A State of Mind Analyser

(Based on a Corpus of Tweets)

About the Project

- “A State of Mind Analyser” is based primarily on the concept of “Geotagging Based Sentiment Analysis”.
- “Geotagging” here refers to addition of geographical identification metadata to a “Corpus of Tweets”
- “Sentiment Analysis” refers to identify, extract, quantify, and study affective(experience of feeling, emotion or mood) states and subjective information.

Scope of the Project

- "A State of Mind Analyser" is a **web application**, in which when inputted "**Geographical Location**" and "**Time Period**" from a given available set, displays the **"overall" sentiment expressed by people** on Twitter, a popular microblogging platform.

- It can be **used to analyse the trends in the dynamic mood of the population.**

Need of "the hour"

•Due to the outbreak of COVID-19, in a country with a large, diverse population like India, there are bound to be instances of **mass hysteria** and **panic** which are further fuelled by **unreliable and sometimes inaccurate data.**

•Social media acts as the bridge between the people, the government, and such organizations.

•Determining the emotions of the citizens by the government would :

1.Provide **insights into the public mind-set**

2.Pave the way for the government and many organizations to **address these situations**

3.Provide **right data and information**

4.Help in **suppressing unnecessary panic** among the people

•**Acts as a sanity check for the effectiveness of the adopted government policies.**

•After Analysis, **amendments can be made to the decisions taken by the regime policies**, and can be made in such a way so as to enhance the sentiment towards a positive outlook.

•**Help NGOs and various organizations** to come forward to help the people.

•**Businesses can adapt their products and services** to match the requirements of the people based on the trending mood of the public, which will not only **help businesses to grow but will also help the public meet their need of the hour.**

•**Shifts in sentiment on social media correlate with shifts in the economics of a country**

•Govt. can thereby make **business and people-friendly rules and laws to help in the betterment of the economy** and the market in these untested times.

Dataset

•Source:

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

•As of December 9, 2020: Consists of 264 csv files, **updated on a daily basis wef from March 20, 2020 to present**, with average size of about (25-30)MB.

•For the sake of simplicity and to test our existing data visualization and summarization techniques, **a part of the whole dataset(1200 entires)** has been taken into consideration.

•The **original dataset consists of 34 attributes** which has been further **truncated down to 7 attributes**, discarding the ones which are out of scope for this project.

```
df.dtypes

created_at     object
id              int64
lang           object
place          object
source         object
text           object
user_name      object
dtype: object
```

Test Plan - Phases

PHASE 1

Data Exploration

Data visualization and statistical techniques to describe dataset characterizations in order to better understand the nature of the **data.**

Data Summarization

PHASE 2

Data Pre-Processing

Labeling the Dataset (Positive, Neutral, Negative Sentiments)

Applying and Optimizing different ML models

PHASE 3

Evaluating and Testing

Database Creation

Backend Web Development

PHASE 4

Dashboard Creation

GUI for project

Deployment

Phase 1 - Data Exploration and Visualisation

1
- Drop/**Truncate** irrelevant attributes.
  (34 -> 7)

2
- Handling **Missing Values**

3
- Exploring **Unique Values**

4
- For Geotagging, grouping by country using package: **pycountry**

5
- **Visualisation:**
  - Top 20 Locations by Tweets – Bar Graph
  - Top 20 Twitter Sources – Bar Graph
  - WordCloud: Most Frequently Occurring Words

Phase 1 - Data Exploration and Visualisation

Fig 6.1Top 20 Locations by Number of Tweets:

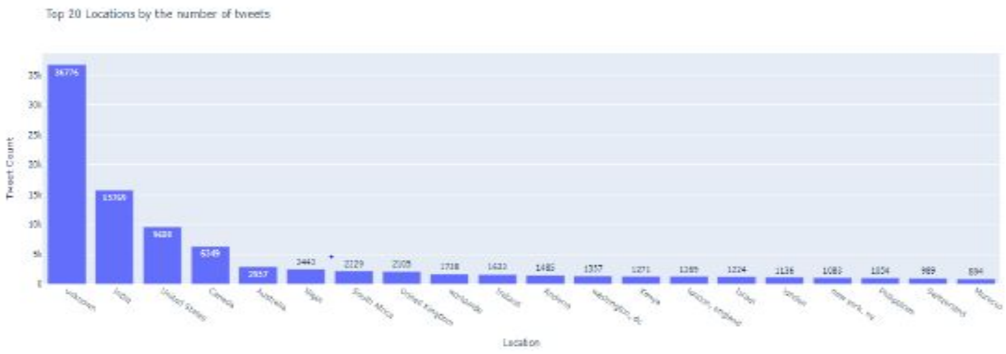


Fig 6.2 Top 20 Sources by Number of Tweets:

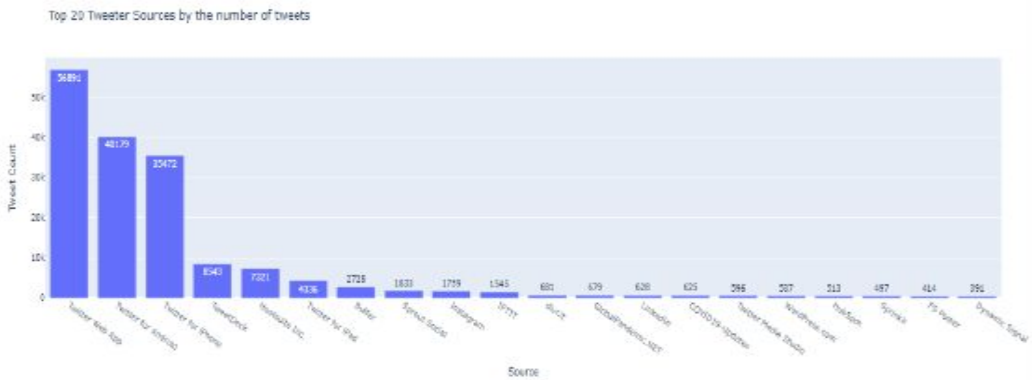
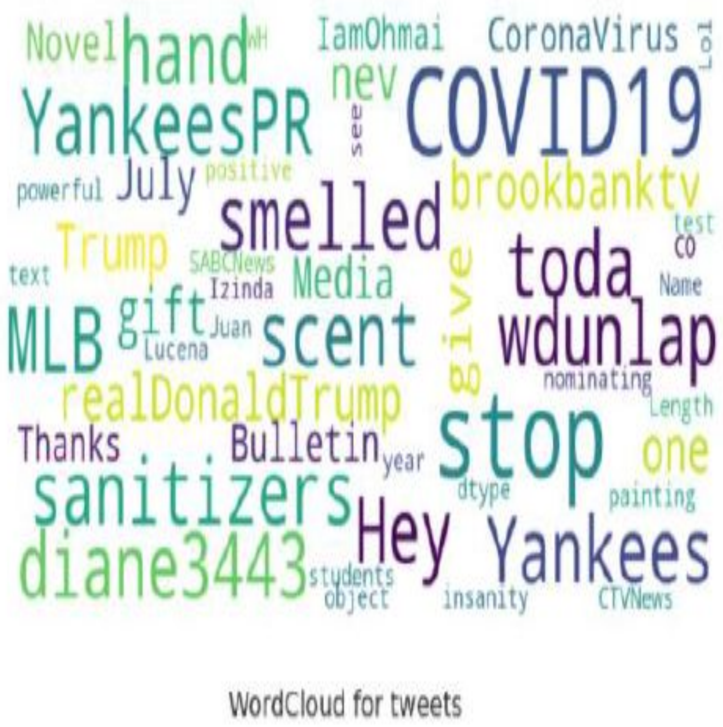


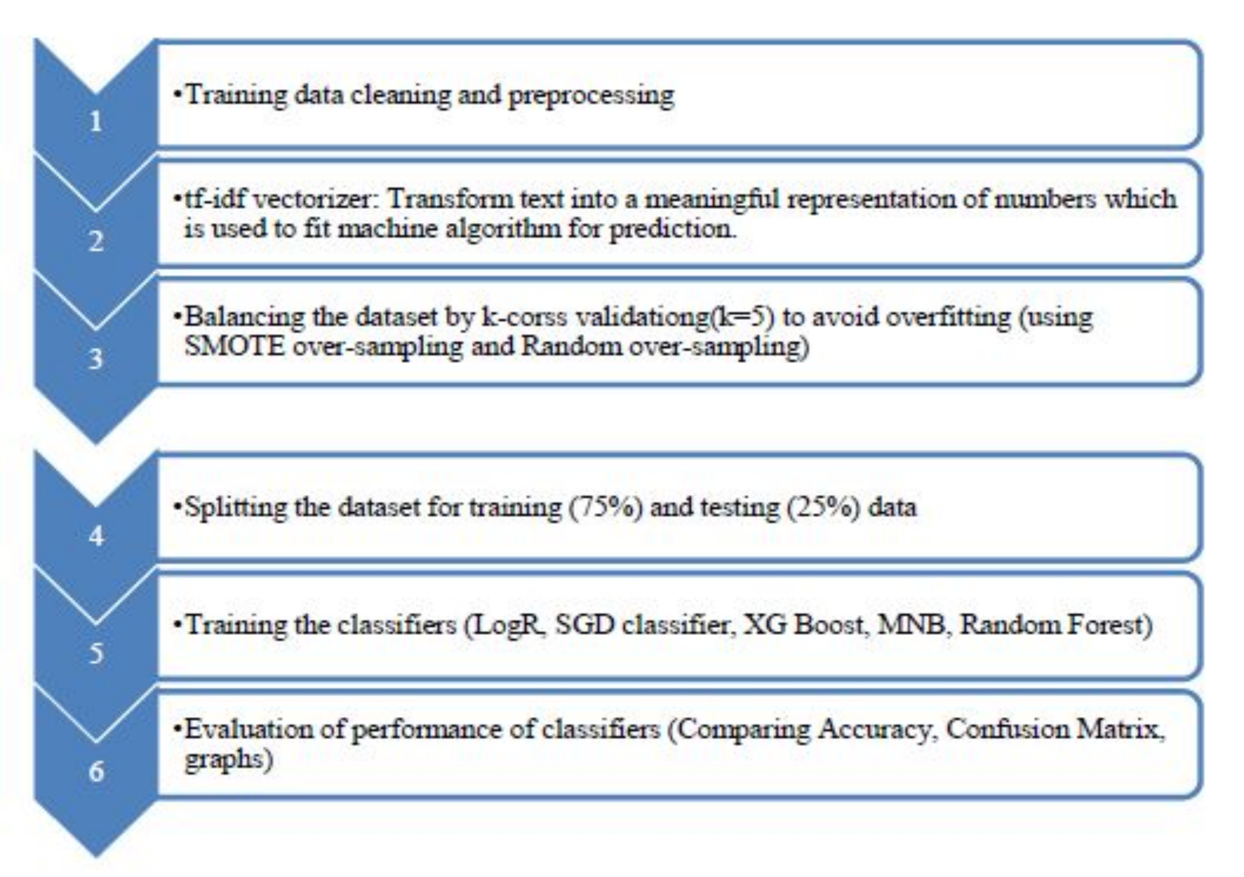
Fig 6.3 Word Cloud for Prevalent Words:



Phase 2.1 - Data Cleaning & Preprocessing

- Using the package **tweet-preprocessor** we have dealt with **cleaning, tokenizing and parsing** dataset:
    - URLs
    - Hashtags
    - Mentions
    - Reserved words (RT, FAV)
    - Emojis
    - Smileys
- From **nltk**, we have imported **Stopwords** from **nltk.corpus** and then removed them from the tweets.

Phase 2.2 - Training the Classifiers



Evaluation Results

Classifiers	No Sampling		SMOTE over-sampling		Random over-sampling	
	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy	Training Accuracy	Testing Accuracy
LogR	84.71	82.72	83.36	74.4	82.02	72.67
SGD	85.37	82.84	84.11	73.41	82.56	73.5
XG Boost	90.65	84.91	90.91	80.7	90.86	81.04
MNB	81.55	80.47	76.39	71.24	75.37	69.77
Random Forest	97.83	88.98	96.89	86.5	95.91	85.48

Table 6.1 : Accuracy of Classifiers (refer to Appendix 2)

Classifiers	Training Accuracy	Testing Accuracy
LogR	84.79	83.99
SGD	85.8	85.1
XG Boost	90.55	85.19
MNB	81.96	81.42
Random Forest	97.82	90.25

Table 6.2 : Confusion Matrix Accuracy of Classifiers (refer to Appendix 3)

Comparison between Classifiers

Results

- We have observed that **Random Forest** without sampling has performed the best with an accuracy of **97.82% (training accuracy) and 90.25% (testing accuracy).**
- The reason being Random Forest is very suitable for dealing with high dimensional noisy data in large databases in text classification
- Other than we have observed that **XGBoost** classifier works well after using oversampling technique with an accuracy of **90.98% (training accuracy) and 80.82% (testing accuracy).**
- The reason being that XGBoost tunes the parameters by itself and works in a parallelized fashion distributed among clusters. Therefore, the balanced the data in the clusters, the better it performs.

Libraries Used

- **Numpy:** used for working with arrays. It also has functions for working in domain of linear algebra, matrices, etc.
- **Pandas:** Data analysis and manipulation. Key data structure is DataFrame. Built upon NumPy.
- **Matplotlib:** plotting library. Draw inline plots for quick data analysis (basic plotting)
- **pycountry:** ISO databases for the standards: Countries, Subdivisions of countries,etc
- **Seaborn:** data visualization library. Drawing attractive and informative statistical graphics
- **WordCloud:** data visualization technique used for representing text data in which the size of each word indicates its frequency
- **Sklearn:** machine learning library
- **Imblearn:** toolbox for imbalanced dataset in machine learning.
- **Stratifiedkfold:** Provides train/test indices to split data in train/test sets
- **Joblib:** provide lightweight pipelining in Python. Simple parallel computing. logging and tracing of the execution, etc.
- **Counter:** count the key-value pairs in an object
- **Pickle:** for serializing and de-serializing a Python object structure. Any object in Python can be pickled so that it can be saved on disk.
- **Tweet-preprocessor:** cleaning, tokenizing and parsing dataset
- **Nltk:** to work with human language data
- **Tfidfvectorizer:** Convert a collection of raw documents to a matrix of TF-IDF features
- **StandardScaler:** standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation.

Risk Analysis

- Tweets are usually coupled with **hashtags, emoticons and links**, creating difficulties in determining the expressed sentiment.
- Requirement of **large datasets of lexical databases** where “emotional words” are associated with “sentiment values” (quantifiable).
- Presence of **unclear or scarce datasets** and **lack of labelled data** can pose a barrier to the advancements in the area of sentiment analysis.
- Problem in **recognizing human aspects of a language** like irony, sarcasm, negotiations, exaggerations, and jokes. This can lead to skewed and incorrect results.
- For example, the terms “fight” and “positive” are used in a negative and positive context respectively, but we observe a **role reversal** in this situation. The identification of such terms and their usage according to the context would be essential.
- **Lack of classifiers for multi-linguistic data, or data in any other language except English.**

Future Scope

•Collection of a **multilingual corpus** of Twitter data and build a **multilingual sentiment classifier.**

•Exploration in **active learning techniques** to detect Twitter sentiments

•**Hidden or veiled sentiment detection, satire detection, comparison or association handling and emoticon detection.**