# Explainable AI via Exhaustive Decision Tree Querying on Mental Health Tech Survey Data

Fizza Tauqeer

*Department of Artificial Intelligence,*
*CureMD, Pakistan*

fizzatauqeer95@gmail.com

## Abstract

*Explainable Artificial Intelligence has slowly gained momentum in the field of research, with many studies extensively touching upon the need of such Knowledge Bases in the exponentially growing arena of machine learning.*

## 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 decisions in relation to everyday events happening in our periphery. Unfortunately, one such event currently taking over these networking applications happens to be COVID-19 - a novel type of contagious coronavirus that attacks the host organism's respiratory system leading to a declining state of health or complete breakdown - which began to illustriously appear in late November of 2019 and has since become a global pandemic in a mere matter of 4 months, lethally wreaking havoc even more so an year later. Many countries have gone into severe and frequent lockdowns since the disease's onslaught to contain transmission for the safety of their nationals. The swift spread and deadly effects of this contagion, along with these lockdowns, have naturally appeared to significantly affect numerous individuals mentally and physically, with primary hypotheses suggesting social confinement, ambush of unemployment, lack of consumer resources and the impeding expectation of an increasing financial crisis worldwide. Consequently, a huge number of social media users are intentionally and unintentionally actively expressing their experiences, thoughts and feelings on various social media platforms, particularly Twitter.

With this massive flow of continuous information, it is imperative to gauge, through the help of these outlets, 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 in how severe solutions or treatments are needed or required, 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 dealing with this crisis. Similarly, it can be anticipated that virtual shopping platforms will improvise and push their platforms given the exponential rise in online shopping in these current times. Such significant mappings were heavily considered as the intended problem is approached.

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 emotion spectrum, this helps in gauging the effect of sentiments and their trends with respect to COVID-19 quite successfully.

This 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. The final solution is meant to be based on a thorough study of the optimized algorithms utilized, the motivation behind each selection and what interventions must be made to improve existing results (such as any experimentations on the metrics and models used or combined), while bringing originality. The system is also focused with ensured data from countries such as Pakistan, India, USA, Italy and Iran as they have been proven to be hit the hardest from COVID-19 and thus can naturally provide a more realistic impact in data and sentiments in tweets. For this purpose, model testing for sentimental analysis of multiple labels will be done initially in order to derive prized graphical conclusions based on numerous emotions. Upon obtaining these results, one can move towards taking advantage of

locational and time-series data identifiers for analytical purposes, all of which will be covered in the 'Results and Conclusion' section. It is also intended to provide a comparison between this set of implementations and previously carried out research alongside this work in order to certify this attempt's standing in the overall approach of sentiment analysis and classification, especially to COVID-19.

Using the methods outlined in the 'Methodology and Experiments' sections, the intent is to bridge the gaps present in studies on such pivotal areas of COVID-19. In the perusal of exploration of studies for the 'Literature Review' section, one is able to quickly understand that research areas were primarily focused on predicting COVID-19 hotspots – along with physical ailments and symptoms – as the pandemic spread more, rather than on equally important factors such as its earlier transmission, swift transcendence and long-lasting effects on emotional health. With this distinct and quite present investigation gap in mind, implementing an analytical system was targeted which could serve as a potential aspect of reliable research yet to be done, while retaining the plausible opportunity of being positively exploited in futuristically predictable social health systems and solutions.

## 2. Literature Review

In order to acquaint myself with how the problem of multi-label sentiment classification had been attempted at in previous literature and research articles, multiple related research works were analyzed which focused on predicting sentiments of Twitter users both theoretically and practically. During this process, it was noticed that the problem of sentiment analysis being addressed specific to COVID-19 had begun to take significant momentum, strongly implying that it was a exponentially growing area of upcoming research. The works mentioned henceforth inspired the designated 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 Naïve-Bayes framework [1]

This study guided me in the selective strategy of scrapping tweets data from Twitter based on multiple emotions, as otherwise I was initially focusing on positive, negative and neutral sentiments by way of this research's categorization, which cannot successfully correlate to a proposed method of various sentiment identifiers. Nevertheless, this research was able to gauge meaningful statistics, such as that of 36% of people holding positive sentiments, while 14% harbored negative views. However, it is to be noted that the study was conducted in April of 2020 when COVID-19 was not being considered as a major threat to health and well-being world-wide and is based on probabilistic methods of computation, 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 it can be assumed that their accuracy on long tweets may have lacked due to the absence of more in-depth modules that Deep Learning can provide in comparison to holistic Machine Learning functionalities.

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

Irene et al. [3] used a pre-trained BERT (Bidirectional Encoder Representations from Transformers) 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 the prime motivation in including the BERT model for the project's classification task, as prior to it the sole focus was on implementing a baseline bidirectional LSTM (Long Short-Term Memory) network with pre-trained global vectors (GloVe) of Twitter embeddings due to the influence of a research titled "A Deep Learning-Based Approach for Multi-Label Emotion Classification in Tweets" [4]. However, due to the idea and usage of Deep Attention models being relatively untested and unexplored, the credibility of this specific research was unreliable, thus providing the motivation and opportunity to explore other model ventures. It also served as an explanatory study 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, such as which location has the most fear-based reactions to the pandemic. A prime example of such insights can be

seen in Fig. 2 below.

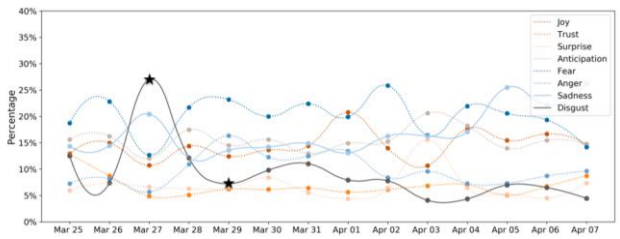



Figure 2: Emotion Trend on the word 'lockdown' from March 25 to April 7, 2020 [4]

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 healthy comparison to add in this section of related work due to the envisioned goal of using a Bidirectional LSTM for breaking down tweet text sentence structures for emotion consumption.

While exploring the decision to utilize both BERT and LSTM modules for this project, an interesting study [6] was come across 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 embedding. 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 clarified a lot of my 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 as well – implying that in order for a successful implementation to take place, there would certainly be trade-offs involved. To decide on what to compromise, it must be imperative as to what certain goal one would want to achieve with each customized implementation. Hence, I decided to leave this project implementation open-ended as a starting methodology which could later be transformed to scenario or context-based executions.

Perhaps most persuasive and important to the cause and effect of this problem is a study focusing on using tweets information to visualize and study subjectivity, alongside modelling emotions brought on by the pandemic. Kabir et al. [7] developed a real-time based application to observe these elements in tweets of residents of USA. The researchers were able to generate several analytics over certain time periods to study the outlined and expected changes with moderation. Though the study did not rely on any modulation or framework to achieve this cause, they were able to exclusively exploit the data relevant to COVID-19 alone and extract meaningful insights that could be further used by a variety of communities and origins to aid in the fight against coronavirus. This research inclined us to exploit the curated and scraped data as well in order to bring about visual insights prevalent in data.

Even though COVID-19 is a harvesting topic of viable research, there is still a lack of study and experimentation with its effect on mental outlook analyzed through social media, especially Twitter. This extensive study was able to yield only this literature relevant to the topic, but with the work influenced through these means, the aim is to add to this scarce pool of resources while providing relevancy and reliability.

## 3. Data Collection and Processing

The goal is to utilize a total of two datasets for this problem, in order to be able to gauge the rate of success and efficiency which each set of data ingestion into the models would provide. Though there is reliance on one already available dataset and one self-scrapped dataset, human labelling was initiated on the scrapped dataset as a data construction strategy for result verifications in the latter half of this research.

The dataset of COVID-19 relevant tweets which is planned to be consumed in my models and experiments was gathered using the theoretical functionalities of Twitter Streaming Application Programming Interfaces (APIs) along with Lookup APIs for Tweet Identifiers (for ID Extraction) in Python language. For this research, Lookup APIs were used specifically to get precise historical tweets related to the coronavirus strain. With regards to locational relevancy, Tweet IDs of 2, 40, 070 tweets (in the aim to represent a realistic approach) were globally extracted from the real-time Coronavirus Tweets Dataset available at IEEE Data Port [8] – the additional sentiments of positive/negative nature provided alongside it was discarded, as my problem was not modelled along this binary sentiment lineage. This data port monitors the real-time Twitter feed for coronavirus-related tweets using more than 90 different keywords and hashtags that are commonly used while referencing the pandemic, while complying with the content redistribution policy set by Twitter. The open-sourced Tweet IDs were extracted by the IEEE Data Port system on the following active keywords and tags: "corona", "coronavirus", "covid", "covid19", "covid-19", "sarscov2", "sars cov2", "sars cov 2", "covid_19", "ncov",

"ncov2019", "2019-ncov", "pandemic", "2019ncov", "quarantine", "flatten the curve", "flattening the curve", "hand sanitizer", "lockdown", "social distancing", "work from home", "working from home", "ppe", "n95", "covidiots", "herd immunity", "pneumonia", "chinese virus", "wuhan virus", "kung flu", "wearamask", "wear a mask", "vaccine", "vaccines", "corona vaccine", "corona vaccines", "face shield", "face shields", "health worker", "health workers", "stay home stay safe", "corona update", "frontline heroes", "corona warriors", "home school", "home schooling", "home tasking", "masks4all", "wfh", "wash ur hands", "wash your hands", "stay at home", "stay home" and "self-isolating."

This dataset held multi-lingual tweets, but for my experiments I mined and concentrated on subsets of English tweets only, dated from March 21, 2020 to April 03, 2020. Data was collected based on this timeline in order to bring an element of ingenuity, as the literature review conducted earlier showed that studies appeared more focused towards the peak of the pandemic (such as in the months of June and July) rather than its onslaught and early trajectory around the world.

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 the process was 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 – one could iterate over a size of simultaneously-occurring 8200 iterations and batches in a singular run. I 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 is defined in terms of significance in Table 1. as follows:

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

| | | |
|---|---|---|
| tweet_id | user_id | Date |
| favourite_ count | retweet_count | hash tags |
| Time | tweet_text | is_quoted |
| symbols | Language | Location |

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 self-built dedicated Python-based pre-processor. This reduced the data size (and possibly the 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 actual tweets text attribute. 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 - using Python - for the purpose of Human Labelling of the emotions identified in these tweets by a sample of 39 annotators. This sample of annotators consisted of students studying Economics with Data Science as their Bachelor's degree at Information Technology University (ITU), Lahore, Pakistan. They were chosen on the basis of being informed enough in regards to the effects of Data Science on large-scale world problems, thus they can be hypothetically assumed as reliant to tag these tweets on the basis of its inherent emotion.

The plan is to use this accumulated, refined and labelled data for the predictive models in terms of Train Set (4937 samples), Validation Set (1064 samples) and Test Set (1085 samples). The 11 emotions which were labelled 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 and were already present in the SemEval Data (the other dataset which is planned to be used). A subset of the data curated is shown in Figs. 3 and 4 in order to provide an understanding of what the final data looks 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 there was on hand. More accurate approximations and results can be achieved if the volume of data scrapped is increased, tagged and later tested. Thus, the work with all such datasets was done 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 | 1243054659991489024 | 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, the challenging SemEval - 2018 - Task 1 data was also used, 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. I 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, which is agreeable.

From this dataset of tweets with emotions, tweets of roughly 10, 983 records were randomly selected for the future 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.

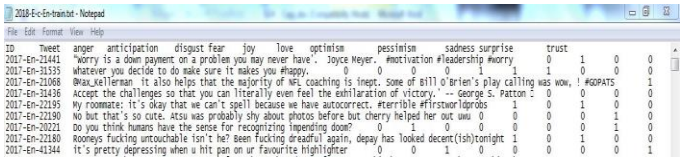



Figure 5: The head outlook of the features of the SemEval - 2018 - Task 1 dataset

The models I plan to engage will output data tagged on the relevant emotions – which they are able to identify – in the format of the curated COVID-19 Tweets Dataset. The evaluation metrics and computations in the following sections rely on these ground truth and predicted datasets.

## 4. Methodology

Initially, I 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. However, due to the Literature Review it was easier to first understand and use other statistical methodologies that were more established and known to provide substantial value as well, achieving superior classification with and without using the traditionally popular method of Bag-of-Words (BOW), which is considered a must in Natural Language Processing for sentiment analysis.

Given these developments, I researched more extensively and found a better alternative in the form of the probabilistic BERT model after the LSTM implementation. However, it will only be accepted as a superior substitute after analyzing the training and testing results on both of the datasets (outlined additively in the 'Results' section). Additionally, I 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 with the sigmoid activation function. To test the efficiencies of each of these implementations, it was imperative to utilize and model on each of the curated and standardized datasets discussed earlier.

Each model was implemented based on a well-versed set of motivated 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 statistical techniques such as precision and recall – more advanced and sophisticated attention-based structures such as the LSTM and BERT model.

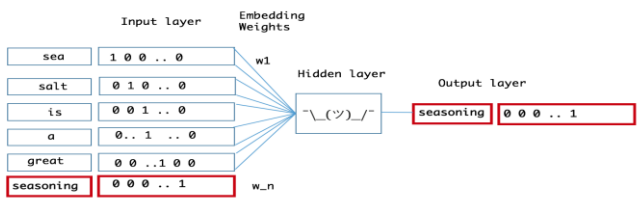



Figure 6: The structure of a typical BOW model used for classification purposes [9]

The LSTM model has always been more dedicated to the task of classifying sentiments when placed 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. Most significantly, 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 striking results than ones achieved in recent researches.

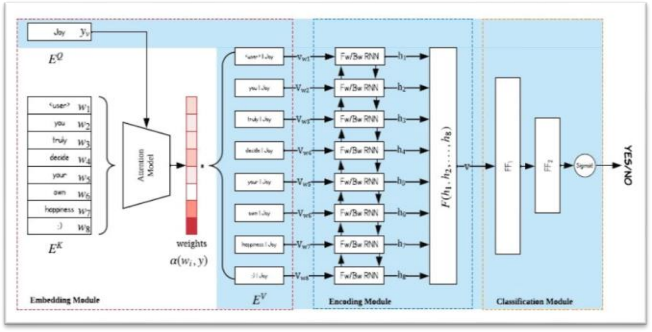



Figure 7: The outline of how a LSTM model ingests input text and identifies with the output labels during training [4]

Conclusively, the BERT model comes moderately 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 has the potential to perform far better out of all three of the models in the evaluation metrics and model performance, and expectantly – if trained on higher quality and quantity of data – could result in even better correlative outcomes. The following example diagram shown in Fig. 8 is the assessment of what the architecture of the BERT model looks like in solving this multi-labelling predictive problem. The goal is to implement this structure for maximal performance gauging.

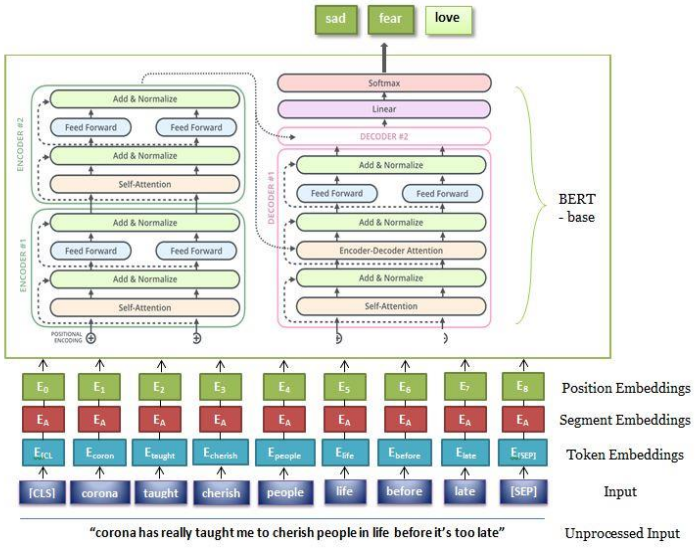



Figure 8: Methodology Diagram – The BERT model while classifying tweets text

To the tokenized input, the inherent Masked Language Model (MLM) is applied in the modelling stage. In Masked LM, some tokens are masked, and then predicted by the model. Next Sentence Prediction (NSP) is sequentially applied, where the model learns to predict the next sentence due to the model's unsupervised nature, which greatly increases the understanding of sentence relationship. For the input data, firstly a probabilistic tokenizer was created so that input data could be fed in the tokenized manner. Using the process of 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 – each such embedding is further cycled to the classifier stage for prediction.

## 5. Experiments and Results

All of the model implementations were attempted with both primary 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 both 'TensorFlow' (in which it performed poorly) and PyTorch (in which it performed faster).

For the Bag-of-Words configuration, I 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 based on statistical configurations such as the t-test.

For the LSTM configuration, I 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, I 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 for the problem being addressed. The Adam Optimizer was used with linear decay setting for this configuration. The singularly trained classification layer was changed from the earlier selected softmax to sigmoid, as it provided a better result for the desired 0 – 1 loss that is required along with mimicking the softmax functionality for selecting outputs, especially in the testing phase.

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

The experimentation results of the 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 the implementations on the 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 the COVID-19 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 the 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 the 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 the 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 the 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

These experiments confirmed the earlier hypothesis that the BERT model would perform the best out of all the implementations, in terms of the evaluation 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 the compilation and curation procedures of mere weeks, but it is a start nonetheless. In analyzing the above obtained trial-based results, it can be seen that experimentations for negative emotions such as anger, anticipation, disgust and sadness were better classified by the models on the data, whereas the emotions anger, disgust, joy,

love, optimism, and sadness performed better on the models using the SemEval – Task 1 Data.

## 6. Analysis and Insights

Prior to embarking on model implementations and experiments, it was important to utilize statistical measures of feature selection. The number of features in both sets of the consumable data was large in quantity and using all features for model training is not a process recommended by data science experts as it is [10]. The aim in using such selection measures is to exploit those variables which have the strongest relationship with the output variable. In this problem, there are multiple such output variables in the form of the emotion attributes; hence there is a need to reiterate the process of feature selection sequentially to keep those features that complement each such attribute individually and conjunctively. As the data was a conglomeration of categorical and non-categorical data equivalently, I employed relevant and applicable techniques for this practice to cater to each data type exclusively. Fig. 11 below shows the results of the Chi-Square statistic measure employed and accordingly ordered on the sample data features. It is to be noted that as the data is sampled, irregularities in feature importance can also occur due to the unbalanced nature of the ever-increasing population data from which it has been extracted.

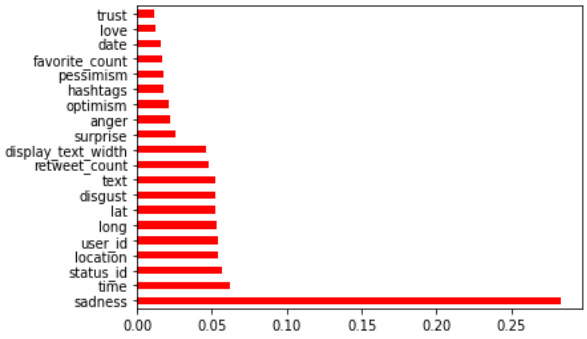



Figure 11: Chi-Square Test Results on the COVID-19 Tweets Dataset

In terms of experimentations that could provide influence for the procedures of visualizing and thus subsequently analyzing the curated and model-tagged dataset, I carried out multiple tests of hypotheses with a 95% confidence interval (computed on the foundation of contingency tables comprising of the attributes for which the aim is to establish and prove a relationship between) based on various aspects that could be meaningful in terms of surveying the sentiments present in these datasets. Four major categories for the hypotheses with regards to the emotion attributes of the data were encountered – the first alternate hypothesis considered was that negative emotions occurred more on the most frequent dates in the dataset (Significance Accepted); the second alternate hypothesis suggestion was that tweets having sentiments of anticipation and hope occurred particularly on a frequent time-stamp in Universal Time Coordinated (UTC) format (Significance Rejected); the third alternate hypothesis was that trusty and optimistic tweets were retweeted more in comparison to other sentiment-tinged tweets (Significance Accepted); the fourth alternate hypothesis was divided into two sub hypotheses - that emotions of surprise were prevalent in India-based Tweets (Significance Rejected) alongside pessimistic tweets being prevalent in America-based Tweets (Significance Accepted). The null hypothesis for each of these cases was naturally the negation and resulting rejection of the alternate hypothesis. This testing procedure was carried out using two forms of Hypothesis Testing in order to accommodate the time-series nature of data while providing a dependable implementation source: the Two Sample Z-Test and the Analysis of Variance (ANOVA) Test (which is based on Linear Regression – also known as Ordinary Least Squares (OLS) – modulations). Both procedures resulted in the same result of hypothesis rejection and acceptance with regards to the null and alternate hypothesis, thus providing a layer of extra assurance on the achieved results.

These claims were finalized based on the context surrounding these hypotheses and can be seen in Tables 6, 7, 8, 9 and 10 below. An archetypal example output for the hypothesis test outlined in Table 6 mapped as an ANOVA Test can also be seen in Fig. 12 below. Tweets with negativity were likely to increase as the number of coronavirus cases increased day by day. Messages of hope and optimism were probable to be spread more due to the depressive state of the pandemic. With India's surrounding neighbours facing almost non-existent cases of COVID-19, it came as a worldwide surprise that cases accordingly surged exponentially in the country itself. Conclusively – with the current state of America's declining measures to stabilize its Public Health, especially its assured provision to their nationals – it was likely that the general populace would feel and express being overwhelmed with the incoming wave of the pandemic. These hypotheses were largely justified and enabled for a more analytical-centered approach to the datasets, as can be seen in the visualizations later on in this section.

| | Accepted | Rejected |
|---|---|---|
| H0 | No | Yes |
| H1 | Yes | No |

Table 6: Test 1: H1 – Fear, Pessimism and Anger on the Rise on Frequently-Occurring Dates

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | fear | R-squared: | 0.006 |
| Model: | OLS | Adj. R-squared: | 0.005 |
| Method: | Least Squares | F-statistic: | 134.3 |
| Date: | Fri, 29 Jan 2021 | Prob (F-statistic): | 5.79e-31 |
| Time: | 23:02:48 | Log-Likelihood: | -13734. |
| No. Observations: | 24184 | AIC: | 2.747e+04 |
| Df Residuals: | 24182 | BIC: | 2.749e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 0.2711 | 0.004 | 72.566 | 0.000 | 0.264 | 0.278 |
| C(anger)[T.1] | -0.0638 | 0.006 | -11.587 | 0.000 | -0.075 | -0.053 |

| | | | |
|---|---|---|---|
| Omnibus: | 4286.865 | Durbin-Watson: | 2.009 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 6059.032 |
| Skew: | 1.196 | Prob(JB): | 0.00 |
| Kurtosis: | 2.460 | Cond. No. | 2.53 |

Figure 12: ANOVA Test 1: Fear on the Rise on Frequently-Occurring Dates

| | Accepted | Rejected |
|---|---|---|
| H0 | Yes | No |
| H1 | No | Yes |

Table 7: Test 2: H1 – Anticipation, Surprise and Optimism on the Rise on Frequently-Occurring Time Values

| | Accepted | Rejected |
|---|---|---|
| H0 | No | Yes |
| H1 | Yes | No |

Table 8: Test 3: H1 – High Retweet Counts were due to Higher Emotional Display in Tweets

| | Accepted | Rejected |
|---|---|---|
| H0 | Yes | No |
| H1 | No | Yes |

Table 9: Test 4.1: H1 – High Rate of Surprise in Tweets based on Indian Locational Identifiers

| | Accepted | Rejected |
|---|---|---|
| H0 | No | Yes |
| H1 | Yes | No |

Table 10: Test 4.2: H1 – High Rate of Pessimistic Outlook in Tweets based on American Locational Identifiers

To gain further intuition into how the models worked based on these hypotheses tests, predictions and the consumed data, I further graphically exploited these notions with tools from Python and Tableau in order to inspect which feature could relate or effect which other features, and how their relationships impacted the overall imprint of the data and model. To keep a healthy balance between each aspect of the following visual analysis, I incorporated insights for both the sample data and the model output data. As the best-model accuracy turned out to be 87% alongside a high ratio of significance acceptances in the hypotheses tests, the visual analysis for the model output data can be considered congruent to the visual analysis of the sample data; hence there is no resorting to redundancy and repetition in the envisioned reporting below.

Fig. 13 below shows the occurrence rate of each emotion label in the predictions of the BERT model, reiterating the earlier assumption and hypothesis that the negative sets of emotions – such as anger, fear and disgust – were more frequent on the spectrum than the positive sets of emotions. This is in direct correlation to the actual labels in the annotated dataset as well as the proposed deduction of the first primary alternate hypothesis (Table 6).

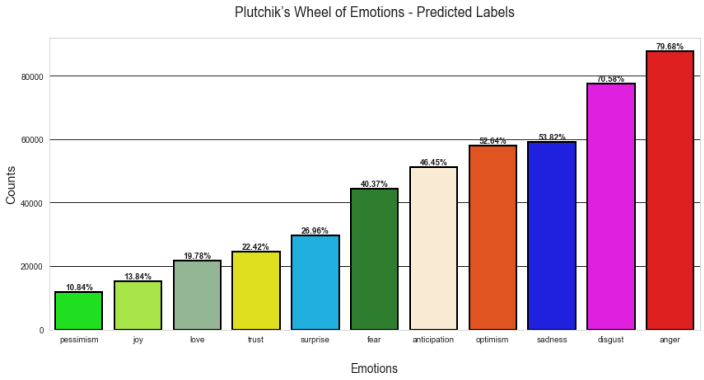



Figure 13: Occurrence of Emotion Labels in the BERT model.

Additionally, the most frequent location areas were also checked – where emotion counts of all labels were more than 60% in terms of occurrence coverage – in order to understand which locations were seemingly providing more towards the cause of the problem being attempted, as presented in Fig. 14. This attempt was also influenced and subsequently verified by the concluding alternate hypothesis tests regarding India and USA's emphasis on certain emotion scales (Tables 9 and 10).

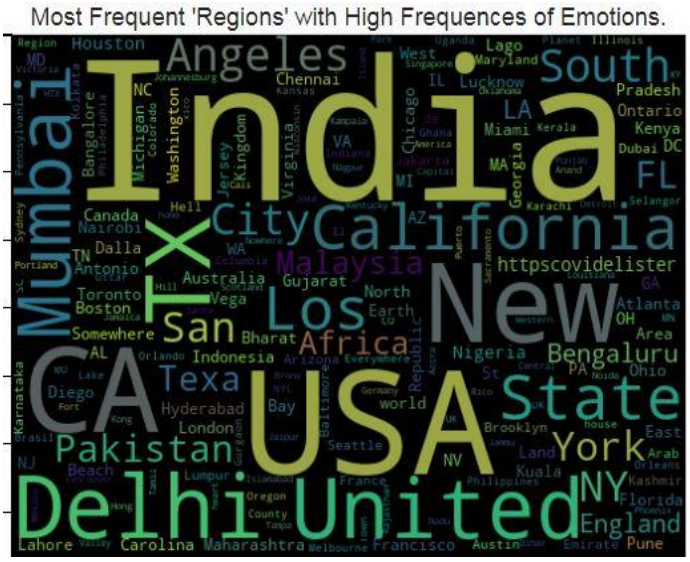



Figure 14: Frequent locating regions that appear to directly affect the predicted emotions.

Similarly, when plotting the Pearson Correlation matrix of the predicted labels as shown in Fig. 15, 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 statistical predictions made can be relied upon. I was able to stratify these alliances of combinations of emotions – such as anticipation and anger, sadness and fear, and more – into their expected clusters on this matrix jurisdiction.

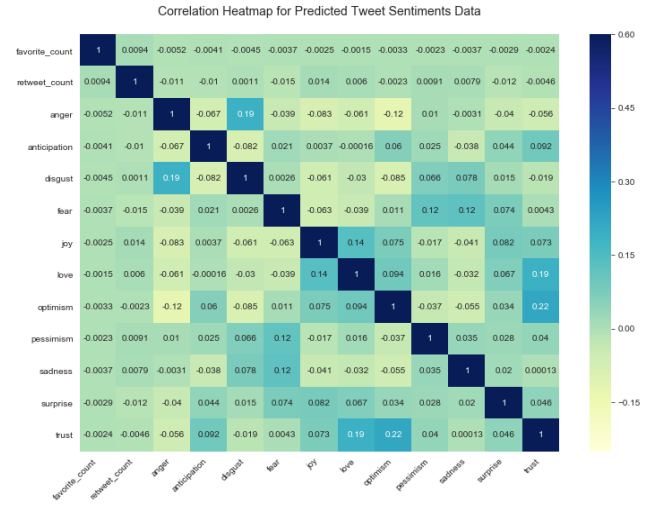



Figure 15: Pearson Correlation Matrix of the Predicted Emotion Labels from the BERT model.

However, this correlation was developed based on the BERT model output, and that too prior to encoding the non-categorical attributes. Performing the Pearson Correlation once again after encoding these features, I was able to extract surprising relations between the attributes which were seemingly over-looked in the initial correlation pass of the model output data, as can be seen in Fig. 16 below. It can be perceived that the 'text' attribute comprising of the actual tweets is heavily influenced by the date and time attribute, which can be anticipated as the nature of tweets will be partial towards the pandemic as it progressed. Similarly 'location,' 'text' and 'time' were also found to be dependant variables. The following visualization of the correlations present in the dataset was found to be in tandem to the visualization of the correlations present in the best-model output data (Fig. 13), as the dependency between negative emotion sets and positive emotion sets can be analytically seen.

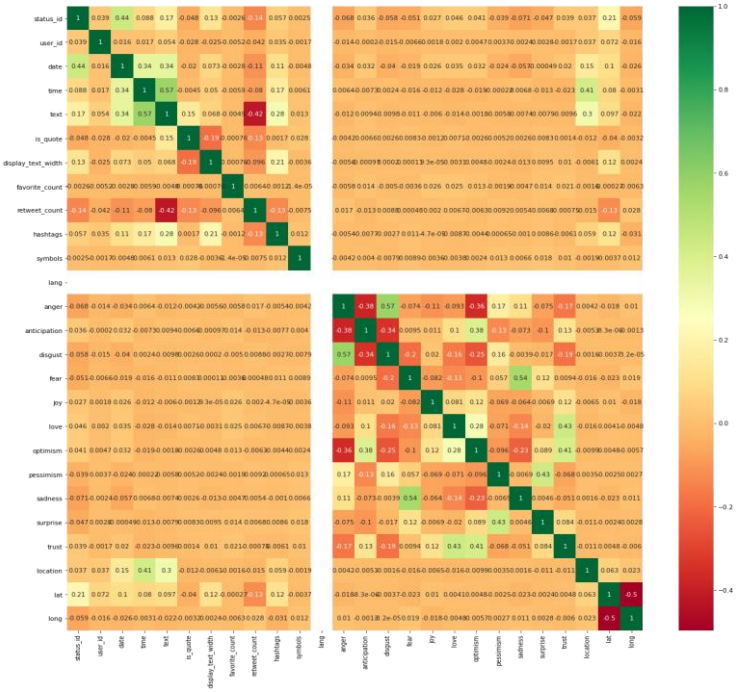


Figure 16: Pearson Correlation Matrix of all Attributes of the COVID-19 Tweets Dataset

While conducting the visual analysis of the COVID-19 dataset, the fact that a major emotion – fear – appeared to exist as a recurrent word in the tweets text was also come across, signalling that it would be likely a major and frequent label in the prediction process (which it was proven to be indeed). 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. 17. In addition to this, the initial hypothesis tests also pointed out and substantiated this large-scale presence (Tables 6 and 8).

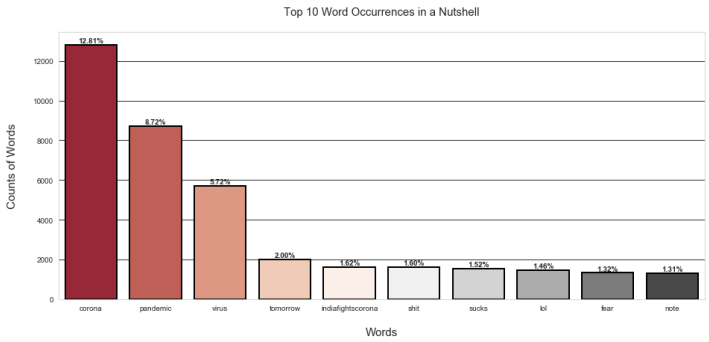



Figure 17: Top 10 Word Occurrences in the tweets of the COVID-19 Tweets Dataset.

In order to map out the trend of emotions of different countries with respect to different dates and months, I set about visualizing interpolated and non-interpolated real-time interactive plots to measure these ideas competitively. I also plotted these concepts without locational restrictions to gain a time-series analysis of the entire set of emotion tags as well. Furthermore, I made subsets of these ideas into positive and negative emotions on the same scale of dates for a spectrum comparison. These approaches can be seen in Figs. 18a, 18b, 19a, 19b, 19c, 20a and 21b below.

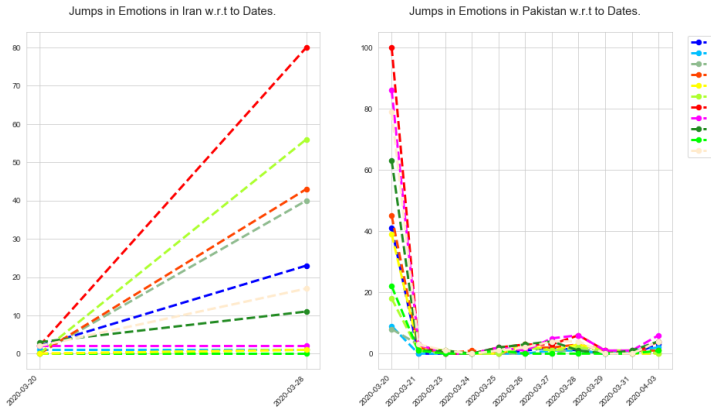



Figure 18a: Comparison of Iran and Pakistan with respect to each emotion and event of date present in the COVID-19 Tweets Dataset

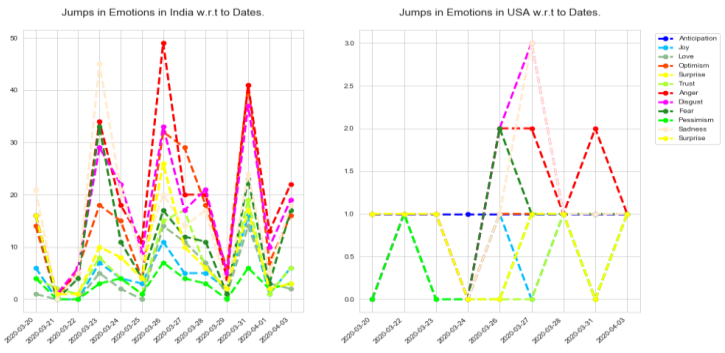



Figure 18b: Comparison of India and USA with respect to each emotion and event of date present in the COVID-19 Tweets Dataset

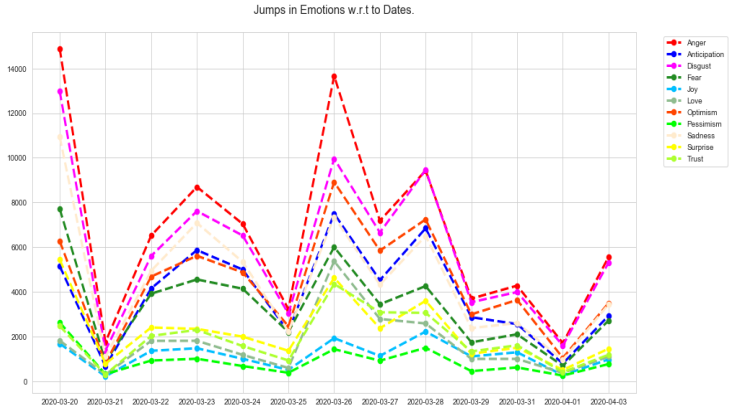



Figure 19a: Comparison of each emotion overall relevant to dates present in the COVID-19 Tweets Dataset

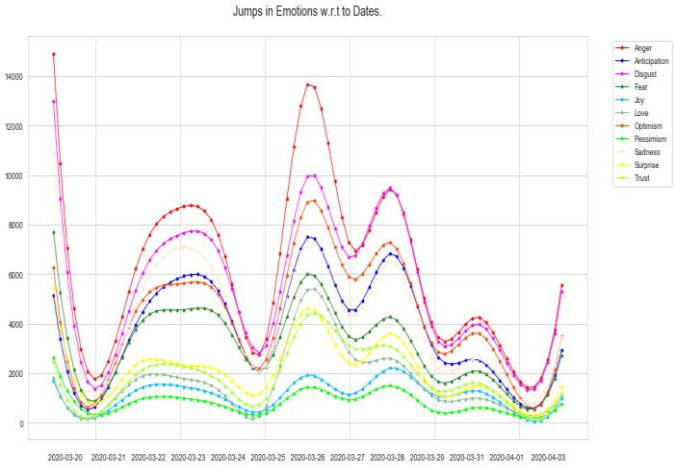



Figure 19b: Interpolated comparison of each emotion overall relevant to dates present in the COVID-19 Tweets Dataset

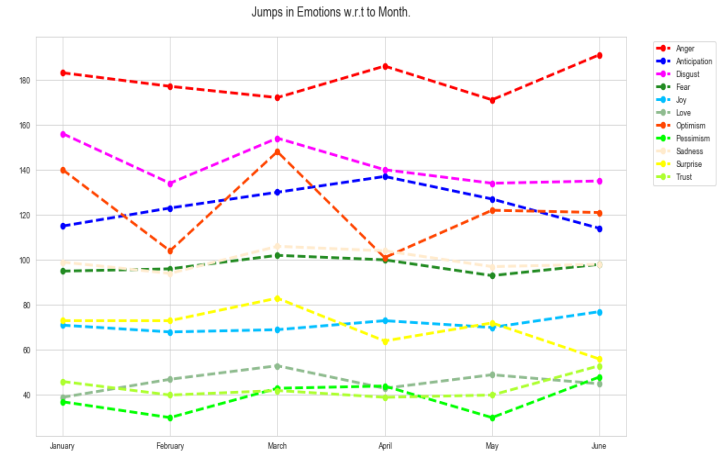



Figure 19c: Comparison of each emotion overall relevant to months present in an earlier scrap of the COVID-19 Tweets Dataset

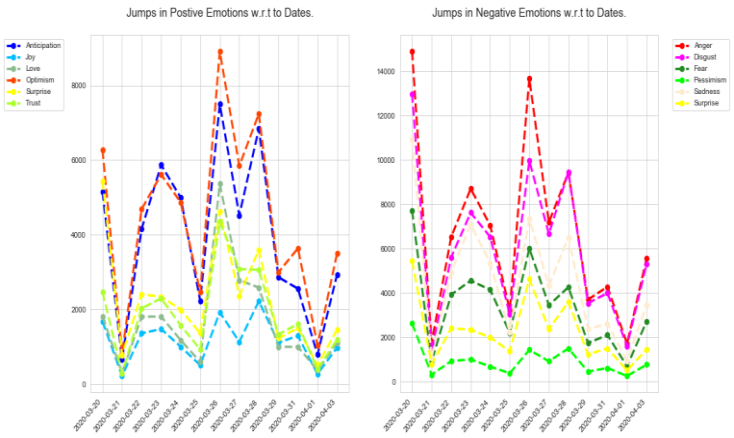



Figure 20a: Comparison of each positive emotion with each negative emotion relevant to dates present in the COVID-19 Tweets Dataset

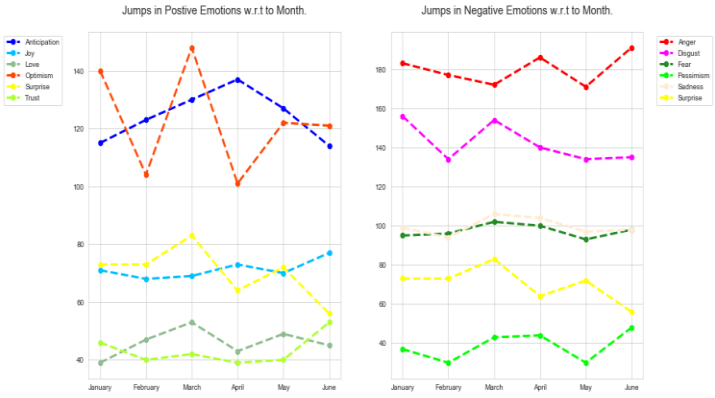



Figure 20b: Comparison of each positive emotion with each negative emotion relevant to months present in an earlier scrap of the COVID-19 Tweets Dataset

To comprehend how distributed the emotions were globally, and which emotion dominated these areas, I also plotted emotion counts with reverence to their longitude and latitude to see what areas contributed to what sentiment clusters. Fig. 21 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 colour 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.

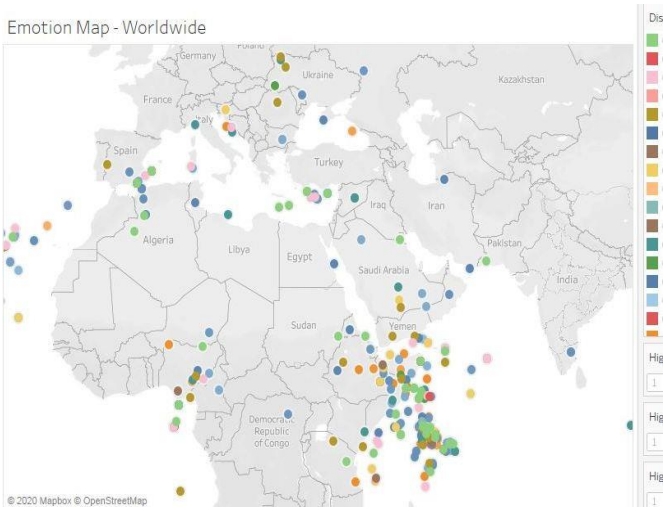



Figure 21: A fragment of the World Map depicting clusters of emotions with intensity

The results of the best performing model – the BERT modular system – also identify and substantiate a pivotal relationship: the optimal prediction of the negative set of emotions (such as fear and anger, for example) directly correlate to an increase in the number of confirmed cases of COVID-19 afflicted patients alongside an increase in the number of deaths due to the ailment. For a proof of concept, the following analyses in Figs. 22a and 22b show these aspects of increments visually for recoveries, confirmed cases and deaths based on data acquired from the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at the John Hopkins University (JHU) [11] for the same timeline as the self-curated data accumulated at my end for solving the predictive problem. It is understandable and generically expected that as these quantifiers increased exponentially, the presence of the negative set of emotions amplified on a similar scale as well.

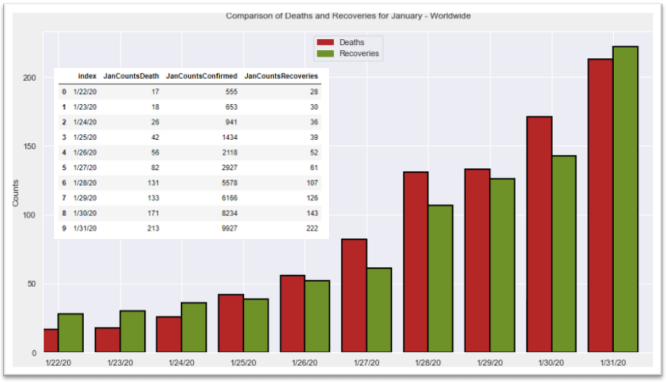
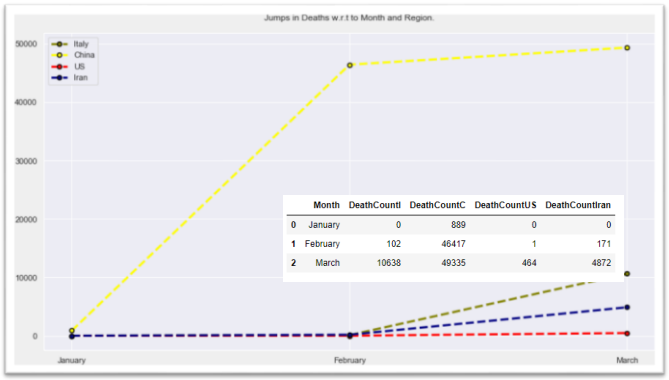



Figure 22a: Tabular and Visual Comparison of Deaths, Recoveries and Confirmed Cases for the month of January, 2020




| | Month | DeathCountI | DeathCountC | DeathCountUS | DeathCountIran |
|---|---|---|---|---|---|
| 0 | January | 0 | 889 | 0 | 0 |
| 1 | February | 102 | 46417 | 1 | 171 |
| 2 | March | 10638 | 49335 | 464 | 4872 |

Figure 22b: Tabular and Visual Comparison of Deaths for the months of January, February and March 2020 of the highly-hit countries of China, Italy, USA and Iran

## 7. Conclusion

Based on the model implementations, performances and analysis, I was 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. The interest lies 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 the conjecture 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 meant 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. It cannot expect to beat or even reach it realistically without the usage of GPUs (which were hard for to be utilized but was managed sparsely using Google Colaboratory). Thus, results are prone to this data directly and can be hence improved with more qualified data. With this supplementary data, it may be able to map the BERT-base model to a BERT-large model which could hypothetically improve on performance.

Another realization of this study was that a BOW model appears to be despondently outdated to the current scenarios of real-time problems. It is a good baseline to work with in accustoms to beginner-friendly terminology, where 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 in present day and usage.

This setup can be modelled better with the above mentioned enhancements, but it currently works at respectable ratios, considering the limitative barriers and resources discussed.

## 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 BERT in the Field of Disaster Management,” 2019.

[7] Md. Y. Kabir and S. Madria, “CoronaVis: A Real-time COVID-19 Tweets Data Analyzer and Data Repository,” July 2020.

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

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

[10] https://www.kdnuggets.com/2017/06/practical-importance-feature-selection.html

[11] https://github.com/CSSEGISandData/COVID-19