FINAL REVIEW

Twitter Emotion Analysis – An LSTM Approach

ROHAN ALLEN (18BCI0247)

GUIDE: DR. MANIKANDAN N

ABSTRACT

The primary goal of this project is to stream real-time COVID-19 related tweets using Twitter and its Python API-Tweepy and then conduct a sentiment and emotion analysis on the same. The results are compared with a corpus of tweets collected two years ago. Initially lexicon-based tools such as Text2Emotion, VADER, TextBlob, etc were considered. To incorporate machine learning concepts, we also utilized some pretrained emotion detection models from the work published by Colnerič and Demšar.
These models were able to predict emotions with regards to some famous psychological models such as Ekman, Plutchik and the Profile of Mood States. However, we decided to build our very own deep learning model due to some deficiencies discovered in the pre-trained models.

INTRODUCTION

“Data is the new oil,” is a common statement touted by many leading scientists and eminent personalities for the past few years. But on closer inspection, one can argue that the power of data far exceeds that of oil, because it is an infinite commodity that will only increase in quantity. Massive amounts of data enable organizations to analyse trends and patterns, draw accurate and meaningful insights into customer preferences, and seamlessly predict the customer’s very next move. The vast range of applications of data science is, quite simply put, frightening, and is something that makes it an extremely exciting and enticing prospect.

Twitter, a social media and microblogging platform is a storehouse of information and data in today’s world. This is where the news first breaks and where millions of people come to voice their opinion on a variety of fields such as politics, sports, entertainment, etc. On average there are nearly 500 million tweets tweeted daily. This is a staggering number, one which makes Twitter an ideal source of data for our experiment.

MOTIVATION

Sentiment Analysis is a very popular tool deployed by many organizations to test the popularity of a product and gauge the customer's preferences. It is used in a wide variety of field such as online shopping, movie reviews, etc. However, the sentiment is described as merely positive or negative which is rather vague. This project aims to delve deeper into the topic of sentiment analysis and classify a sentiment into specific emotions such as love, anger, fear, joy, etc. This would certainly help in providing a more nuanced understanding of the mood of the public towards a certain product/idea.

OBJECTIVE

To compare and contrast the sentiment and emotions of two corpora of COVID-19 tweets using both lexicon-based approach and deep learning based approach.

STRATEGY

- To use Big Data and Data Mining concepts to collect a corpus of 265,000 tweets using Tweepy.
- To explore NLP concepts thoroughly and use the same to pre-process the tweets.
- Performing Sentiment and Emotion Analysis using existing tools.
- To research extensively on deep learning/machine learning and experiment with different frameworks and train a bi-directional LSTM-RNN model using TensorFlow to predict emotions of tweets.
- Delving into the topic of hyper parameter tuning and constantly experimenting with numerous parameters to improve the accuracy and the performance of the model.
- Producing numerous histograms, word clouds, etc using the matplotlib python package to effectively visualize the tweets.

Retrieving Tweets

Tweepy a simple and user-friendly python library for accessing the Twitter API was used to stream real time tweets. The tweets were filtered based on the language(only English tweets were retrieved) and certain covid related hashtags.

List of Hashtags used to filter the tweets

['#covid19','#longcovid','#coronavirus','#stayhome','#socialdistancing','#covid-19','#covid2019','#coronavirusoutbreak', '#sarscov2', '#virus', '#covidisnotover', '#covidvaccines','#vaccinated', '#longcovid', '#omicron', '#cases', '#covid', '#pandemic', '#coronaviruspandemic', '#mask', '#deltacron', '#covidiots']

Explanation of Corpora used in Project

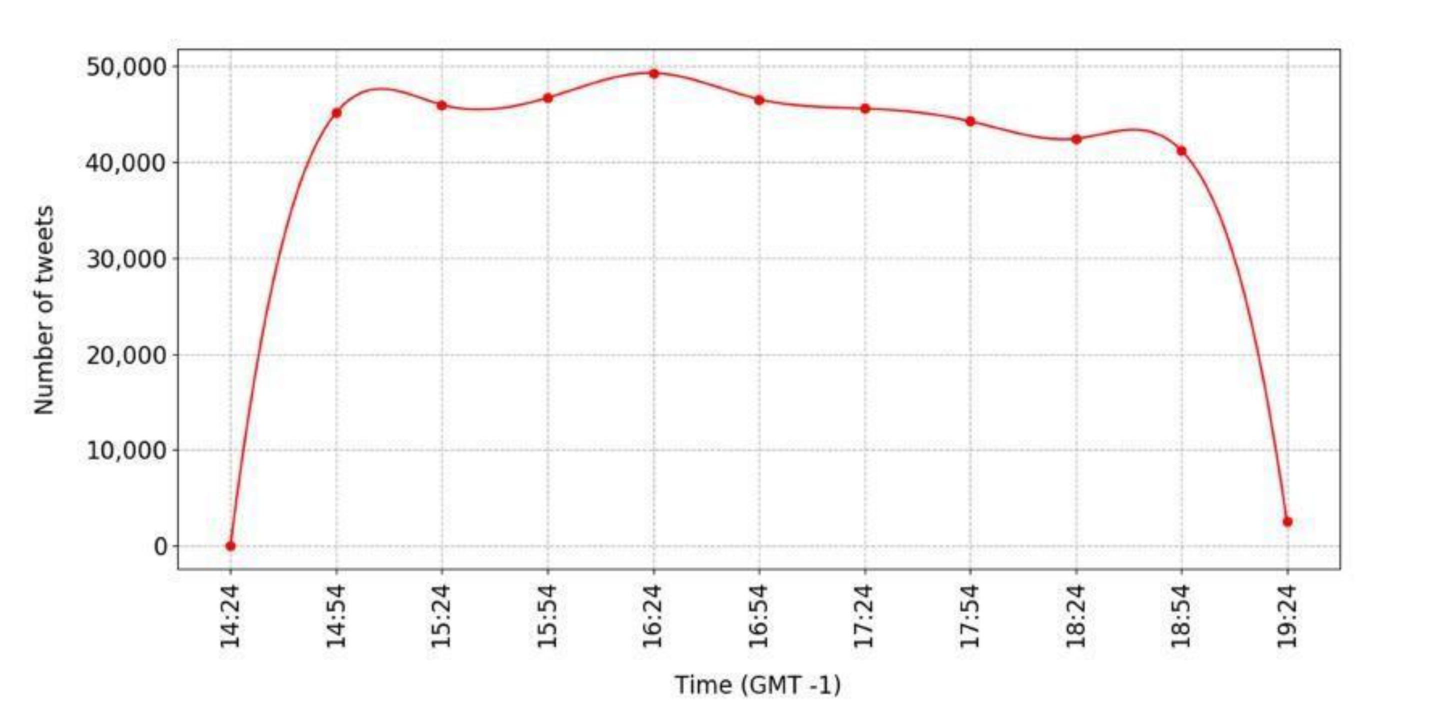
The tweet collection commenced on the 24th of March, 2022 at 23:30 GMT and stopped on the 29th of March, 2022 at 00:30 GMT. The total duration of tweet collection was ninety-six hours and a total of 265,108 tweets were collected.

The corpus of tweets collected by Palomino and Varma on 22nd April 2020 from 15:30 to 19:30 (GMT) are also taken into account. Nearly two years have passed and it is not unreasonable to assume that the public sentiment towards the covid pandemic has changed. We aim to compare and contrast the sentiment of the tweets collected two years ago and as of now and see if there is any change.

Hashtag Distribution

Hashtag	No. of Tweets
#covid19	238,432
#coronavirus	116,557
#stayhome	31,820
#covid_19	11,068
#socialdistancing	6510
#covid-19	4636
#covid2019	2341
#flattenthecurve	2124
#coronavirusoutbreak	2058
#sarscov2	1861
#virus	1211

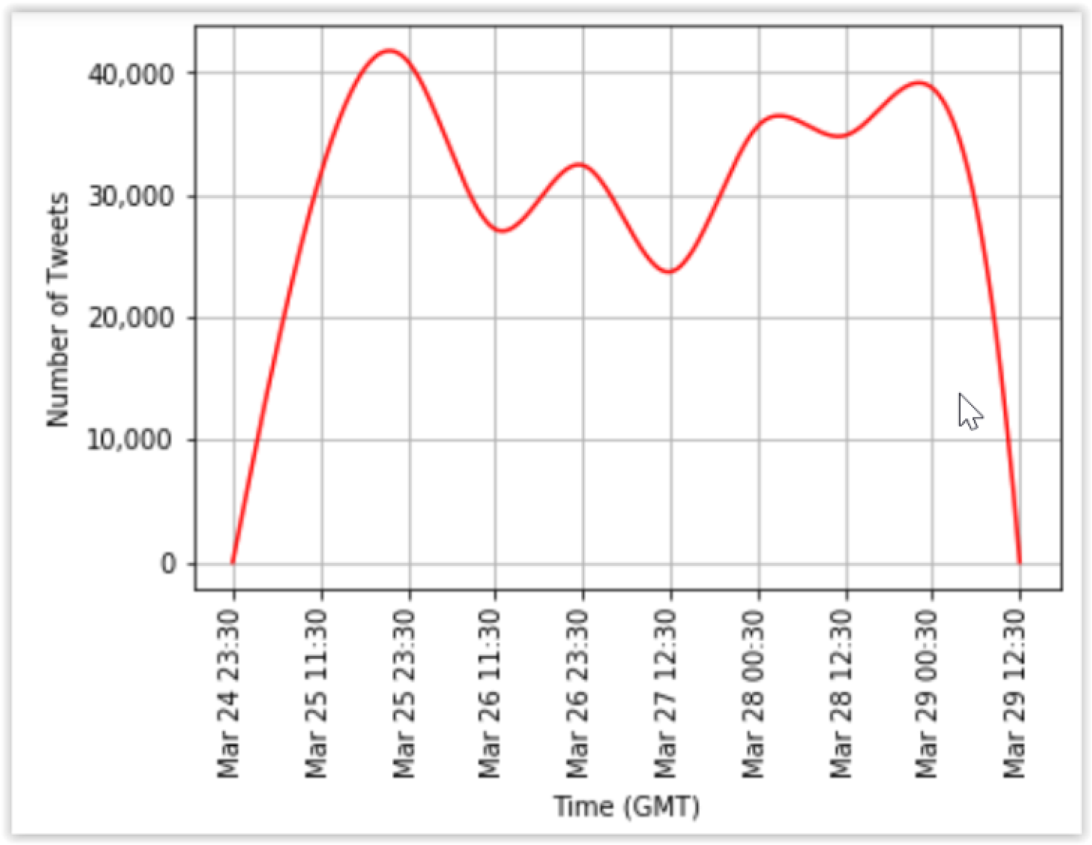
Tweet Distribution over Time



Hashtag Distribution

Hashtag	No. of Tweets
#covid	182,686
#covid19	125,434
#longcovid	34,565
#covidisnotover	25,010
#omicron	15,608
#coronavirus	9487
#covid-19	8443
#pandemic	8429
#mask	6627
#sarscov2	2950
#stayhome	2118
#virus	2032
#vaccinated	1320
#covidiots	1038
#socialdistancing	502
#deltacron	283
#cases	273
#covidvaccines	206
#covid2019	196
#coronaviruspandemic	188
#coronavirusoutbreak	153

Tweet Distribution over Time



SENTIMENT ANALYSIS

When it comes to tools that may be used to calculate a text's sentiment score, there are several options. This decision is crucial since the entire research is skewed in this direction because to the basic differences in the algorithms' foundations. Due to the lack of consensus in sentiment analysis tools we experiment with three such tools VADER, TextBlob and SentiStrength to compare and contrast the results.

EMOTION DETECTION

The bedrock of Emotion Detection (ED) systems are emotion models, which determine how emotions are expressed. The models presume that emotions exist in different states, justifying the need to differentiate between them. When engaging in any ED-related task, it is essential to initially determine the emotional model to be used.

The Paul Ekman model groups emotions into six types. According to his theory, there are six basic emotions that arise from discrete neurological systems as a result of how a person interprets a situation, and hence emotions are independent. Happiness, sadness, anger, disgust, surprise, and fear are the basic emotions. However, the combination of these feelings can result in more complicated emotions like guilt, shame, pride, desire, and greed, etc.

The Robert Plutchik model, much like Ekman, proposes that there are just a few primary emotions that occur in opposite pairings and combine to form complicated emotions. Supplementing Ekman's six core emotions, he named two more, including trust and anticipation, making a total of eight essential emotions. Joy vs. sadness, trust vs. disgust, anger vs. fear, and surprise vs. anticipation are the eight opposing emotions. As per Plutchik, there are differing levels of intensities for each emotion determined by how an event is interpreted by an individual.

EMOTION DETECTION

Profile of Mood States or POMS is a standard validated psychological test formulated by McNair et al to assess the psychological state or mood of a person. It classifies emotions into seven categories; anger, confusion, depression, fatigue, friendliness, tension and vigour.

Building on Colnerič and Demšar’s work we use their pre-built models to classify the emotions on our tweet collections using Ekman, Pluthick and POMS emotion models. Furthermore, to supplement this we decided to use Text2Emotion, which is lexicon based python library to attain emotions of tweets. The five emotions that are assessed in Text2Emotion are Happy, Angry, Sad, Surprise and Fear.

TECHNICAL SPECIFICATIONS

Hardware Requirements:

System: intel i7

Hard Disk: 1 TB

Monitor: 15.6’’ LCD

Ram: 8GB

TECHNICAL SPECIFICATIONS

Software Requirements:

Operating System: Windows 10

Coding Language: Python

Coding Environment: Google Collaboratory

Framework: Keras/TensorFlow

Deep Learning Bi-Directional RNN-LSTM

The results after using the pre-trained Ekman and Plutchik model are quite suspicious, as feelings of joy and trust are some of the dominant emotions, which seems highly unlikely. Therefore, in order to address this, we have trained our very own Bi-Directional RNN-LSTM to predict emotions.

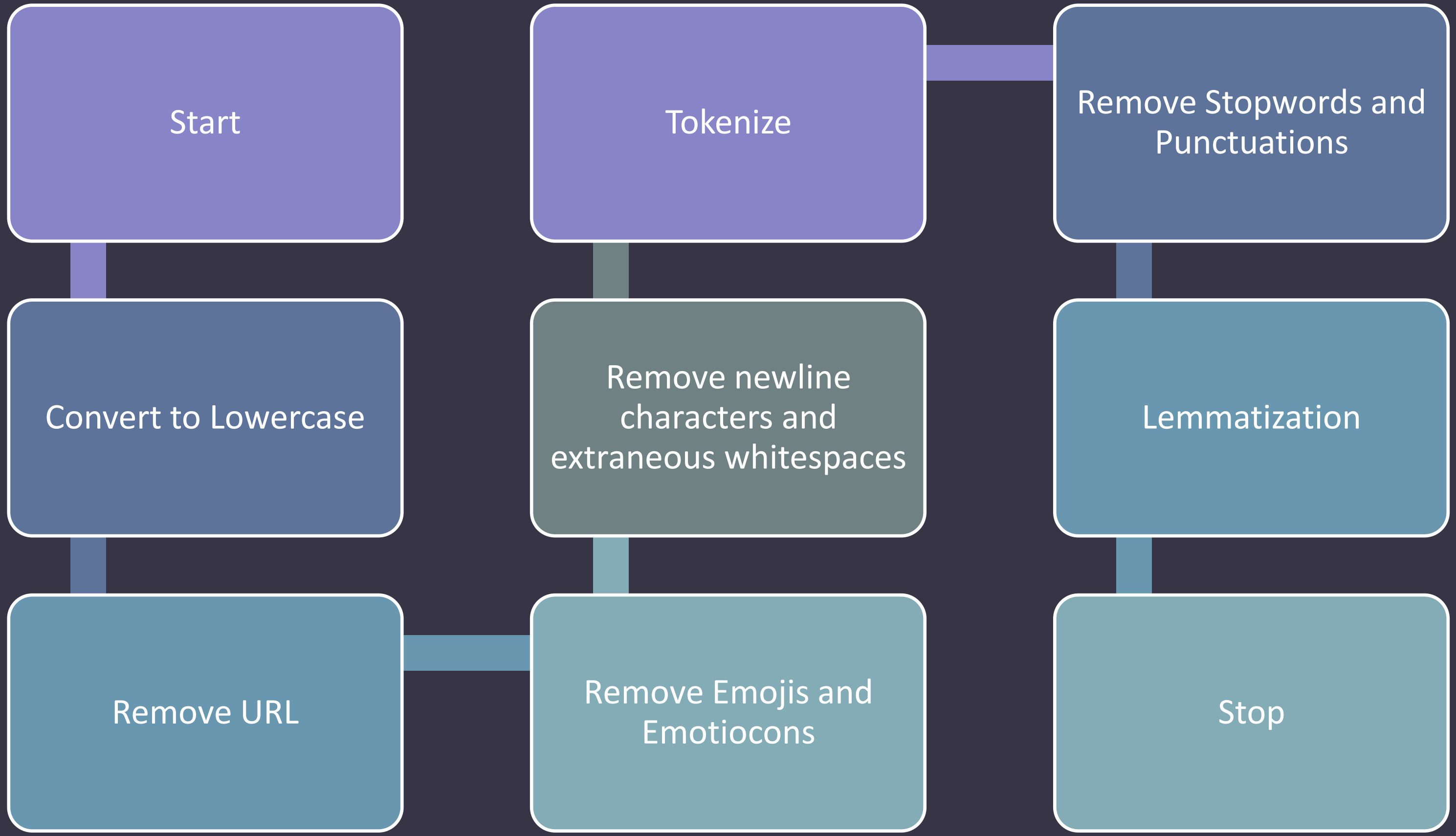
RNN

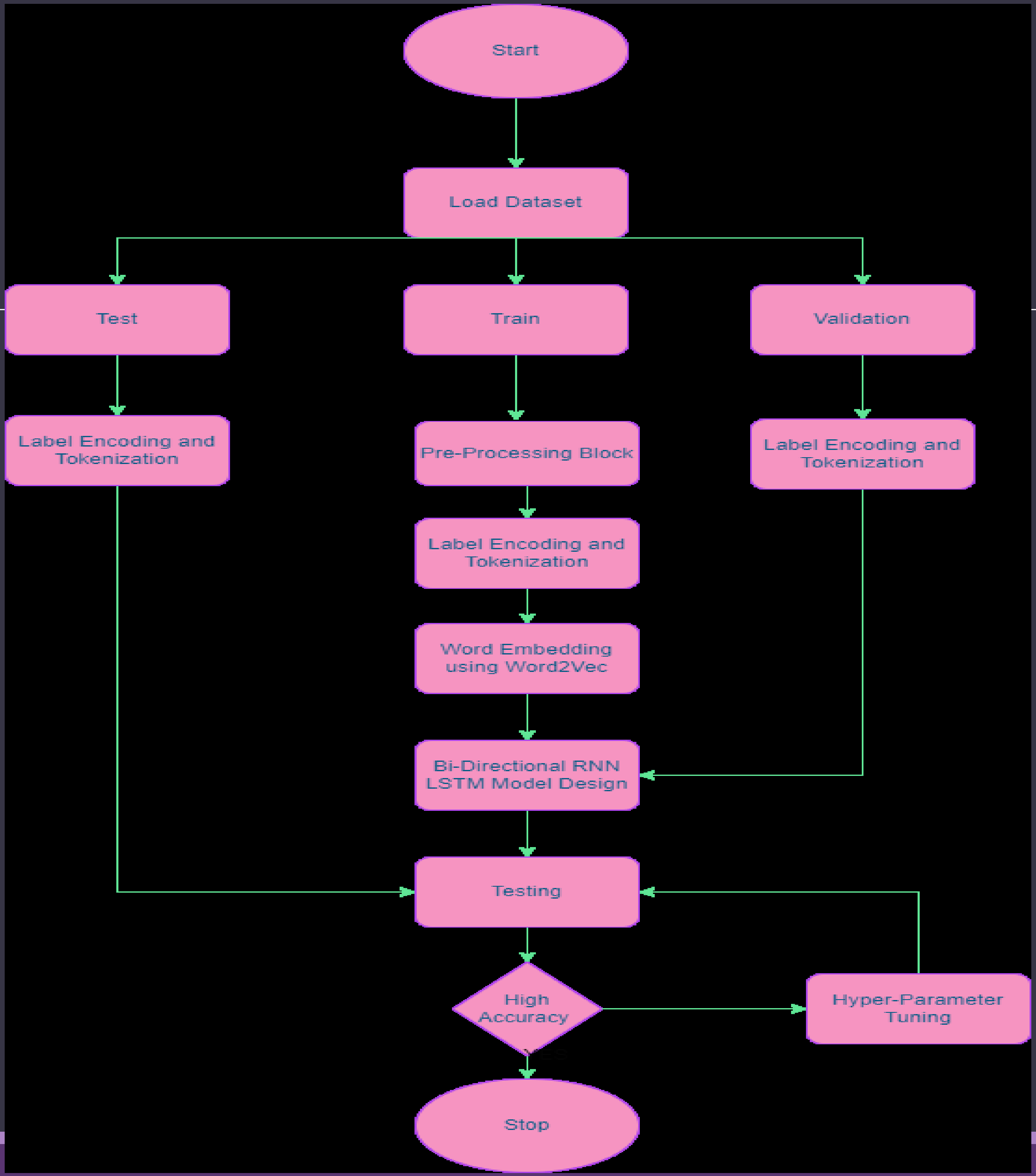
Regular artificial neural networks do not depend on the sequence or order of input. For example, in an artificial neural network to detect if a credit card transaction is fraudulent or not, the output does not depend on the order of the input parameters of the neural network. However, for natural language processing tasks such as sentiment analysis, this is not the case. All NLP tasks are sequence modelling problems, where the order of the words play an important role in determining the meaning and context of a sentence, and thus the sentiment.

LSTM

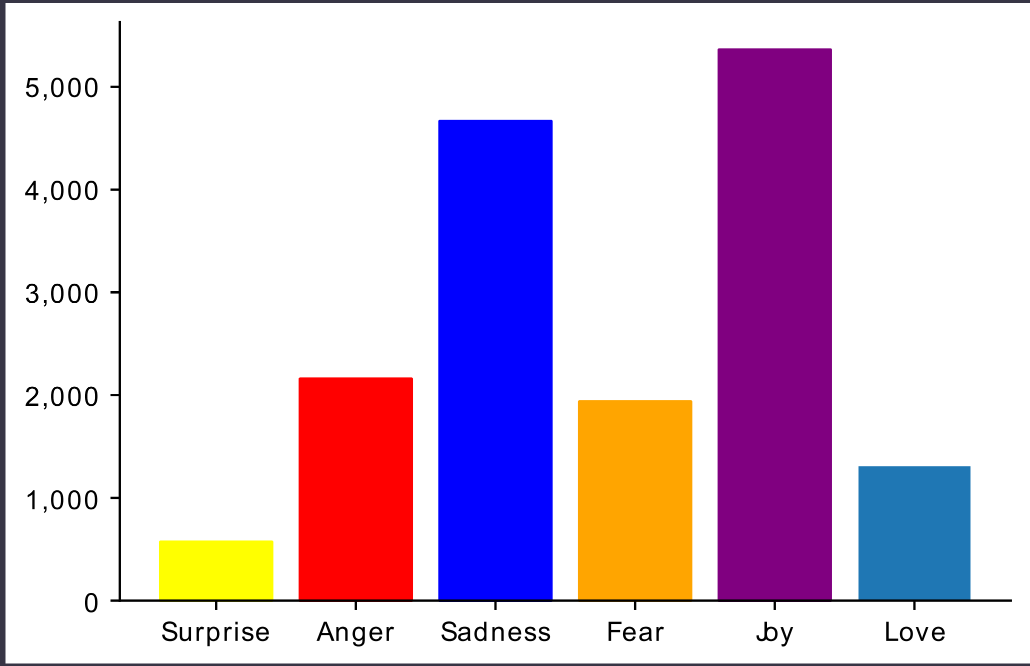
A common problem faced with RNN's is that it has an extremely short-term memory. Thus, words encountered at the beginning of a sentence are quickly forgotten and have little influence on the later and final outputs of the neural networks. This is due to the vanishing gradient problem. After the forward pass of the neural network, the cost and then the total loss is calculated. During the backpropagation step, the loss is differentiated with respect to the weights of the input parameters so that the weights can be adjusted accordingly and the model will learn and minimize the loss value, maximizing accuracy. In a deep neural network however when adjusting the weights of the inputs of the first layer, the value of the gradient is extremely small, due to the constant multiplication of minute numbers. Hence, the change of their values are negligible, and soon they are forgotten by the RNN.

Pre-Processing Block




Start
Load Dataset
Test
Train
Validation
Label Encoding and Tokenization
Pre-Processing Block
Label Encoding and Tokenization
Label Encoding and Tokenization
Word Embedding using Word2Vec
Bi-Directional RNN LSTM Model Design
Testing
High Accuracy
Hyper-Parameter Tuning
Stop

Training Data



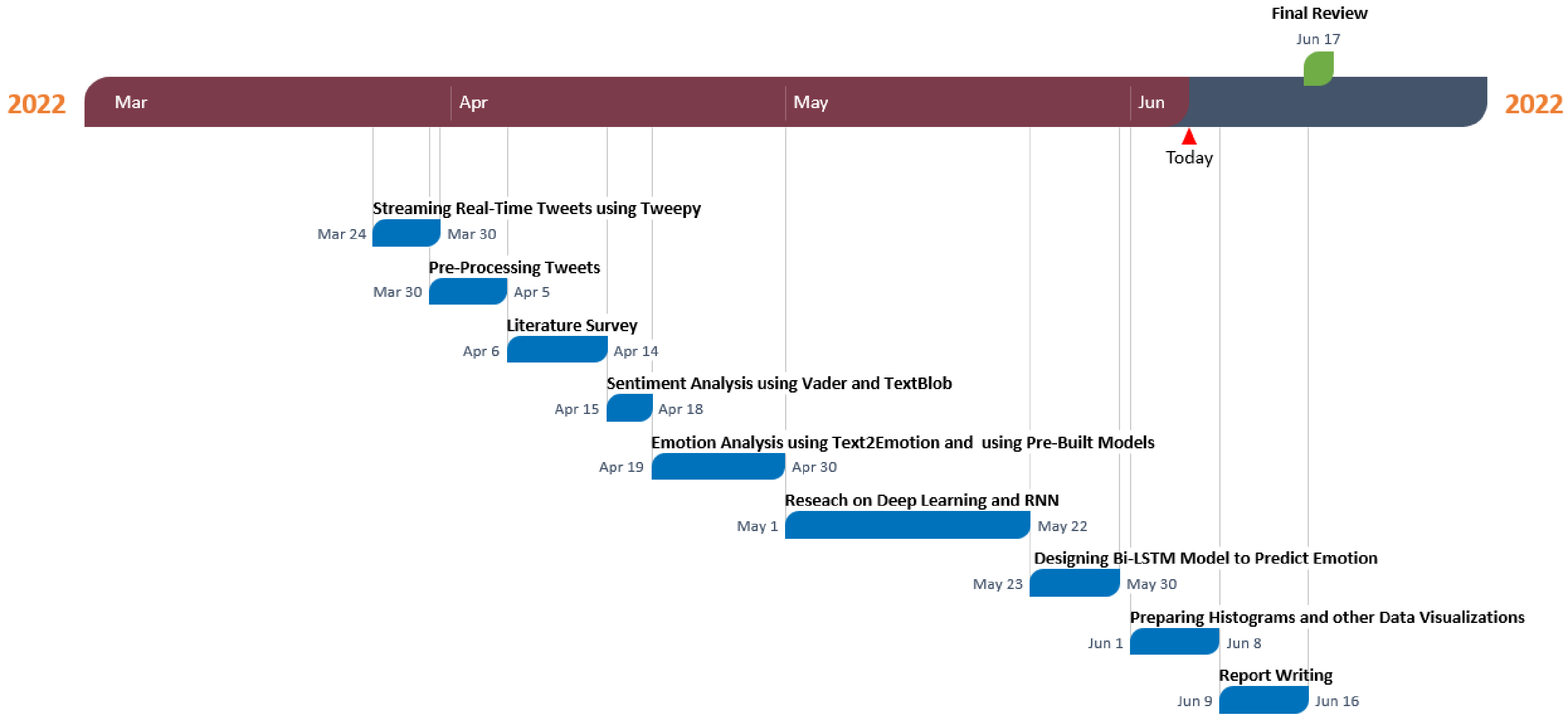
Model Design

```
EMBEDDING_DIM = 100
class_num = 6
model = Sequential()
model.add(Embedding(input_dim = num_words,
output_dim = EMBEDDING_DIM,
input_length= X_train_pad.shape[1],
weights = [gensim_weight_matrix],trainable = False))
model.add(Dropout(0.2))
model.add(Bidirectional(CuDNNLSTM(100,return_sequences=True)))
model.add(Dropout(0.2))
model.add(Bidirectional(CuDNNLSTM(200,return_sequences=True)))
model.add(Dropout(0.2))
model.add(Bidirectional(CuDNNLSTM(100,return_sequences=False)))
model.add(Dense(class_num, activation = 'softmax'))
model.compile(loss = 'categorical_crossentropy', optimizer = 'adam',metrics = 'accuracy')
```

MODEL SUMMARY

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding (Embedding)       (None, 300, 100)          1000000

 dropout (Dropout)           (None, 300, 100)          0

 bidirectional (Bidirectiona  (None, 300, 200)         161600
 l)

 dropout_1 (Dropout)         (None, 300, 200)          0

 bidirectional_1 (Bidirectio  (None, 300, 400)         643200
 nal)

 dropout_2 (Dropout)         (None, 300, 400)          0

 bidirectional_2 (Bidirectio  (None, 200)              401600
 nal)

 dense (Dense)               (None, 6)                 1206

=================================================================
Total params: 2,207,606
Trainable params: 1,207,606
Non-trainable params: 1,000,000
_________________________________________________________________
```

SCHEDULE TASKS AND MILESTONES

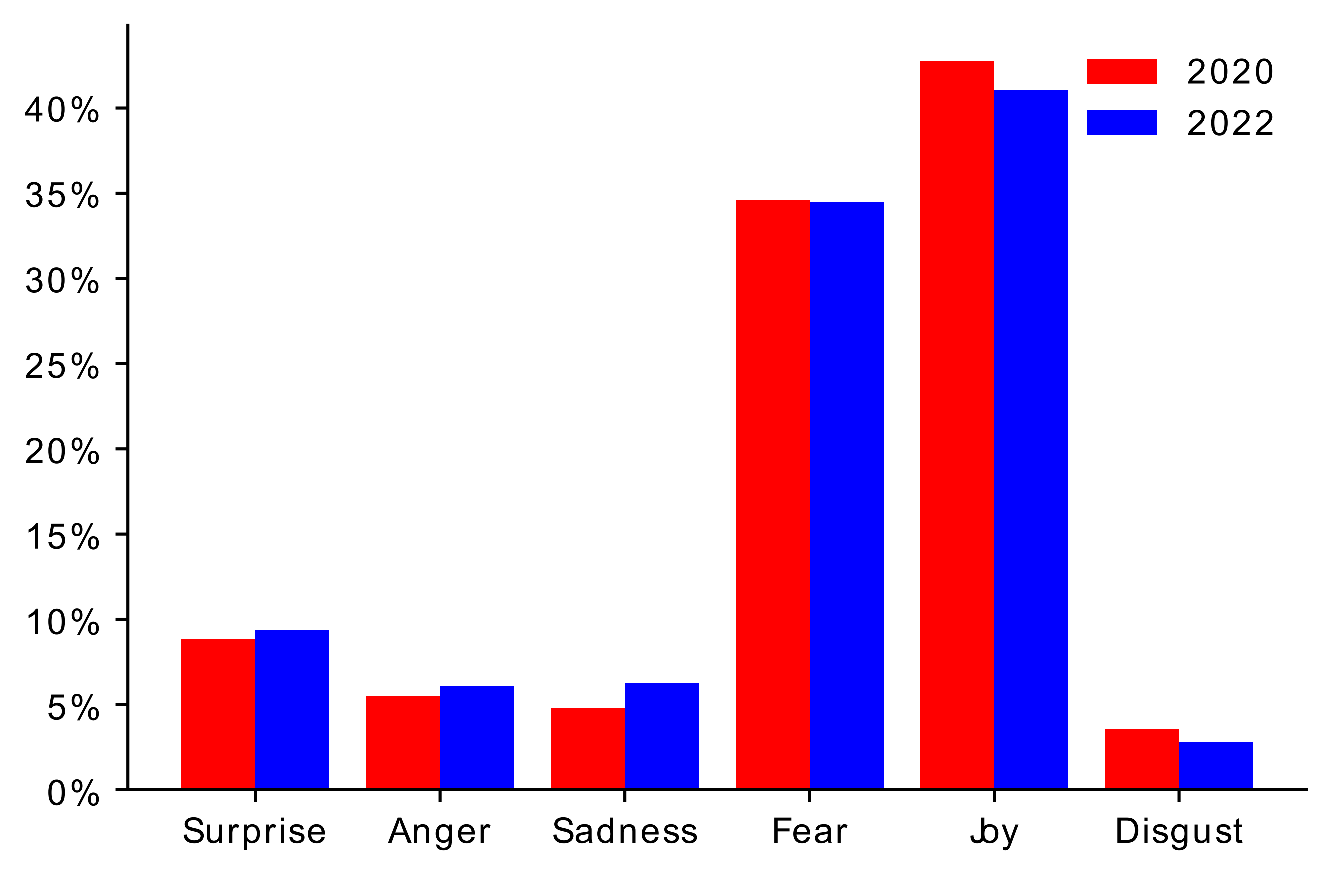


SCHEDULE TASKS AND MILESTONES

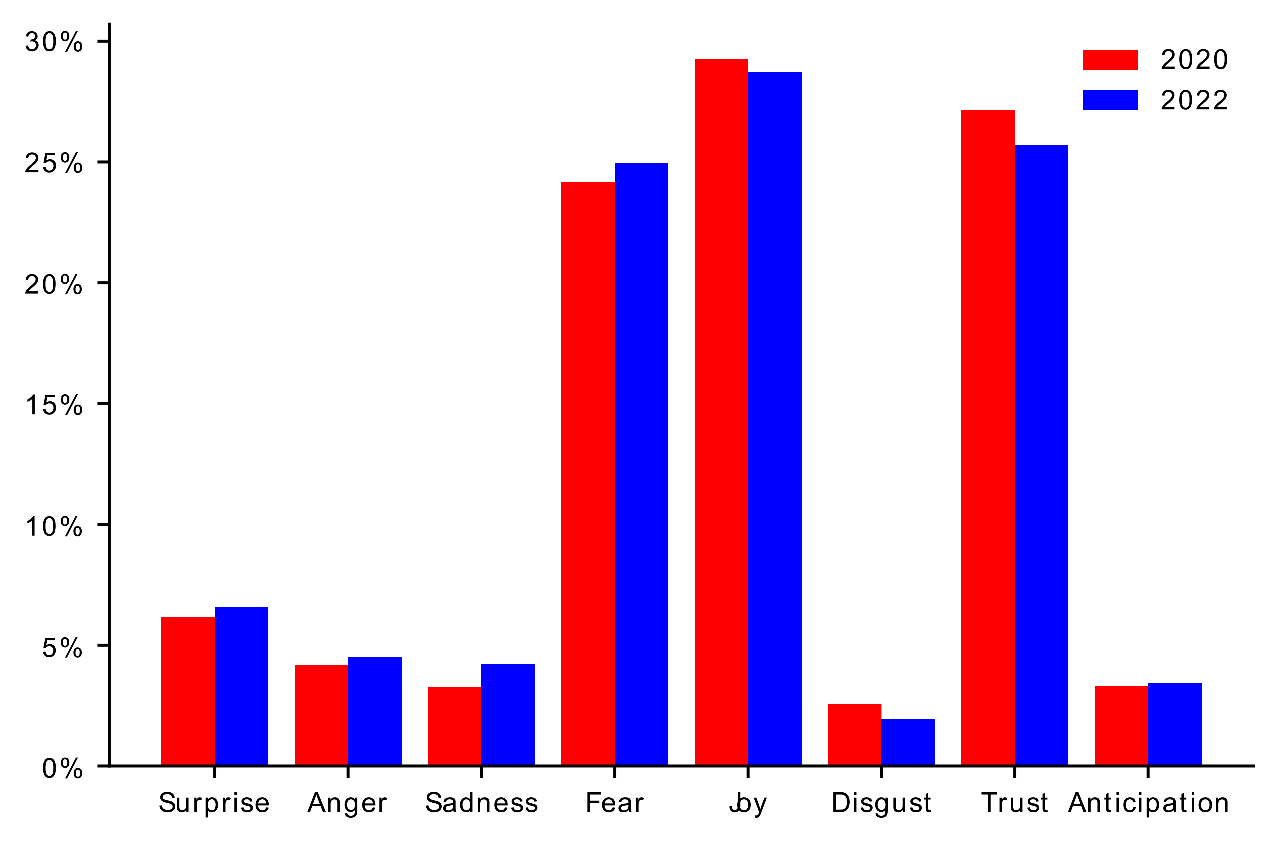
Task/Milestone	Start Date	End Date	Duration(Days)
Streaming Real-Time Tweets using Tweepy	24-03-2022	30-03-2022	6
Pre-Processing Tweets	30-03-2022	05-04-2022	6
Literature Survey	06-04-2022	14-04-2022	8
Sentiment Analysis using Vader and TextBlob	15-04-2022	18-04-2022	3
Emotion Analysis using Text2Emotion and using Pre-Built Models	19-04-2022	30-04-2022	11
Reseach on Deep Learning and RNN	01-05-2022	22-05-2022	21
Designing Bi-LSTM Model to Predict Emotion	23-05-2022	30-05-2022	7
Preparing Histograms and other Data Visualizations	01-06-2022	08-06-2022	7
Report Writing	09-06-2022	16-06-2022	7
Final Review	09-07-2022		1

RESULTS

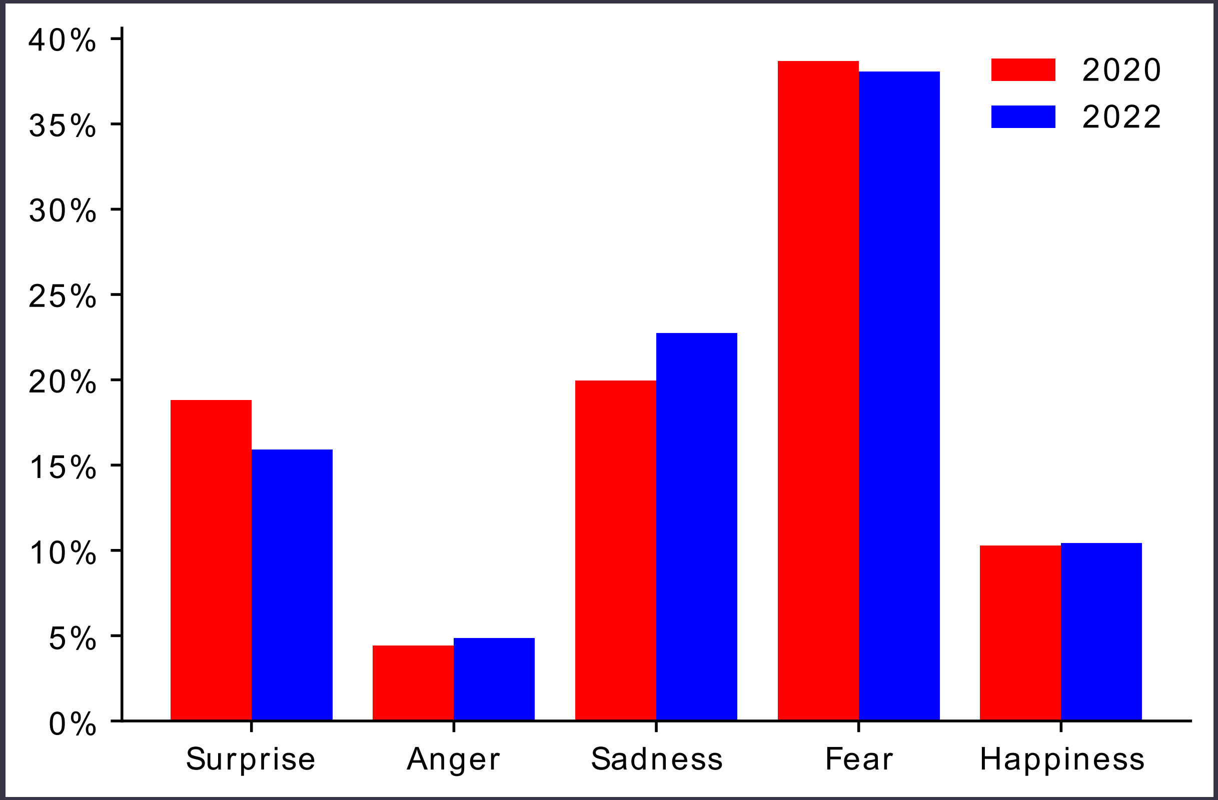
Cumulative distribution of the probabilities of each of Ekman's emotions.



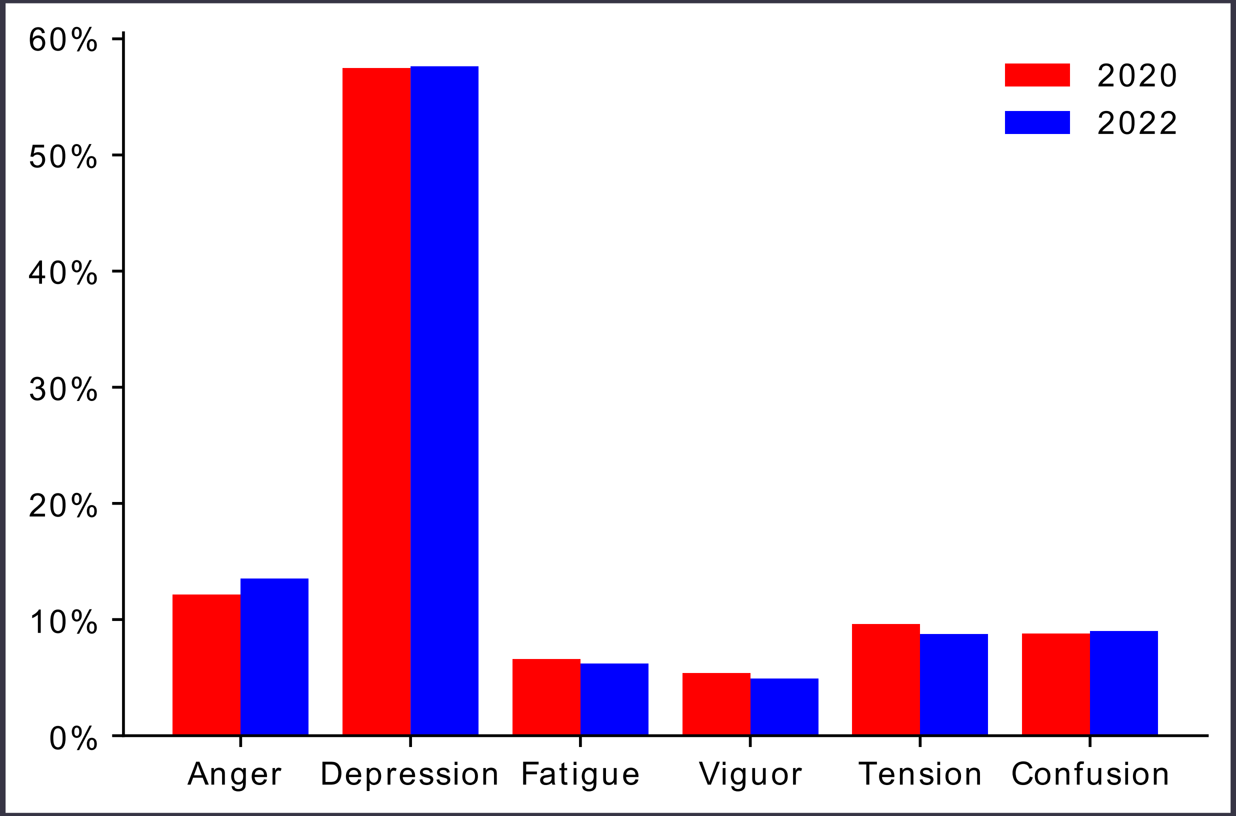
Cumulative distribution of the probabilities of each of Plutchik's emotions.



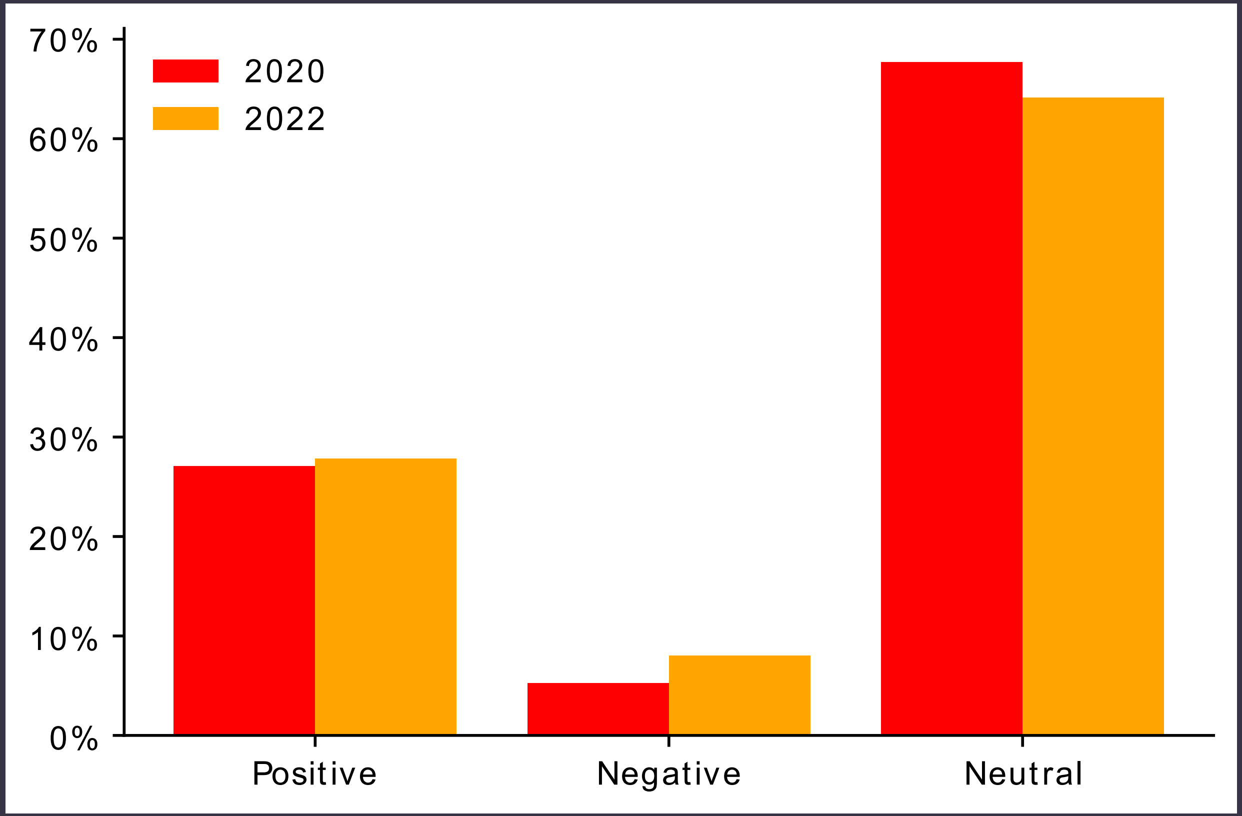
Cumulative distribution of the probabilities of each emotion in Text2Emotion.



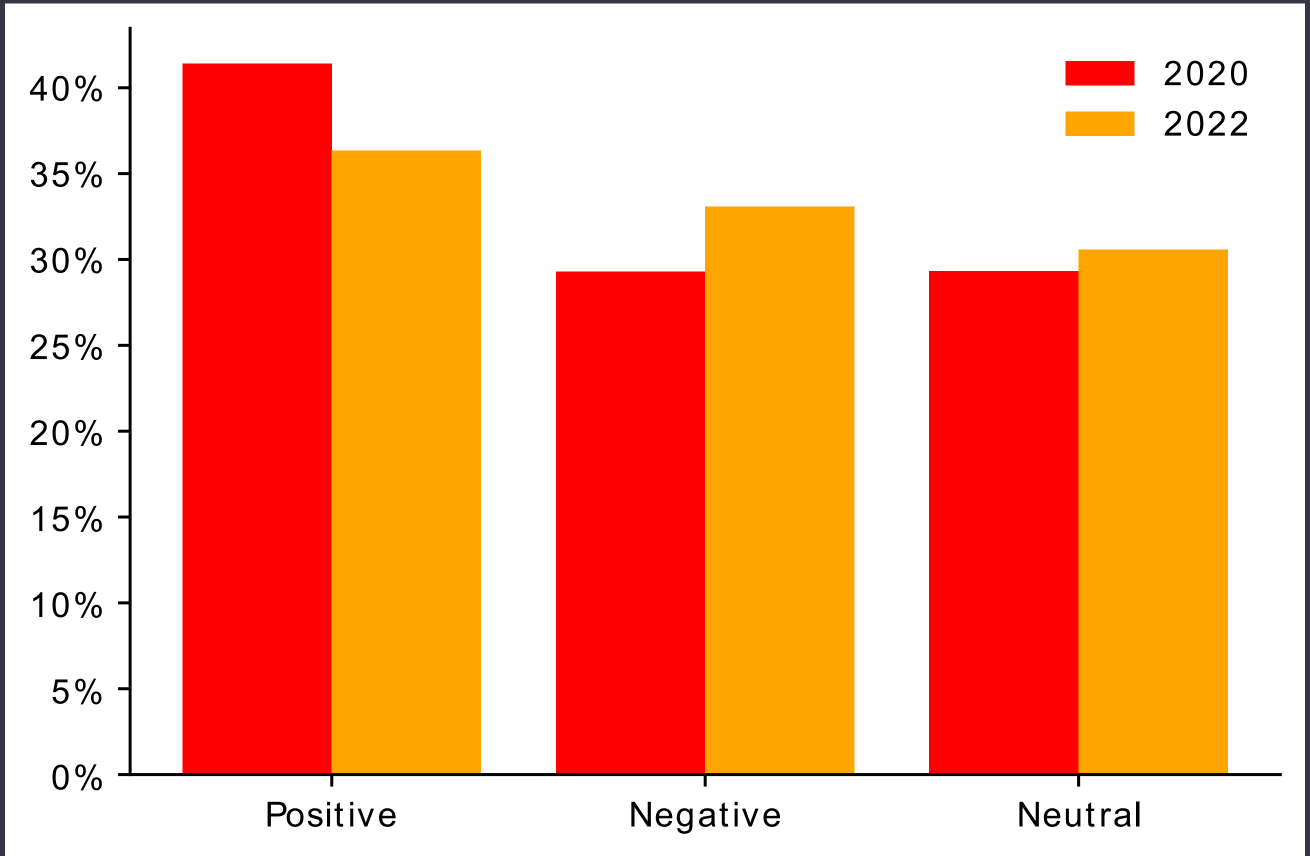
Cumulative distribution of the probabilities of each of POMS's emotions.



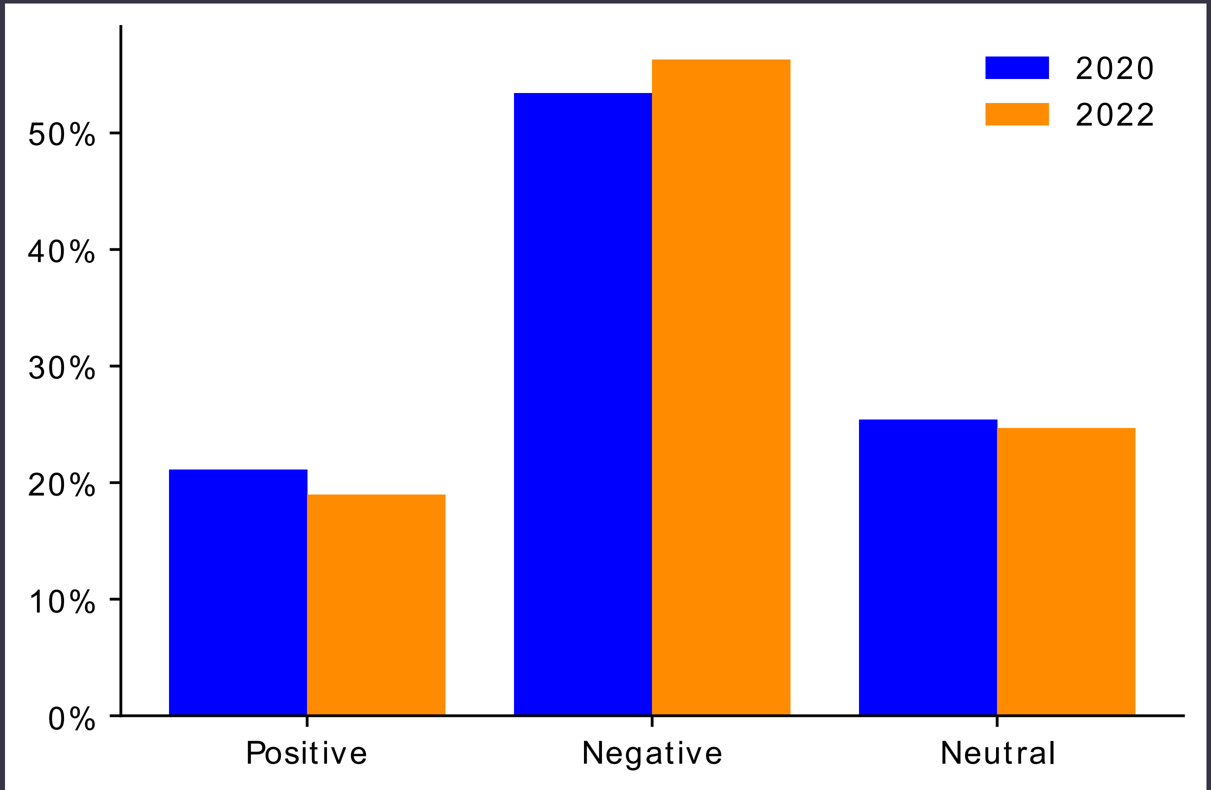
Frequency distribution of tweets according to TextBlob.



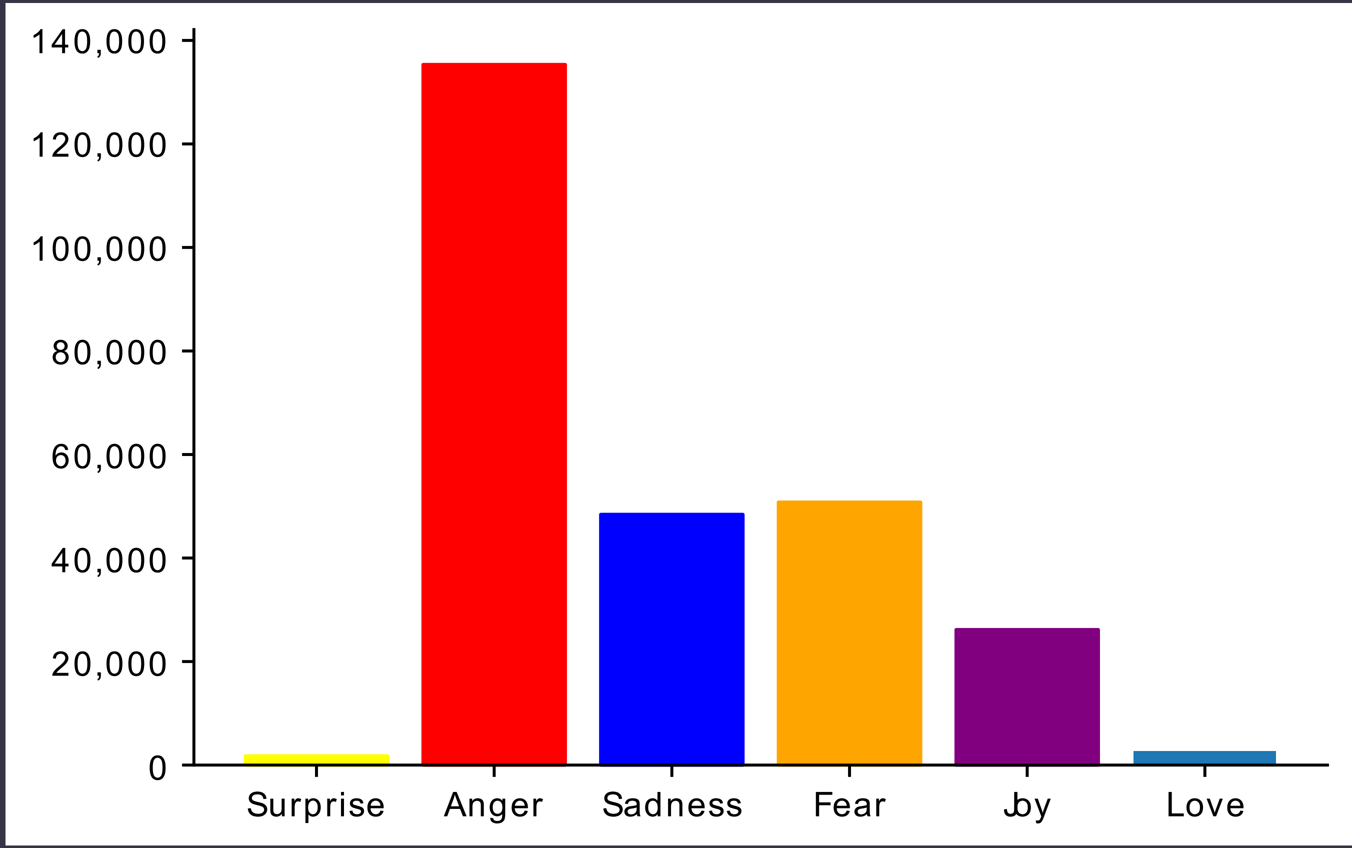
Frequency distribution of tweets according to VADER.



Frequency distribution of tweets according to SentiStrength.



Frequency distribution of tweets by emotions classified by LSTM Model.



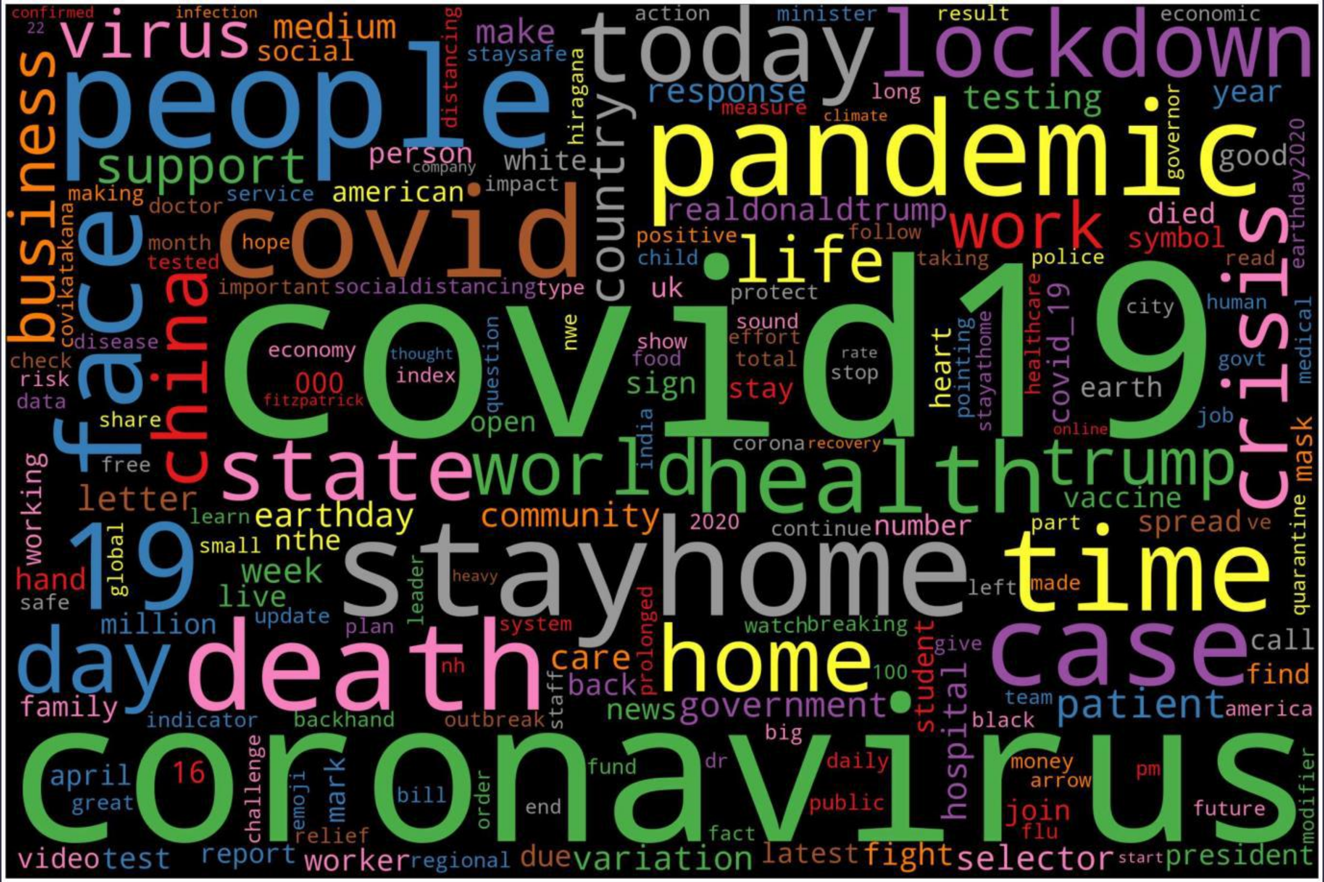
TF-IDF Score 2020

Term	TF-IDF Score
covid19	44186.93
coronavirus	30987.27
people	12507.46
stayhome	11591.37
pandemic	10836.62
covid	10018.76
death	9582.968
time	8636.371
face	8414.087
case	8098.131
today	8072.942
health	7999.169
lockdown	7264.073
home	7140.71
day	7002.378
crisis	6670.096
state	6634.335
world	6533.159
china	6238.218
trump	6235.947

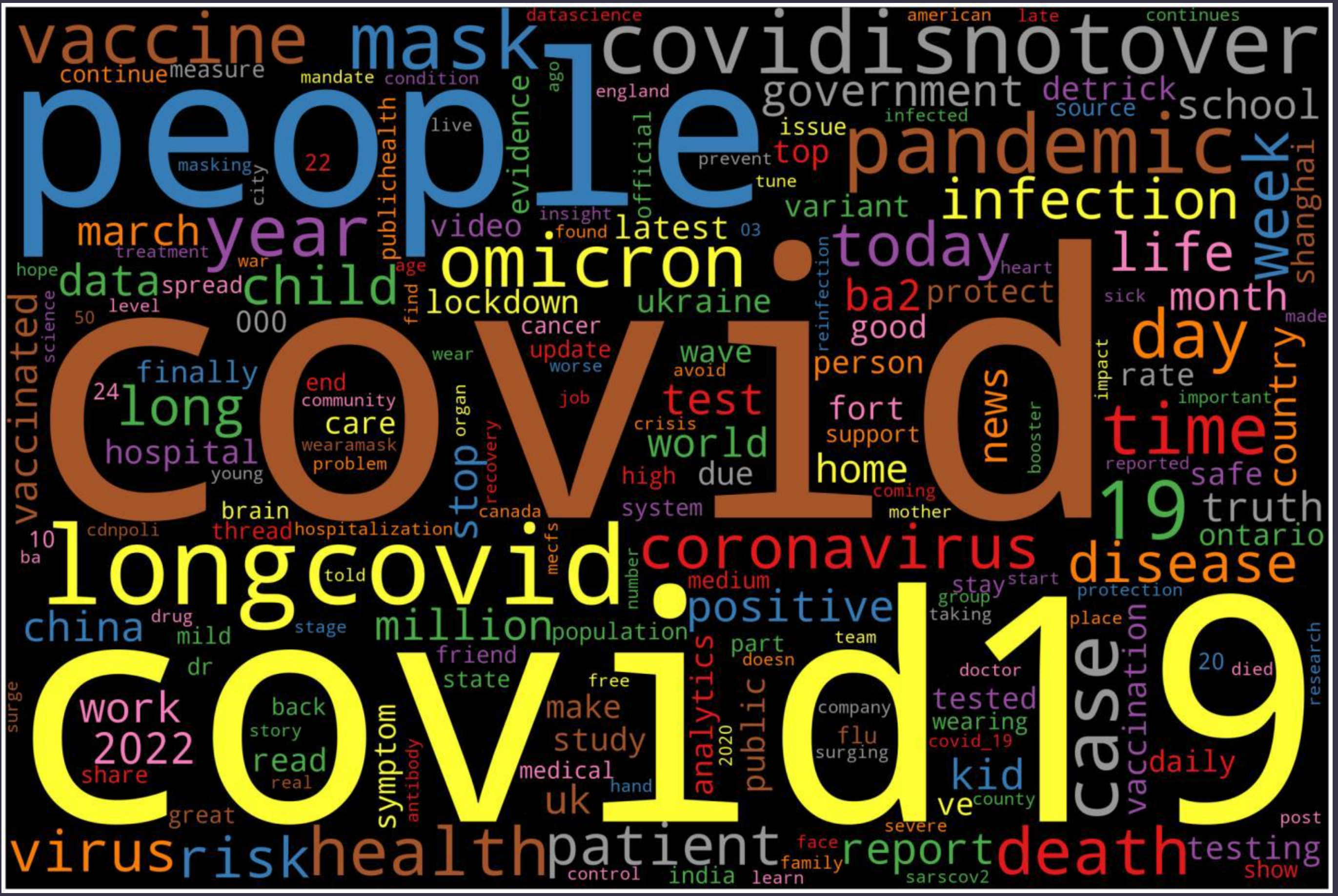
TF-IDF Score 2022

covid	24479.99
covid19	24153.37
people	10173.2
longcovid	9681.456
covidisnotover	9355.41
pandemic	8398.545
mask	8001.38
case	7680.718
omicron	6746.918
year	6197.592
time	6165.984
vaccine	6087.344
day	5636.658
health	5593.961
death	5279.887
week	4805.976
today	4568.455
coronavirus	4556.032
patient	4176.897
virus	4138.394

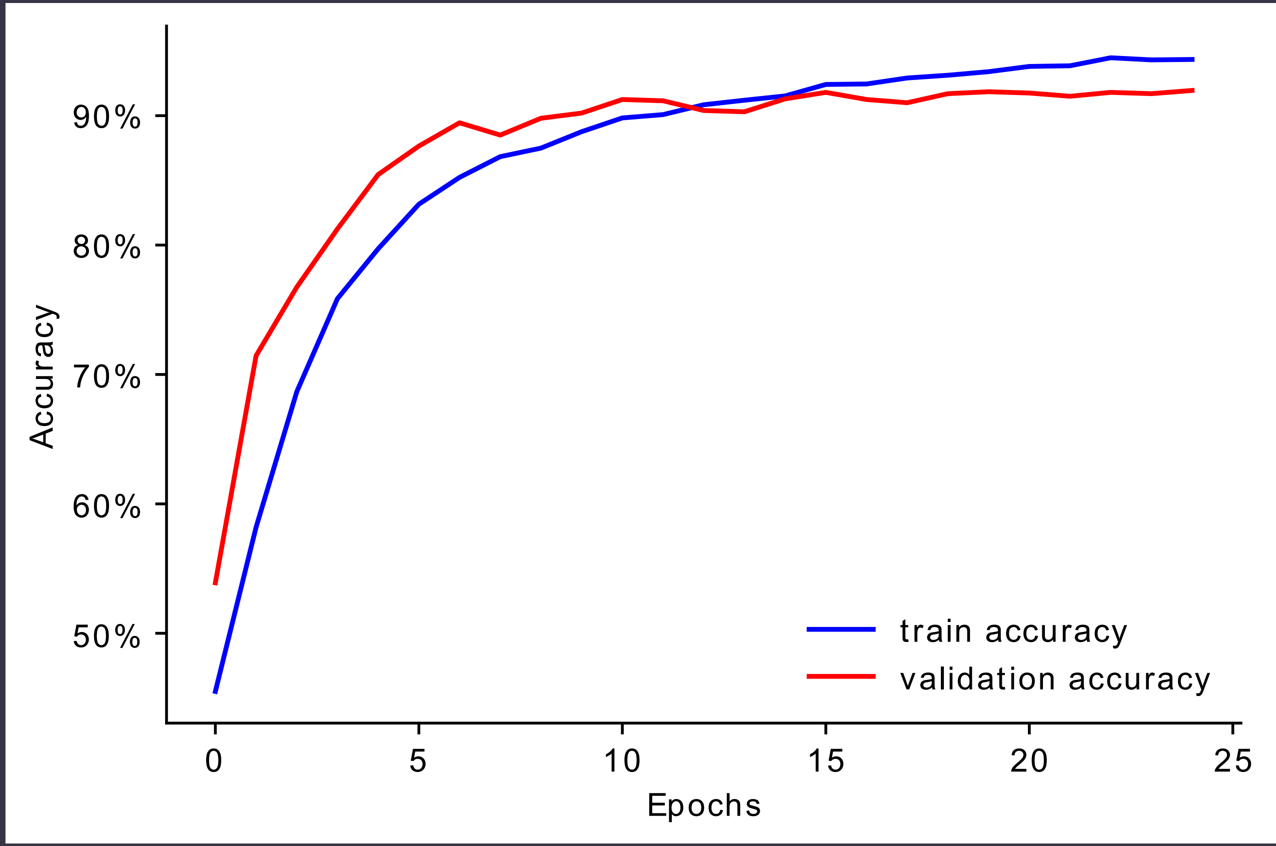
WordCloud 2020



WordCloud 2022



Train vs Validation Accuracy

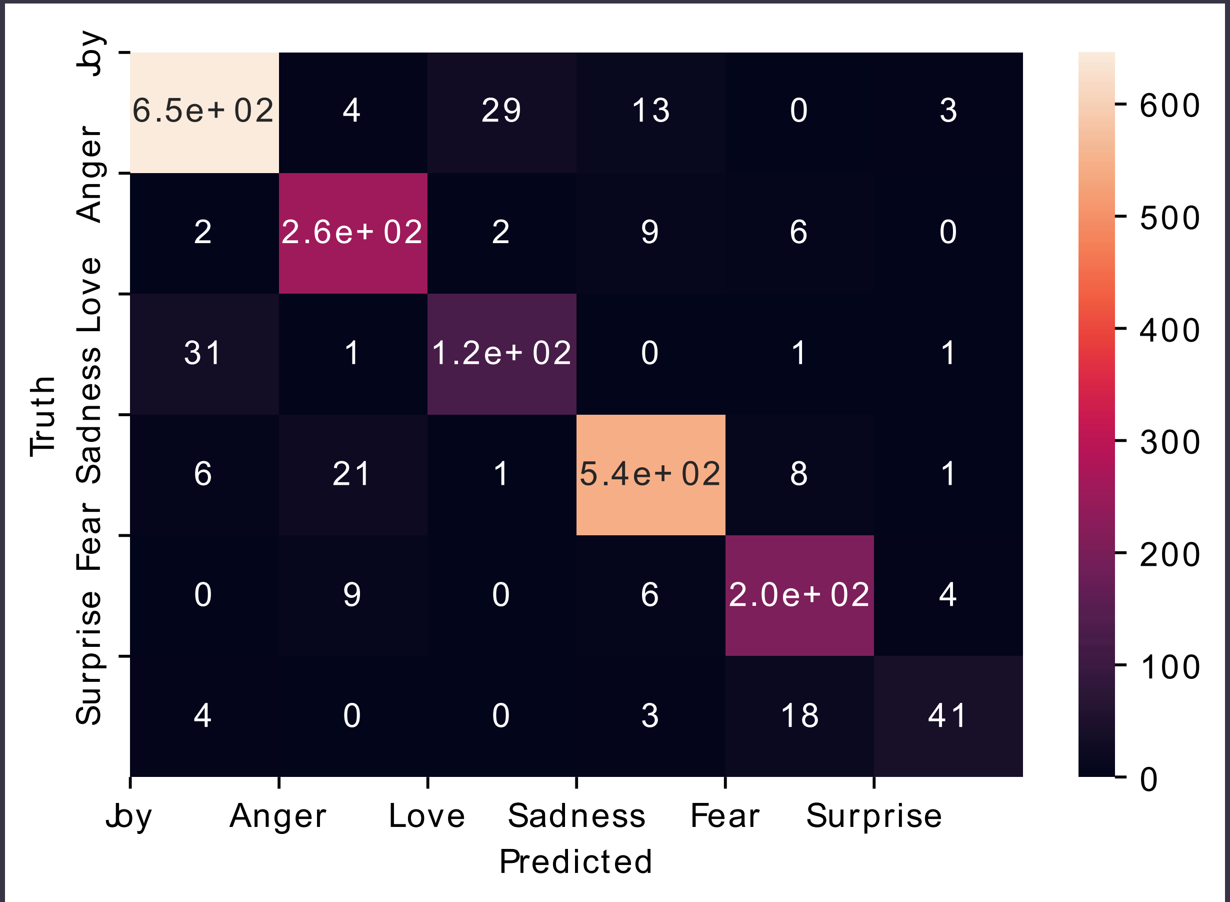


Performance Metrics

```
y_pred =   np.argmax(model.predict(X_test_pad), axis  =  1)
y_true = np.argmax(y_test, axis = 1)
from sklearn import metrics
print(metrics.classification_report(y_pred, y_true))
```

```
              precision    recall  f1-score   support

           0       0.93      0.94      0.93       689
           1       0.93      0.88      0.90       291
           2       0.79      0.80      0.79       157
           3       0.94      0.95      0.94       575
           4       0.92      0.86      0.89       238
           5       0.62      0.82      0.71        50

    accuracy                           0.91      2000
   macro avg       0.85      0.87      0.86      2000
weighted avg       0.91      0.91      0.91      2000
```

Confusion Matrix



CONCLUSION

Analysing the various histograms and figures related to Ekman, Plutchik, POMS and Text2Emotion tools we can clearly see that there is not much change in the mood of people towards COVID-19 from 2020 to 2022. The same emotions which were dominant in 2020 continue to remain dominant in 2022. What was fascinating to observe was that Joy was the most dominant emotion according to the Ekman and plutchik models. This is quite odd as COVID-19 was associated with death, sickness, discomfort and a departure from the normal way of life, not something that denotes 'Joy'. Text2Emotion painted a gloomy picture with fear being the most dominant emotion in both 2020 and 2022, and POMS denoting that depression being the most dominant emotion in both 2022 and 2020. There are slight differences between the mood from 2020 and 2022, but the general trend remains constant.

When looking at the sentiment, according to Textblob most of the tweets were classified as neutral in both 2020 and 2022, however according to SentiStrength most of the tweets were classified as negative and positive according to VADER.

CONCLUSION

The most prominent terms in 2020 wordcloud include covid19, coronavirus, pandemic, lockdown, virus, stayhome, death, etc. The most prominent terms in 2022 wordcloud include covid19, longcovid, covidisnotover, omicron, vaccine, mask, etc. Clearly, after two years the terms such as stayhome have lost prominence.

When it comes to the RNN model it classifies the majority of tweets as anger, followed by fear and sadness. This is major departure from the emotions denoted by our Ekman and Text2Emotion models.

We were successful in building a model that predicts the emotions of given tweets, with an exceptional accuracy score. One pertinent observation we noted was that our RNN model predicted anger to be the most dominant emotion in 2022. Compared to 2020, fear was the most dominant emotion according to Text2Emotion. This makes sense as with the passage of time, people may have gotten accustomed to and fed up of COVID-19.

AREAS FOR IMPROVEMENT

Even though this project was successful in designing a model to predict emotions to a reasonably high degree of accuracy, in hindsight, there are still some measures we can undertake to enhance the project. The first one would be to balance the training dataset as some emotions were overrepresented. This would have further improved the classification accuracy of the model. Another approach would be to use a larger dataset to train our model. The current dataset used to train our model only has 16,000 labelled records. This would further boost the performance of our model. Also, if the training dataset was domain specific and related to covid-19 it would have made the resulting model extremely powerful. Another modification we could have made was to try the BERT model to predict our emotions. BERT which is a transformer model, is a recent development that can be tested for our purpose. Lastly, we could have user other metadata collected from twitter such as location, number of followers and used that to perform data analysis and group emotions based on these factors.

CHALLENGES

Research on emotion detection on text pales in comparison to emotion detection, on facial expressions, body language, etc. The main reason is that textual data has many variables such as context, sarcasm, slang words, grammatically incorrect words, etc. This hinders any emotion detection model to a considerable extent.

FURTHER APPLICATIONS

Even though the primary objective of our model was to gauge the emotions of people regarding covid-19, this emotion model can be utilized in a variety of domains to tackle difficult problems. It can be used in the field of mental health to detect signs of depression, sadness and so on. Furthermore, it can be integrated in chatbot applications to provide psychiatric counselling and support to patients and thereby play an enormous role in suicide prevention. It can be also used on social media sights to flag hate speech and racist content. This will greatly help in attenuating the problem of cyberbullying, which is becoming widespread nowadays.

REFERENCES

[1] M. A. Palomino and A. Padmanabhan Varma, "Any Publicity is Good Publicity: Positive, Negative and Neutral Tweets Can All Become Trends," 2020 39th International Conference of the Chilean Computer Science Society (SCCC), 2020, pp. 1-8, doi: 10.1109/SCCC51225.2020.9281266.

[2] N. Colnerič and J. Demšar, "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model," in IEEE Transactions on Affective Computing, vol. 11, no. 3, pp. 433-446, 1 July-Sept. 2020, doi: 10.1109/TAFFC.2018.2807817.

[3] Saravia, Elvis & Liu, Hsien-Chi & Huang, Yen-Hao & Wu, Junlin & Chen, Yi-Shin. (2018). CARER: Contextualized Affect Representations for Emotion Recognition. 3687-3697. 10.18653/v1/D18-1404.

THANK YOU!