analysis framework on the magnanimous COVID-19 topic, further influencing the analysis to other forums of social media – ultimately Twitter. The authors were able to achieve an accuracy of 81.15% with GloVe embedding in the LSTM framework.

As data on Reddit is quite different than data extracted from Twitter, it was a nice comparison to add in our related work due to our own envisioned goal of using a Bidirectional LSTM for breaking down tweet text sentence structures for emotion consumption.

While exploring our decision to utilize both BERT and LSTM modules for our project, we came across an interesting study [6] in which the standard BERT architecture was used for classification of tweets from the famous Crisis Lex and Crisis NLP datasets on the basis of managing disasters, along with several other customized trained BERT architectures to compare with the baseline bidirectional LSTM, again with pre-trained Glove Twitter embeddings. Results showed that the BERT and BERT-based LSTM performed at par from the baseline model by a value of 3.29% on micro, macro-averaged F-1 scores. It was an interesting study which was found after we had completed our own implementation, but it nonetheless clarified a lot of our own inhibitions, assumptions and hypotheses of the predictions occurring, particularly why certain modulations were performing better or worse in comparison to each other. The authors were able to cement the fact that ambiguity and subjectivity affected the performance of these models considerably, while in some fine-tuned experiments the models were able surpass set standards of human performance.

3. Data

We utilized a total of two datasets for our problem, in order to be able to gauge the rate of success and efficiency that each set of data ingestion into our models would provide.

The dataset of COVID-19 relevant tweets which we utilized in our models and experiments was gathered using Twitter Streaming Application Programming Interfaces (APIs) along with Lookup APIs for Tweet Identifiers (ID extraction). For our research, Lookup APIs were used specifically to get these precisely historical tweets related to the coronavirus strain. Globally, we extracted Tweet IDs of 2, 40, 070 tweets (in our aim to represent a realistic approach) from the real-time Coronavirus Tweets Dataset available at IEEE Data Port [7] – we discarded the additional sentiments of positive/negative nature provided alongside it as our problem was not modelled along those sentiment lines. The open-sourced Tweet IDs were extracted by the IEEE Data Port system on the following keywords: ‘corona,’ ‘coronavirus,’ ‘covid,’ ‘pandemic,’ ‘lockdown,’ ‘quarantine,’ ‘hand sanitizer,’ ‘ppe,’ ‘n95,’ different yet highly possible variants of ‘sarscov2,’ ‘nCov,’ ‘covid-19,’ ‘ncov2019,’ ‘2019ncov,’ ‘flattening the curve,’ ‘social distancing,’ and ‘working from home.’ This dataset holds multi-lingual tweets but for our experiments we mined and concentrated on subsets of English tweets only, dated from March 21, 2020 to April 03, 2020.

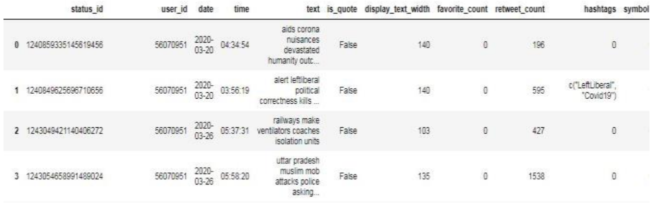
After extracting the Tweet IDs, the actual tweets data was scrapped from Twitter using the ‘rTweet’ library in R language. The reason of shifting to R from Python was due to the fact that we were limited by iteration and batch sizes of scrapping tweets in Python incredibly – only a size of 100 iterations and batches were possible in one run of the scrapping script. In R – on the contrary – we could iterate over a size of simultaneously-occurring 8200 iterations and batches in one run. We also extracted the twitter user's data location in the same framework using similar Lookup API methodologies inherent in rTweet for locational data exploitation further down the line. The numerous attributes selected with the defined purpose of classification and visualization are defined in Table 1. as follows:

tweet_id	user_id	Date
favourite_ count	retweet_ count	hash tags
time	tweet_text	is_quoted
symbols	language	Location

Table 1: The features extracted for the COVID-19 Tweets Dataset

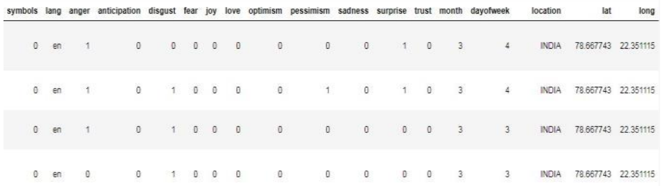
This accumulated data was further pre-processed extensively in both R and Python frameworks with multiple processing passes to remove textual errors that are liable to cause trouble while training the data. Lowercasing and Normalization were heavily done: retweet URLs or any URL for that matter were deleted, user-mentions and tags were removed, all tweet texts were lower-cased, along with removal of special characters and redundant spaces using the dedicated, coded pre-processor. This reduced the data size (and computation later on) without losing information and integral data. For example, the processing was such that even the non-ASCII characters along with stop-words would be removed from the tweets text feature. Using regular expressions for recognition, metadata information such as Twitter markup, emoticons, dates, times, currencies, acronyms, hashtags, user mentions, URLs, retweet counts and words with emphasis were removed successfully and proficiently from the tweets text feature.

From this final dataset of computed COVID-19 tweets, tweets dated April 02, 2020 and of roughly 8000-records were randomly selected for the purpose of Human Labeling of the emotions identified in these tweets by 39 annotators. Our models were then trained (4937 samples), validated (1064 samples) and tested (1085 samples) on this labelled data. The 11 emotions which were labeled and later classified were anger, anticipation, disgust, fear, joy, love, optimism, pessimism, sadness, surprise and trust – emotions that exist on the Plutchik's Wheel of Emotions. A subset of the data curated is shown in Figs. 3 and 4. in order to gain an understanding of what the final data looked like. It is imperative to note here that data is subjective to the techniques employed, which are of course limited in nature due to the resources and time we had on hand. More accurate approximations and results can be achieved if the volume of data scrapped is increased, tagged and tested. Thus, we worked with all such datasets with a grain of salt in retrospective.



	status_id	user_id	date	time	text	is_quote	display_text_width	favorite_count	retweet_count	hashtags	symbol
0	1240859335145619456	56070951	2020-03-20	04:34:54	aids corona nuisances devastated humanity outc...	False	140	0	196	0	
1	1240849625696710656	56070951	2020-03-20	03:56:19	alert leftliberal political correctness kills ...	False	140	0	595	c("LeftLiberal", "Covid19")	
2	1243049421140406272	56070951	2020-03-26	05:37:31	railways make ventilators coaches isolation units	False	103	0	427	0	
3	1243054658991489024	56070951	2020-03-26	05:58:20	uttar pradesh muslim mob attacks police asking...	False	135	0	1538	0	

Figure 3: The head outlook of the first few features of the curated COVID-19 Tweets Dataset

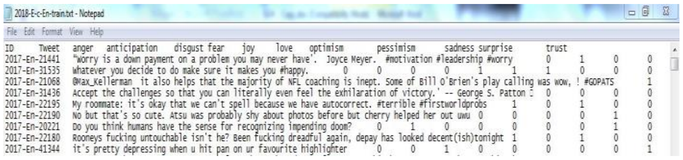


symbols	lang	anger	anticipation	disgust	fear	joy	love	optimism	pessimism	sadness	surprise	trust	month	dayofweek	location	lat	long
0	en	1	0	0	0	0	0	0	0	0	1	0	3	4	INDIA	78.667743	22.351115
0	en	1	0	1	0	0	0	0	1	0	1	0	3	4	INDIA	78.667743	22.351115
0	en	1	0	1	0	0	0	0	0	0	0	0	3	3	INDIA	78.667743	22.351115
0	en	0	0	1	0	0	0	0	0	0	0	0	3	3	INDIA	78.667743	22.351115

Figure 4: The head outlook of the last few features of the curated COVID-19 Tweets Data

Apart from this curated dataset, we also used the challenging SemEval - 2018 - Task 1 data, which contained the same emotion wheel labelling for an extensive set of disconcerting tweets in the face of emotional analysis. The dataset represents a considerably good quality of Affect in Tweets – the dataset was curated by earlier researchers with the specific goal of multi-labelling the 11 emotions on the same wheel scale. We did not carry out any pre-processing on this data, as its curative authors suggested it be used as it is for achieving better performance on challenging tasks – a theme around which this data is centered quite significantly. The argument presented forth was that realistic data fed into realistic systems does not come pre-processed and heavily edited, with which we did agree

From this dataset of tweets with emotions, tweets of roughly 10, 983-records were randomly selected for our models to thus be congruently trained (6938 samples), validated (886 samples) and tested (3259 samples) on this labelled data. A subset of the data utilized from this dataset is shown in Fig. 5 below.



```
ID        Tweet    anger   anticipation    disgust fear    joy     love    optimism        pessimism       sadness surprise        trust
2017-En-21441   "Worry is a down payment on a problem you may never have'.  Joyce Meyer.  #motivation #leadership #worry       0       1       0       0
2017-En-31535   Whatever you decide to do make sure it makes you #happy.        0       0       0       0       1       1       1       0       0       0
2017-En-21068   @Max_Kellerman  it also helps that the majority of NFL coaching is inept. Some of Bill O'Brien's play calling was wow, ! #GOPATS         1
2017-En-31436   Accept the challenges so that you can literally even feel the exhilaration of victory.' -- George S. Patton    0       0       0       0
2017-En-22195   My roommate: it's okay that we can't spell because we have autocorrect. #terrible #firstworldprobs      1       0       1       0       0
2017-En-22190   No but that's so cute. Atsu was probably shy about photos before but cherry helped her out uwu  0       0       0       0       1       0
2017-En-20221   Do you think humans have the sense for recognizing impending doom?      0       1       0       0       0       0       0       1       0
2017-En-22180   Rooneys fucking untouchable isn't he? Been fucking dreadful again, depay has looked decent(ish)tonight  1       0       1       0       0
2017-En-41344   it's pretty depressing when u hit pan on ur favourite highlighter       0       0       1       0       0       0       0       0       1
```

Figure 5: The head outlook of the features of the SemEval - 2018 - Task 1 dataset

4. Methods

Initially, we had planned to approach the problem of classifying multiple emotions for any given tweet solely on the basis of a bidirectional LSTM model, as it is considered a baseline standard for analyzing and classifying sentiments – an established rarity when it comes to multi-labelling, but not attempted commonly in the domain of tweets sentiment analysis. This decision was suggestively influenced by our choice of paper presentation [4] as it had implemented a similar methodology and claimed to beat the industry standard by achieving an accuracy of 59%. However, as pointed out by our course instructor, the paper implementation was relatively new and published to a lenient paper conference (MDPI). Given these inhibitions, he encouraged us to look for other methodologies and studies that were more established and known to provide substantial value. He also pointed out that it was hard to gauge how the paper's methodology was able to achieve the classification without using any underlying Bag-of-Words technique, as it is a popular attempt at natural language processing with sentiment analysis as the goal.

Given these scenarios, we researched more extensively and found a better alternative in the form of the BERT model. However, it was only deemed a superior competitor after analyzing the training and testing results on both of the datasets (this is touched upon in the 'Experiments' section). Additionally, we also carried on with implementing the bidirectional LSTM model and – as a base comparator – the BOW model in order to comprehend which model would fit to the problem better and bring forth more genuine results. To test the efficiencies of each of these implementations, it was imperative that some form of ground truth be correlated with in order to sustain credibility, thus we decided on the human annotation of the accumulated COVID-19 Tweets Dataset with the specific labels defined earlier.

Each model was implemented based on a set of individual motivating factors. The BOW model represents the most rudimentary state-of-the-art guidelines for feature representation in multi-label classification and has always been considered effective in terms of a 'general' methodology while analyzing sentiments, as shown in Fig, 6. However, it is limited in the way of its structure to carry out maximally performing emotion analysis, as it does not take into account the word order and context in its proceedings. Thus, it can never exceed and outperform – in terms of precision and recall – more advanced and sophisticated attention-based structures such as the LSTM and BERT model.

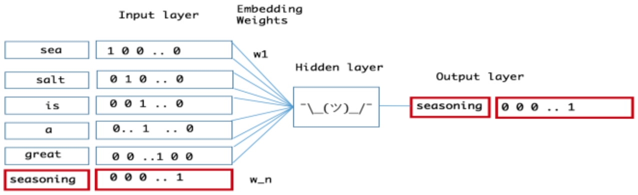


Figure 6: The structure of a typical BOW model used for classification purposes [8]

The LSTM model has always been more dedicated to the task of classifying sentiments in comparison to the BOW model and other Machine Learning methodologies such as the Naïve Bayes Classifier, with guaranteed better results and dedicated attention to word order, vectorization and context. Perhaps most importantly it takes care of label correlation and inter-dependency due to its nature of preserving history and sequence of words. The LSTM approach reads text sequentially and stores relevant information to the task at hand, such as plurality, gender, negation and so on. Fig. 7 below shows how an LSTM construction makes use of the words ingested to classify it into multiple labels when detection occurs. Though results are better than most implementations focused on sentiment analysis, it is usually done in conjunction with data of singular emotion labels topology rather than multiple. If explored more frequently and fine-tuned to the requirements of each individual applicable problem, it is likely to provide grander results than ones achieved in recent researches.

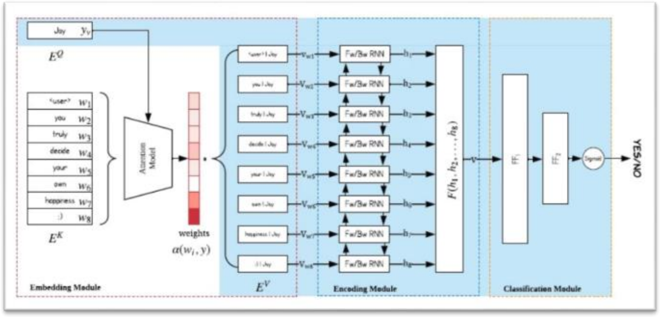


Figure 7: The outline of how a LSTM model ingests input text and identifies with the output labels during training [4]

Alas, the BERT model comes pre-trained for unsupervised tasks such as masked language modeling and next sentence prediction, and along with these major advancements it also happens to be a swifter bidirectional model based on its integrated transformer architecture which has slowly overtaken the standards set by previous techniques of transfer learning. The most significant characteristic of the BERT model, however, happens to be its replacement of the sequential nature of the LSTM structure with a much faster Attention-based strategy. It indeed did perform far better out of all three of the models and expectantly – if trained on higher quality and quantity of data – could result in even better correlative outcomes than the current peak of 87% accuracy. The following example diagram shown in Fig. 8 shows a nice estimate of what the architecture of our BERT model looks like.

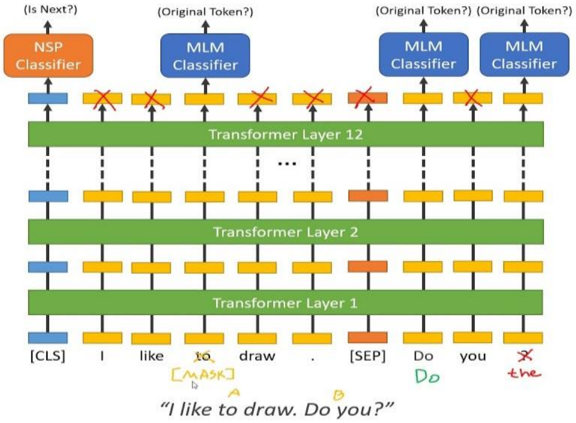


Figure 8: The BERT model while classifying tweets text [9]

To the tokenized input, the inherent Masked Language Model (MLM) is applied in the modeling stage. In Masked LM, some tokens are masked, and then predicted by the model. Next Sentence Prediction (NSP) is applied the sequentially applied, where the model learns to predict the next sentence thanks to the model's unsupervised nature, which greatly increases the understanding of sentence relationship. For our input data, firstly tokenizer was created so that input data could be fed in the tokenized manner. Using tokenization, features are then generated from input data, which is then input to the culminating stage of the BERT model. This feed-forward action creates 3 types of embedding: token embedding, segment embedding, and position embedding to further cycle each such embeddings to the classifiers stage for prediction.

5. Experiments

All of our model implementations were attempted with both sets of data for comparative and performance evaluation purposes. The BOW model was implemented using 'PyTorch,' the LSTM model was implemented using 'Keras,' while the BERT model was implemented in 'TensorFlow' (in which it performed too slow) and PyTorch (in which it performed faster) both.

For the Bag-of-Words configuration, we implemented a configuration using a Multi-layer Perceptron (MLP) Classifier through the 'scikit-learn' Python module, as it aids in distinguishing data that is not linearly separable.

For the LSTM configuration, we implemented a 1-D Convolution and Max Pool Layer with Sequentially Bidirectional dense input embedding and configurations, along with the option of dropout to optimize performance.

For the BERT transformer configuration, we used the pre-trained configuration of BERT-base uncased with a learning rate of 0.00003 and batch size of 20 for model training. To fine tune the model, 110/199 BERT encoding layers were unfrozen but it did not result in any improvements in performance as shown in Table 2. The numbers of epoch were found to be unrelated to performance – increasing or decreasing them did not aid in performance. The Adam Optimizer was used with linear decay setting for this configuration. The singularly trained classification layer was changed from softmax to sigmoid, as it provides the desired 0 – 1 loss that we require along with mimicking the softmax functionality for outputs.

In order to enhance the results of our models different tasks were performed by changing the learning rate, batch size, initial seed, and number of unfreeze encoding layers. We also utilized two loss functions meant for multiple labels – Binary Cross Entropy (BCE) with Logits Loss and Focal Loss – in order for these experiments to understand and differentiate between the performative statistics.

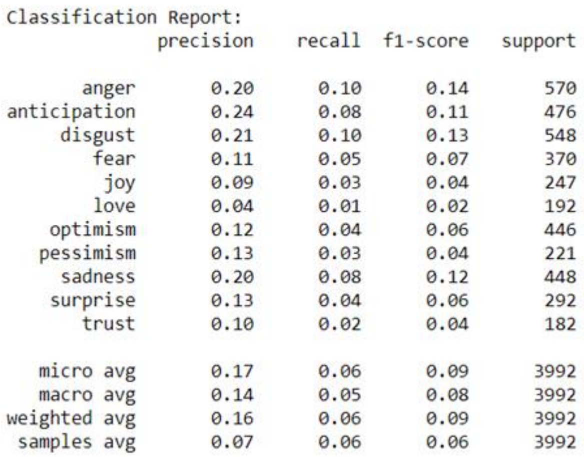
The results of our implementations are given in the Tables 2, 3, 4 and 5, along with Figures 9a, 9b, 9c, 10a, 10b and 10c respectively with regards to each dataset.

	Jaccard Index	Precision	Recall	F1 Score
BOW	0.109	0.166	0.063	0.092
LSTM	0.121	0.152	0.205	0.175
BERT	0.379	0.806	0.391	0.526

Table 2: Results of Evaluation Metrics for our implementations on our COVID-19 Tweets Dataset

	Accuracies	BCE with Logits Loss
BOW	0.176	0.65
LSTM	0.548	0.443
BERT	0.873	0.370

Table 3: Results of Accuracy and Loss on our COVID-10 Tweets Dataset

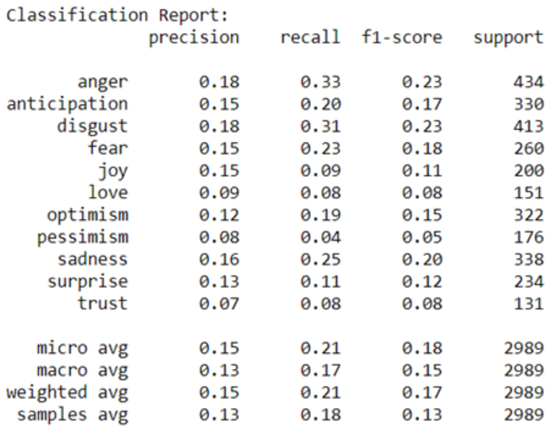


```
Classification Report:
              precision    recall  f1-score   support

       anger       0.20      0.10      0.14       570
anticipation       0.24      0.08      0.11       476
     disgust       0.21      0.10      0.13       548
        fear       0.11      0.05      0.07       370
         joy       0.09      0.03      0.04       247
        love       0.04      0.01      0.02       192
    optimism       0.12      0.04      0.06       446
   pessimism       0.13      0.03      0.04       221
     sadness       0.20      0.08      0.12       448
    surprise       0.13      0.04      0.06       292
       trust       0.10      0.02      0.04       182

   micro avg       0.17      0.06      0.09      3992
   macro avg       0.14      0.05      0.08      3992
weighted avg       0.16      0.06      0.09      3992
 samples avg       0.07      0.06      0.06      3992
```

Figure 9a: A detailed classification report for the BOW implementation on the emotion labels of our COVID-19 Tweets Dataset

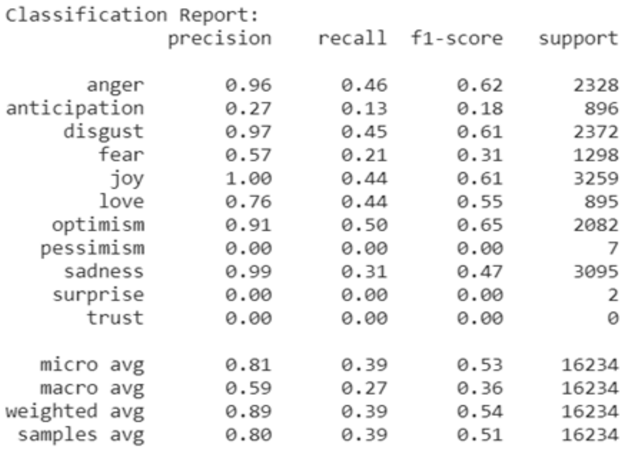


```
Classification Report:
              precision    recall  f1-score   support

       anger       0.18      0.33      0.23       434
anticipation       0.15      0.20      0.17       330
     disgust       0.18      0.31      0.23       413
        fear       0.15      0.23      0.18       260
         joy       0.15      0.09      0.11       200
        love       0.09      0.08      0.08       151
    optimism       0.12      0.19      0.15       322
   pessimism       0.08      0.04      0.05       176
     sadness       0.16      0.25      0.20       338
    surprise       0.13      0.11      0.12       234
       trust       0.07      0.08      0.08       131

   micro avg       0.15      0.21      0.18      2989
   macro avg       0.13      0.17      0.15      2989
weighted avg       0.15      0.21      0.17      2989
 samples avg       0.13      0.18      0.13      2989
```

Figure 9b: A detailed classification report for the LSTM implementation on the emotion labels of our COVID-19 Tweets Dataset



```
Classification Report:
              precision    recall  f1-score   support

       anger       0.96      0.46      0.62      2328
anticipation       0.27      0.13      0.18       896
     disgust       0.97      0.45      0.61      2372
        fear       0.57      0.21      0.31      1298
         joy       1.00      0.44      0.61      3259
        love       0.76      0.44      0.55       895
    optimism       0.91      0.50      0.65      2082
   pessimism       0.00      0.00      0.00         7
     sadness       0.99      0.31      0.47      3095
    surprise       0.00      0.00      0.00         2
       trust       0.00      0.00      0.00         0

   micro avg       0.81      0.39      0.53     16234
   macro avg       0.59      0.27      0.36     16234
weighted avg       0.89      0.39      0.54     16234
 samples avg       0.80      0.39      0.51     16234
```

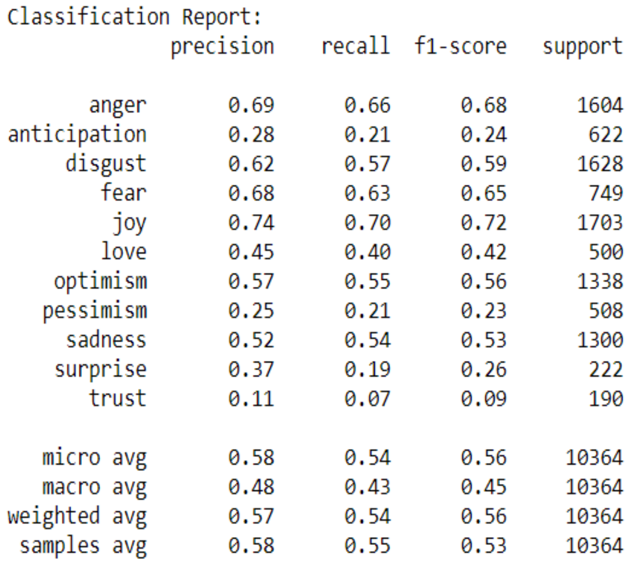
Figure 9c: A detailed classification report for the BERT implementation on the emotion labels of our COVID-19 Tweets Dataset

	Jaccard Index	Precision	Recall	F1 Score
BOW	0.430	0.584	0.538	0.560
LSTM	0.367	0.403	0.591	0.479
BERT	0.370	0.806	0.391	0.526

Table 4: Results of Evaluation Metrics for our implementations on the SemEval – Task 1 Dataset

	Accuracies	Focal Loss
BOW	0.146	0.655
LSTM	0.679	0.630
BERT	0.828	0.413

Table 5: Results of Accuracy and Loss on the SemEval – Task 1 Dataset

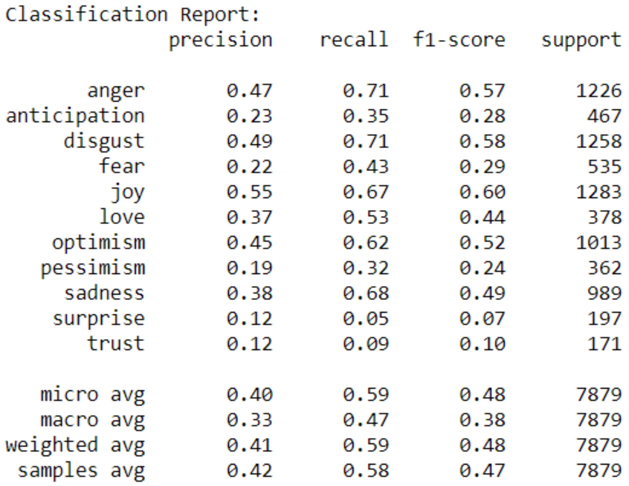


```
Classification Report:
              precision    recall  f1-score   support

       anger       0.69      0.66      0.68      1604
anticipation       0.28      0.21      0.24       622
     disgust       0.62      0.57      0.59      1628
        fear       0.68      0.63      0.65       749
         joy       0.74      0.70      0.72      1703
        love       0.45      0.40      0.42       500
    optimism       0.57      0.55      0.56      1338
   pessimism       0.25      0.21      0.23       508
     sadness       0.52      0.54      0.53      1300
    surprise       0.37      0.19      0.26       222
       trust       0.11      0.07      0.09       190

   micro avg       0.58      0.54      0.56     10364
   macro avg       0.48      0.43      0.45     10364
weighted avg       0.57      0.54      0.56     10364
 samples avg       0.58      0.55      0.53     10364
```

Figure 10a: A detailed classification report for the BOW implementation on the emotion labels of the SemEval – Task 1 Dataset

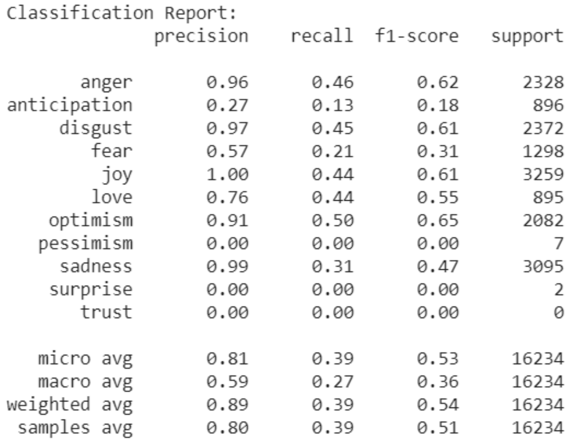


```
Classification Report:
              precision    recall  f1-score   support

       anger       0.47      0.71      0.57      1226
anticipation       0.23      0.35      0.28       467
     disgust       0.49      0.71      0.58      1258
        fear       0.22      0.43      0.29       535
         joy       0.55      0.67      0.60      1283
        love       0.37      0.53      0.44       378
    optimism       0.45      0.62      0.52      1013
   pessimism       0.19      0.32      0.24       362
     sadness       0.38      0.68      0.49       989
    surprise       0.12      0.05      0.07       197
       trust       0.12      0.09      0.10       171

   micro avg       0.40      0.59      0.48      7879
   macro avg       0.33      0.47      0.38      7879
weighted avg       0.41      0.59      0.48      7879
 samples avg       0.42      0.58      0.47      7879
```

Figure 10b: A detailed classification report for the LSTM implementation on the emotion labels of the SemEval – Task 1 Dataset



```
Classification Report:
              precision    recall  f1-score   support

       anger       0.96      0.46      0.62      2328
anticipation       0.27      0.13      0.18       896
     disgust       0.97      0.45      0.61      2372
        fear       0.57      0.21      0.31      1298
         joy       1.00      0.44      0.61      3259
        love       0.76      0.44      0.55       895
    optimism       0.91      0.50      0.65      2082
   pessimism       0.00      0.00      0.00         7
     sadness       0.99      0.31      0.47      3095
    surprise       0.00      0.00      0.00         2
       trust       0.00      0.00      0.00         0

   micro avg       0.81      0.39      0.53     16234
   macro avg       0.59      0.27      0.36     16234
weighted avg       0.89      0.39      0.54     16234
 samples avg       0.80      0.39      0.51     16234
```

Figure 10c: A detailed classification report for the BERT implementation on the emotion labels of the SemEval – Task 1 Dataset

It can be seen that the BERT model performed the best out of all the implementations in terms of the results of the evaluations metrics, accuracies and loss functions on both the datasets. It is also to be noted that the results of the models are far better on the SemEval – Task 1 data due to its curative nature – the list of annotators and expertise for this specific data over a range of years is not comparable to our compilation and curation procedures of mere weeks, but it is a start. In analyzing the above obtained results, we can see that negative emotions such as anger, anticipation, disgust and sadness were better classified by the models on our data, whereas the emotions anger, disgust, joy, love, optimism, and sadness performed better on our models using the SemEval – Task 1 Data.

To gain further insight into how of each these models worked based on predictions and the consumed data, we also graphically exploited these notions with tools from Python and Tableau in order to inspect what feature related to what other features, and how their relationships impacted the overall imprint of the data and model. Fig. 11 below shows the occurrence rate of each emotion label in the predictions of the BERT model, reiterating the earlier assumption that negative emotions were more frequent on the spectrum than others.

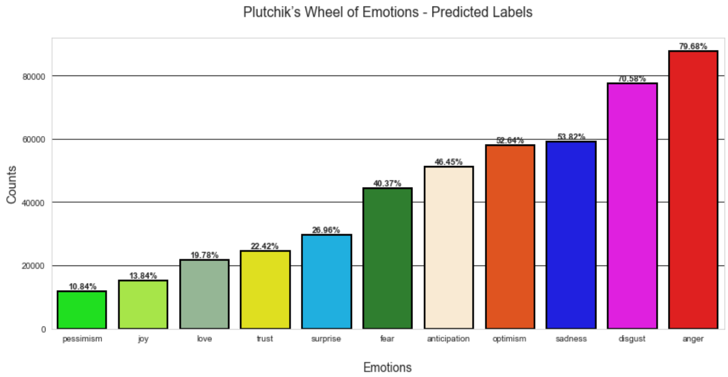


Figure 11: Occurrence of Emotion Labels in the BERT model.

Additionally, we also checked the most frequent location areas where emotion counts of all labels was more than 60% in terms of occurring coverage in order to understand which locations were seemingly providing more towards the cause of the problem being attempted, as presented in Fig. 12.

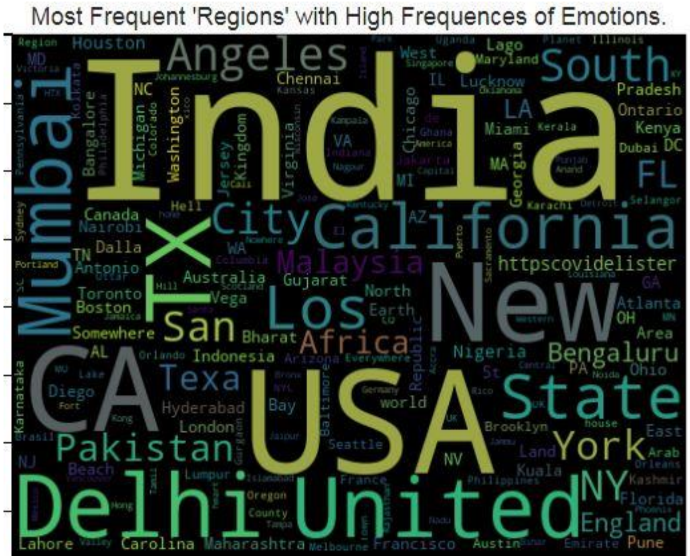


Figure 12: Frequent locating regions that appear to directly affect the predicted emotions.

Similarly, when plotting the correlation matrix of the predicted labels as shown in Fig. 13, it could be seen that emotions that naturally fall in to their individual spectrum of positive and negative emotions have higher correlation, depicting that the prediction is indeed reliable.

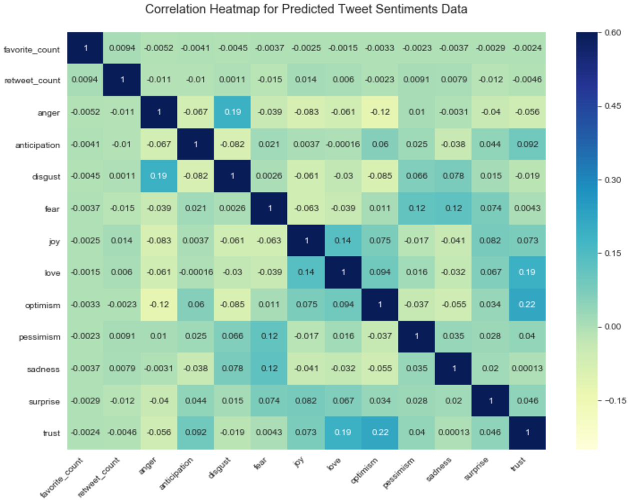


Figure 13: Correlation Matrix of the Predicted Emotion Labels from the BERT model.

We also found out while conducting our visual analysis of the COVID-19 dataset that a major emotion – fear – appeared to exist as a recurrent word in the tweets text, signaling that it would be likely a major and frequent label in the prediction process (which it was proven to indeed be). This was discovered while extracting the top 10 most common words present in the accumulated tweets data from the tweet text attribute, depicted in Fig. 14.

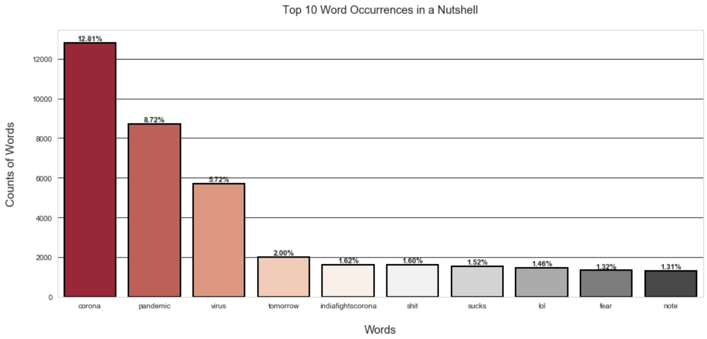


Figure 14: Top 10 Word Occurrences in the tweets of the COVID-19 Tweets Dataset.

In order to map out the trend of emotions of different countries with respect to different dates and months, we set about visualizing interpolated and non-interpolated real-time interactive plots to measure these ideas competently. We also plotted these concepts without locational restrictions to gain a time-series analysis of the entire emotion tags as well. Furthermore, we made subsets of these ideas into positive and negative emotions on the same scale of dates for a spectrum comparison. These methods can be seen in Figs. 15a, 15b, 16a, 16b, 16c, 17a and 17b.

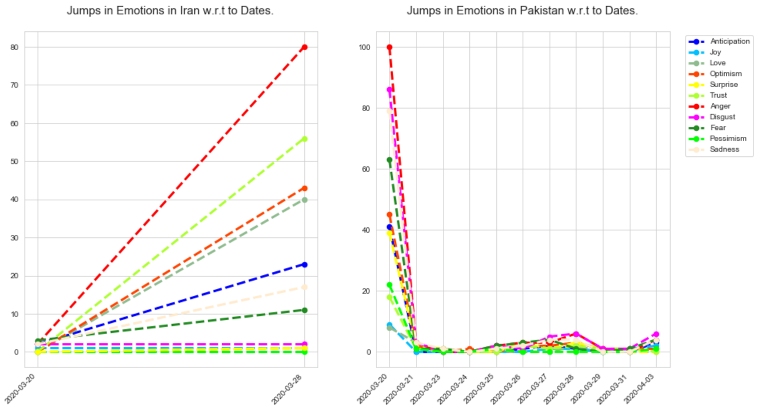


Figure 15a: Comparison of Iran and Pakistan with respect to each emotion and event of date present in the COVID-19 Tweets Dataset

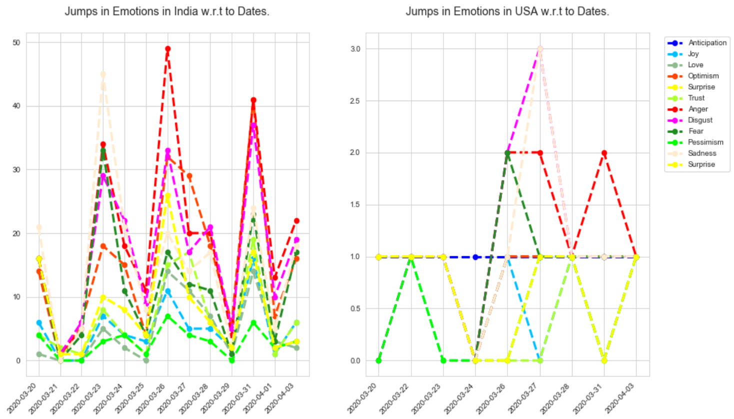


Figure 15b: Comparison of India and USA with respect to each emotion and event of date present in the COVID-19 Tweets Dataset

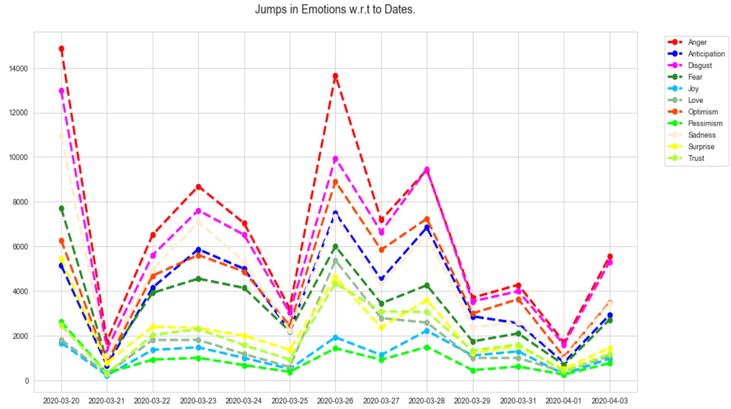


Figure 16a: Comparison of each emotion overall relevant to dates present in the COVID-19 Tweets Dataset

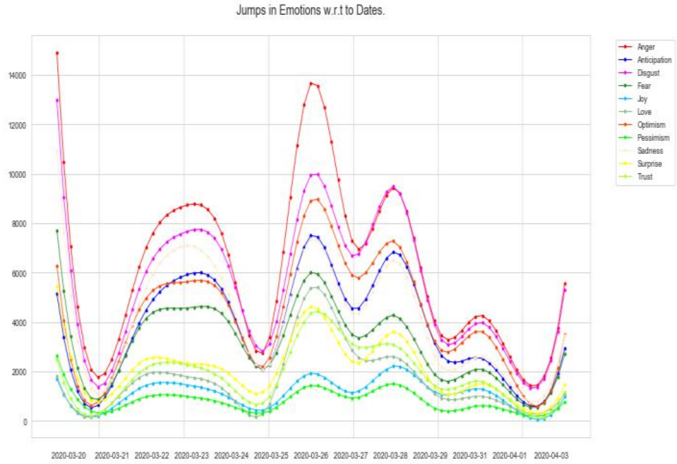


Figure 16b: Interpolated comparison of each emotion overall relevant to dates present in the COVID-19 Tweets Dataset

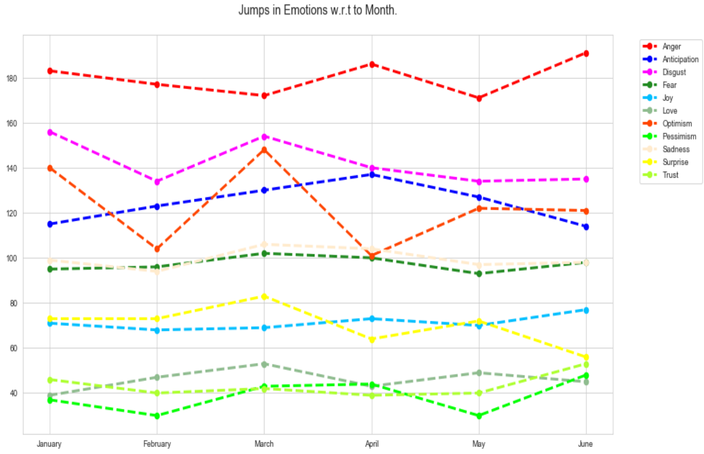


Figure 16c: Comparison of each emotion overall relevant to months present in an earlier scrap of the COVID-19 Tweets Dataset

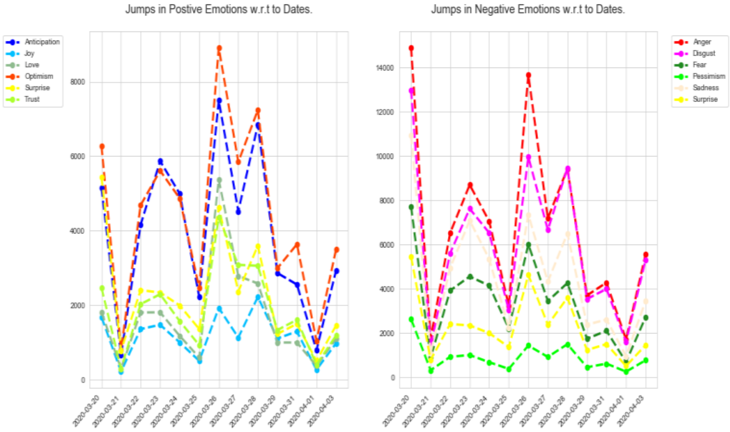


Figure 17a: Comparison of each positive emotion with each negative emotion relevant to dates present in the COVID-19 Tweets Dataset

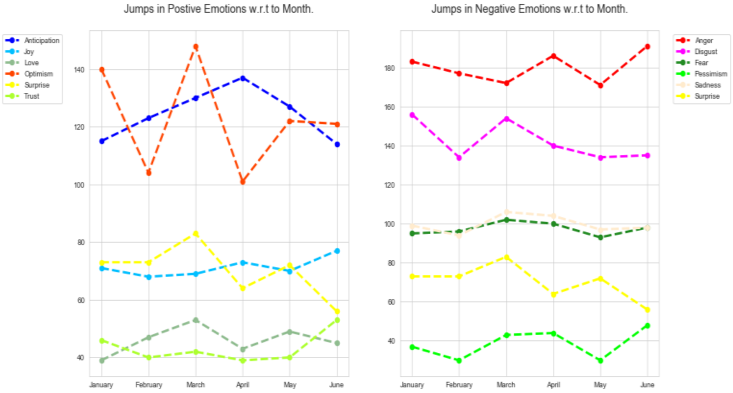


Figure 17b: Comparison of each positive emotion with each negative emotion relevant to months present in an earlier scrap of the COVID-19 Tweets Dataset

To comprehend how distributed the emotions were globally, and which emotion dominated these areas, we also plotted emotion counts with reverence to their longitude and latitude to see what areas contributed to what sentiment clusters. Fig. 18 below is a remnant of the entire global map, in which it can be seen that emotions of fear and pessimism (represented by the green and blue color scheme) are rife in the Middle East, correlating to the increased number of cases and infections resulting from the COVID-19 outbreak. It is accurate to assume that emotions will range higher in areas where social communication is common but with an exponential escalation of cases.

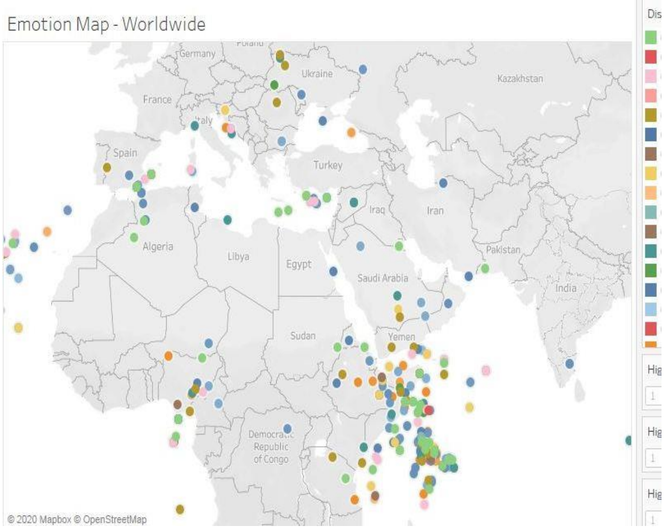


Figure 18: A fragment of the World Map depicting clusters of emotions with intensity

6. Conclusion

Based on our model implementations, performances and analysis, we were able to create a competent and optimized system that detects and predicts numerous emotions with an accuracy of 87% at best for now. If optimized with better data collection, processing and ensured emotion labelling, it can be expected to provide even better results. We are interested in exploring further reasoning of why the BERT was able to oust the LSTM model, when by convention it is expected by an LSTM model to perform better. With rare implementations of such a case study, it was hard to comprehend and understand these differences fully but our guess is that using a pre-trained model with fine-tuning and ground truth of labels for an otherwise unsupervised task may have led to this performance.

Any result or visual analysis of this project is mean t to be taken with the underlying fact that the data ingested for this process is limited – practically data of this nature is real-time and of a streaming nature. We cannot expect to beat or even reach it realistically without the usage of GPUs (which were hard for us to utilize but was managed sparsely). Thus, results are prone to this data directly and can be hence improved with more qualified data. With more data, we may be able to map our BERT-base model to BERT-large which could hypothetically improve on performance.

Another realization of this study was that a BOW model appears to be, sadly, outdated to the current scenarios of real-time problems. It is a good baseline to work with in beginner friendly terms, and the intrinsic theoretical work at play is certainly being used in sentiment-based analytical models. However, to rely solely on a BOW model to achieve supreme results in the field of sentiment analysis appears to be impractical at best for now.

This setup can be modelled better with the above mentioned enhancements, but for now works at respectable ratios, considering the limited data and resources.

Note: This report is meant for the Deep Learning Course Spring 2020 at Information Technology University (ITU). The coding modules that were utilized were created by the author members alone, apart from the BERT model which has hints of the official coded model tutorials and has thus been referenced accordingly [10].

References

[1] K. H. Manguri, R. N. Ramadhan, and P. R. Mohammed Amin, “Twitter Sentiment Analysis on Worldwide COVID-19 Outbreaks,” April 2020.

[2] J. Samuel, G. Nawaz, M. Rahman, E. Esawi, and Y. Samuel, “COVID-19 Public Sentimental Insights and Machine Learning for Tweets Classification,” June 2020.

[3] I. Li, Y. Li, T. Li, S. Alveraz-Napagao, D. Gracia and T. Suzumura, “What are We Depressed about When We Talk about COVID19: Mental Health Analysis on Tweets using Natural Language Processing,” June 2020.

URL: https://arxiv.org/abs/2004.10899

[4] M. Jabreel and A. Moreno, “A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets”, March 2019.

[5] H. Jelodar, Y. Wang, R. Orji and H. Huang, “Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach,” April 2020.

[6] G. Ma, “Tweets Classification with BE11RT in the Field of Disaster Management,” 2019.

[7] https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset

[8] https://medium.com/@gunjanagicha/word-embeddings-ee718cd2b8b5

[9] https://mccormickml.com/assets/BERT/padding_and_mask.png

[10] https://github.com/huggingface/transformers