Livability Index of Indian States

Yash Mishra	16BCE0747
Vijay Kumawat	16BCE0913
Vaibhav Bangwal	16BCE0964

Submitted to
Professor Anuradha G, SCOPE

School of

School of Computer Science and Engineering

LIST OF ABBREVIATIONS

Abbreviation	Expansion
SVM	Support Vector Machine
LCI	Livable City Index

1. Introduction

By 2050 India is projected to add 416 million urban dwellers to the world's urban population and will be home to about 58% of the total global population.

Structurally, urbanization is advantageous to India on several fronts. Urban areas contribute to approximately 62-63% of India's Gross Domestic Product (GDP), which is estimated to reach 75% by 2030. McKinsey research estimates that cities could generate 70% net of all new jobs by 20306. It presents an opportunity to reduce social inequities which are much less pronounced in urban agglomerations compared with rural areas, since hierarchies are driven more by economic (rather than social) standing in cities. It also serves as a natural focal point for the adoption of new technologies and innovation en masse. Additionally, it creates large markets with critical mass for a variety of goods and services, catalyzing the overall economy and prosperity of a developing country like India.

1.1 Background

"Home is where the heart is, but what if your heart doesn't know where it should be?"

From low crime rates to a great education system, there are many variables to consider when choosing that perfect place that you and your family can call home. To help you make this important decision, we have taken up this project to help you find a home that suits the needs of you and your family. Begin by determining what is most important to you and your family. If you're single, living in a bustling city might be an ideal choice for your next home. If you have a family, on the other hand, a small town offers amenities that your kids will love. Therefore, this question is very user-dependent and no solid or concrete answer can be given. It is up to the client to understand all the work done in this project and weigh-in the factors which is important to him.

1.2 Motivation

However, the rapid pace of urbanization and the increased number of urban dwellers could exacerbate existing challenges like pollution, overcrowding, rising crime levels, poor access to water supply and sanitation facilities, and congestion, among others. This warrants a greater focus on improving the governance and the quality of urban infrastructure and service delivery, which have a direct bearing on the quality of life offered by the cities to its citizens.

1.3 Aim of the Project

Finding the livability index of Indian States and compare all the other states on the basis of index.

2. Literature Review

2.1 The structure of online social networks mirrors those in the offline world

Abstract: The authors of this paper begin by saying that species with larger neocortices manage to maintain coherence in larger groups than with smaller neocortices. They further state studies which claim humans to have strong dyadic relationships. They support this claim by studies which show that communities have strong socialization with 200 members and members higher than this number tend to be passive in their contributions. Studies have also found that humans do not combine and distribute their social effort evenly which further supports the above claims.

The objective of the study was to find how truly online relationships represent the offline ones. After stating known facts about offline relationships (mentioned above), the authors created three datasets (two from Facebook and one from Twitter) to study online relationships of humans.

Facebook data which was not restricted by privacy regulations were used and this resulted in more than 3 million nodes and 23 million edges and above. Due to privacy rules, 44% profiles came under the non-public domain and to tackle this issue, 44% of the nodes were randomly selected and declared as non-public. The magnitude of the dataset being so large, nullified any discrepancies. The dataset was divided into four dynamic moving windows for analysis. When ego networks were extracted with DBSCAN (density based clustering) with clustering size taken 4 for Facebook data and {4,5} for Twitter data, it was that most of the ego networks had size less than 100 and these ego networks were further refined using the offline convention of one message per annum. For Twitter, frequency of contact between two users (f_r) was said to be the ratio between Number of replies to duration of relationship. The authors discounted company and brand profiles to only include true human interactions.

The Complementary Cumulative Distribution Function of the contact frequency showed a long tail which indicated low contact rate for most of the distribution but few relationships have very high level of interactions. This finding resonated with the result of studies done on offline relationships. Three Layers were expected according to minimum frequency of contact but a new layer, Layer 0, was a revelation with it showing almost 1.5 unit (meaning people have very high investment with approximately 1.5 other people).

To conclude, the tendency found in the online study resonated with the tendency for individuals to have one or two intensely intimate friends which is evident in other smaller datasets where contact frequencies have been plotted in order. Though this work is pretty exhaustive in nature but suffers from minor setbacks like gender identification and its impact in social interactions.

2.2 Sentiment Analysis for Social Media: A Survey .

Abstract: The paper was taken from IEEE Xplore where it is stated to have been published in 2015 2nd International Conference on Information Science and Security (ICISS).

It provides a comprehensive insight of how sentiment analysis has developed over the years and what its present state is in the world of science. It states that since SA depends heavily on "positive" and "negative" words, it is unwise to go for pure lexicon analysis and recommends the use of various classification algorithm like Support Vector Machine, Naïve Bayes etc. Other approaches with very good results were Hybrid Approach-where Machine learning algorithm was mixed with classical lexicon approach- and Latent Dirichlet Allocation (LDA) which uses two tables, where table1 represents the chance of selecting a particular part when sampling a particular topic and table2 represents the chance of selecting a particular part when sampling the document. Unlike K-means where something can belong to only one cluster, LDA creates soft clusters. The paper also discusses the challengers faced by Sentimental Analysis, in which incorporation of new data set and tackling the ever-changing sentiments pose the biggest problem.

2.3 Mining Social Media Data for Understanding Students' Learning Experiences

Abstract: This paper was published in: IEEE Transactions on Learning Technologies (Volume: 7, Issue: 3, July-Sept. 2014)

This paper goes beyond the use of Sentiment Analysis to create a holistic understanding of the educational difficulties faced by students as unlike room activities, focus groups and surveys which are expensive and unrealistic for a high frequency practise, this paper uses social media as source for understanding as it provides unmonitored behavioural outlook.

The authors have evaluated their classification on human feedback and interpretation which has enabled a deeper understanding of the data. Since Sentimental Analysis only tells us about positivity, negativity or neutrality in text, it is insufficient to gather deeper insight into student's problem which lead the authors to create a Naïve Bayes based multi-label classification model where tweets were allowed to fall into more than one label at a time. The dataset was collected from Radion6, a social monitoring site, and official dataset of a study conducted by researchers of Purdue University. Like previous paper, they have also used Latent Dirichlet Allocation for soft cluster formation.

In conclusion, this paper provides us with a comprehensive study of the issue it is dealing with along with the various novel techniques incorporated by them like using statistical measures like Cohen's Kappa, Scott's Pi, Fleiss Kappa, Krippendorf's Alpha along with machine learning classifiers.

2.4 Development of a Liveable City Index (LCI) Using Multi Criteria Geospatial Modelling for Medium Class Cities in Developing Countries

Abstract: The authors tried to develop a livability city index(LCI) based on various factors and GIS techniques. A three-stage survey was performed to evaluate the nine significant factors (Safety, Economy, Environment, Education, Health, Transportation, Recreation, Population Density, and Public Utility) using the analytic hierarchy process (AHP).

AHP is a structured technique for organizing and analyzing complex decisions, based on mathematics and psychology. It was developed by Thomas L. Saaty in the 1970s and has been extensively studied and refined since then. It consists of 5 major steps:

1. Model the problem as a hierarchy containing the decision goal
2. Establish priorities among the elements of the hierarchy by making a series of judgments based on pairwise comparisons of the elements.
3. Synthesize these judgments to yield a set of overall priorities for the hierarchy.
4. Check the consistency of the judgments.
5. Come to a final decision based on the results of this process.

Their study implemented a GIS based multi-attribute decision making analysis for developing LCI.

The first stage of the survey (out of three), identifies marked variations in attitudes towards the liveability of a city. In the second stage, experts and city managers were invited to rank each factor and decide on sub-factors for each factor to be used in the AHP process. In the third stage, a questionnaire was developed to generate a multi-factor evaluation of the factors: Safety, Environment, Healthcare, Transportation, Public Utility, Education. Economy, Population Density, Recreation and their sub-factors.

Limitations

They faced three prominent limitations:

2.5 The study is centered around the physical factors, avoiding human emotions.
2.6 Second, the recreation factor can be different for different people, thus it is highly subjective which makes it difficult to be evaluated.
2.7 Third, the physical data which is collected and analyzed is highly volatile and changes frequently, so will not be up to data if the study time is too long.

To deal with the third limitation, a plugin can be developed in any of the GIS open source environment (e.g. White box Gat, Saga GIS) which will automate the process and thus addressing the issue of volatility of the physical data.

2.5 Massive Social Network Analysis: Mining Twitter for Social Good

Abstract: The authors used GraphCT to analyse massive public data from twitter for news dissemination by ranking actors within the conversations, thus providing a much smaller subset for analysis. Graph has been treated as undirected and an algorithm similar to the Kahan's algorithm is used, for parallelism, environment such as OpenMp is used.
In uenza H1N1 Tweets in September 2009, Atlanta Flood Tweets in September 2009, are examples of the tweets on which the analysis is performed.

Algorithm: GraphCT extracts connected components from the graph though a technique similar to Kahan's algorithm. In the first phase searches breadth-first simultaneously from every vertex of the graph to greedily color neighbours with integers.
The higher labelled colors are then coloured into lower labelled colors and the process goes on once there are no more collisions.

Limitations: quantifying significance and confidence of approximations over noisy graph data.

2.6 Monitoring Public Health Concerns Using Twitter Sentiment Classifications

Abstract: In this paper, the authors have presented tweet classification approach to identify negative sentiments of personal health tweets to measure the degree of concern (DOC) for monitoring the public sentiments for a disease using Twitter.

The paper emphasizes on the degree of concern for a disease and its spread.
Epidemic Sentiment Monitoring System (ESMS) monitors two primary bacterial diseases 1. Listeria 2. TB and two viral infectious (swine flu, measles). Clue based

searching and Machine Learning classification methods are used to filter out the tweets into personal and non-personal. Four-point Likert Scale was used instead of two-point Likert scale because positive emotions can arise as a result of relief about an epidemic subsiding. P-corpus results are better than W-corpus' results for personal tweets because p-corpus has a better identification rate of personal tweets. Multinomial Naïve Bayes gave the best classification result among all the Machine learning classifiers (Support Vector Machine, Naïve Bayes). This trend of Multinomial Naïve Bayes outperforming other classifier was not at all surprising because all other previous papers which our team has studied follows the same trend. The authors have extended their study by generating concern maps for geographic visualization which enhances one's understanding of the sentiments people regarding the disease regionally.

To conclude, this paper introduced us with a new classifier model, Multinomial Naïve Bayes, to classify the tweets and our team needs to further study the key differences between Naïve Bayes and its elder sibling to see which would fit best in our project work.

2.7 Social Media as Research Instrument for Urban Planning and Design.

Abstract: In this paper by Wu Wien and Wang Wei, the impact of social media has been assessed and how easy it has become for urban model planning and design with the massive dataset that today's pervasive usage of social media could provide. The authors have rightfully and shrewdly pointed out how Social media today enjoys a fundamental stand in data acquisition platforms for its distinctive user generated content (UGC) and the researchers have to simply acquire them with the necessary tools. Dataset collected from Social Media serves as a perfect research sample for urban planning and design in that subjective descriptions, images and coordinates are accumulated with objective quantity.

Study on the data of Social Media has confirmed that "Six Degree" theory by Milgram who stated that everyone in the world is connected to one another by six people and not surprisingly, researchers have confirmed this by officially declaring – "Median of Five Intermediate acquaintances is needed for first person to know the other."

Presence of social media is not just restricted to Twitter and Facebook which find their utility in sentiment analysis and opinion studies among masses but applications like UBER also tells the researchers in which part of the city the researchers are most likely to go to or the route they frequently travel on, thus helping them augment traffic control systems.

Nevertheless, the authors have admitted that on-situ interviews have an edge over

Social Media data but their non-feasibility outweighs their accuracy.

In conclusion, our recognition of the urban orders cannot be fixed as time changes and technology advances. In this Big Data Era, social media reveals human activities and patterns in the living environment regardless of time and space.

2.8 Talking about Climate Change and Global Warming

Abstract: In this paper, the author, Hayley J. Flower, discusses about the two terms ‘Climate Change’ and ‘Global Warming’. Both, Climate change and Global Warming are unilateral terms which indicate the level of awareness. Frequency can’t measure the popularity because it does not indicate if a search term is common or uncommon. Relative Search Volume(RSV) is used for checking that which term is relatively used more. Tweets containing these two phrases is collected from the Twitter API and this data is pre-processed for sentimental analysis.

“Semantria” was used for sentimental analysis as it helps in text analysis in Microsoft Excel sheet. It identifies sentiment-laden phrases and scores them from -10 to 10 on a logarithm scale. Statistical inference of these tweets is determined and polarity is assigned (Positive, Negative, Neutral).

N-gram method is used to compute the tweet term frequency for both phrases. Pettit and MK tests were used to identify changes in distribution using the software XLSTAT. Temporal trends within the time series were analysed with Spearman’s non-parametric correlation analysis. Pearson’s chi square test was performed for checking the accuracy and reliability of the sentiment analysis.

To conclude, it was analysed that the term global warming is used more frequently and is reflecting more negative sentiment over the people than the term climate change. Our team also learned about two potentially useful software: XLSTAT and Semantria from this paper which we can use for our J-Component.

2.9 Detection of Influential Nodes Using Social Networks Analysis Based on Network Metrics

Abstract: In this paper, the author has studied various parameters which are used to detect the key players, also known as, influential nodes. The parameters he studied were: Coefficient clustering, Density, Closeness, Centrality, Degree Centrality, Page Rank, Eigen Vector Centrality. The choice of dataset the author chose was strikingly very different as he took the relationship data of Indonesian Noordin’s Crisis which was an illegal network and required a Covert Data Analysis. Covert Networks are those where if node A has a relationship with node B then node B must have a relationship with node A. The data on the network is available out in the public and consists of 75 nodes and 397 edges. Microsoft's NodeXL tool was used to study all the above-mentioned centrality parameters. The author weighed in all the parameters to decide his most influential node in this case and it turned out to be the leader of this terrorist group and his name was "Noordin Mohmmad."

To conclude, the fact that all the parameters, if considered, directed towards the leader of terrorist group to the most influential node, in itself validates the study. Our team found this work very beneficial as many new parameters which can decide the most influential node were introduced and we plan on incorporating them in our J-component work.

2.10 Analysis of Twitter Hashtags: Fuzzy Clustering Approach

Abstract: In this paper, the author analyses importance of Twitter hashtag.
A hashtag is a label used on social media sites that makes it easier to find information with a theme or specific content. The data can be obtained from the "Hashtagif.me". Fuzzy clustering methodology is used to analyse the hashtag. The points quite inside the cluster have the high values of memberships while the close to the boundaries have the small value of memberships. For quality measurement of cluster silhouette width is calculated. The silhouette width allows us to identify closest cluster to a point *i* outside the cluster *k*. Positive values of silhouette indicate good separation of clusters. For visualization of multi-dimensional CLUSPLOT approach is used which is based on a reduction of the dimension of the data by Principal Component Analysis(PCA). Each cluster is drawn as a spanning ellipse. To analyse pattern of hashtags, change in the percentage of popularity is calculated and compared with the others.

To conclude this paper, fuzzy clustering helps to recognise the most popular hashtags which can help to understand user's interest and many other useful information like facto,

3. Methodology

a) Sentimental Analysis of each state and union territories of India (total thirty-five in count) using Twitter to calculate the sentiments on ten (10) categories:

i. Health Facility
ii. Job Opportunity
iii. Education
iv. Crime Rate

v. Transportation
vi. Food
vii. Climate
viii. Culture
ix. Environment
x. Estate Value

b) Official Data from Indian Government and other reliable data source like Kaggle to check the reliability of sentimental data with actual data.

c) Representation of scores of each state and its city by the official Indian Government's scheme run under the shelter of Ministry of Urban Development to rank each city.

4. Tools Used

4.1 R Studio for Sentimental Analysis

Sentiment analysis is the computational task of automatically determining what feelings a writer is expressing in text. Sentiment is often framed as a binary distinction (positive vs. negative), but it can also be a more fine-grained, like identifying the specific emotion an author is expressing (like fear, joy or anger).

Sentiment analysis is used for many applications, especially in business intelligence. Some examples of applications for sentiment analysis include:

- Analyzing the social media discussion around a certain topic
- Evaluating survey responses
- Determining whether product reviews are positive or negative

Sentiment analysis is not perfect, and as with any automatic analysis of language, you will have errors in your results. It also cannot tell you why a writer is feeling a certain way. However, it can be useful to quickly summarize some qualities of text, especially if you have so much text that a human reader cannot analyze all of it.

There are many ways to do sentiment analysis. Many approaches use the same general idea, however:

i. Create or find a list of words associated with strongly positive or negative sentiment.
ii. Count the number of positive and negative words in the text.
iii. Analyse the mix of positive to negative words. Many positive words and few negative words indicate positive sentiment, while many negative words and few positive words

I ndicates negative sentiment.

4.1.1 Libraries used in R-studio:

- library(twitteR): R Based Twitter Client Description: Provides an interface to the Twitter web API. Version: 1.1.9
- library(httr): Makes interaction with HTTP possible
- library(ROAuth): Used for twitter dev account verificatoin
- library(wordcloud): Helps in word cloud formation, ultimately making sentiment analysis possible
- library(ggmap): Download a static map from server to location tweet geo-location
- library(Rcurl): Allows one to compose general HTTP requests and provides convenient functions to fetch URIs, get & post forms, etc. and process the results returned by the Web server.

4.1.2 R Code

```
 library(twitteR) library(RCurl) library(httr) library(ROAuth)

 sentimentfun = function(tweettext, pos, neg, .progress='non')

 { scores = lapply(tweettext,

 function(singletweet, pos, neg)

 { singletweet = gsub("[[:punct:]]", "", singletweet)

 singletweet = gsub("[[:cntrl:]]", "", singletweet) singletweet = gsub("\\d+", "", singletweet)

 tryTolower = function(x)

 {                              y = NA

           try_error = tryCatch(tolower(x), error=function(e) e) # if not an error

if (!inherits(try_error, "error")) y = tolower(x)

 # result return(y)

 }
```

```
    singletweet = sapply(singletweet, tryTolower) word.list = str_split(singletweet, "\\s+")

    words = unlist(word.list) pos.matches = match(words, pos) neg.matches = match(words, neg) # we

    just want a TRUE/FALSE pos.matches = !is.na(pos.matches) neg.matches = !is.na(neg.matches)


score = sum(pos.matches) - sum(neg.matches) return(score)

    }, pos, neg, .progress=.progress )

    sentiment.df = data.frame(text=tweettext, score=scores) return(sentiment.df)

    }key="sbaHOPi6RjR350FEFnkJYwejd"

    secret="7LSUnaGIYsaJgvuDujzhYYkm1PlB0EDTFLB4wCIS29cKFVckxa" atoken="386648903-

    iTp7sjF595vkcnD98WCCpcq4JDFru3el9eZkswnp"

    asecret="8kIqzWQdy0h9uHGe02qhuszeRzqYdF3IsLznltAm5cZ9o"

    setup_twitter_oauth(key,secret,atoken,asecret)

    pos = readLines("positive-words.txt") neg = readLines("negative-words.txt") library(wordcloud)

    library(ggmap) tweets=searchTwitteR("andaman", n=9000) tweettext=sapply(tweets,function(x)

    x$getText())

    tweettext=lapply(tweettext, function(x) iconv(x, "latin1","ASCII", sub=""))

    tweettext=lapply(tweettext, function(x) gsub("htt.*",' ',x)) tweettext=lapply(tweettext, function(x)

    gsub("#",",x)) tweettext=unlist(tweettext)

    scores = sentimentfun(tweettext, pos, neg, .progress='text') tweetdate=lapply(tweets, function(x)

    x$getCreated())

    tweetdate=sapply(tweetdate,function(x) strftime(x, format="%Y-%m-%d %H:%M:%S",tz =
    "UTC"))

    isretweet=sapply(tweets, function(x) x$getIsRetweet()) retweetcount=sapply(tweets, function(x)

    x$getRetweetCount()) favoritecount=sapply(tweets, function(x) x$getFavoriteCount())

    data=as.data.frame(cbind(ttext=tweettext, date=tweetdate, isretweet=isretweet,
    retweetcount=retweetcount,
```

favoritecount=favoritecount, score = scores$score, product = "complete", state =

"Andaman", country = "India")) data2 = duplicated(data[,1])

data$duplicate = data2 write.csv(data, file= "andaman.csv")

4.2 Tableau for graph Visualization

Data visualization refers to the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization is an accessible way to see and understand trends, outliers, and patterns in data.

Measure names and Measure values are the two fields created in Tableau by default. These fields are created when a data set is imported into Tableau.

Steps:

1. Drag 'Measure Names' into Columns.
2. Drag 'Measure Values' into Rows.

It creates a visual for all measures present in the data set. By default, Tableau creates a bar chart showing all the measure names and their values. The graph type can be changed as there are different pattern available.

To make it more easy to understand filters and different colors for representing the names can be used.

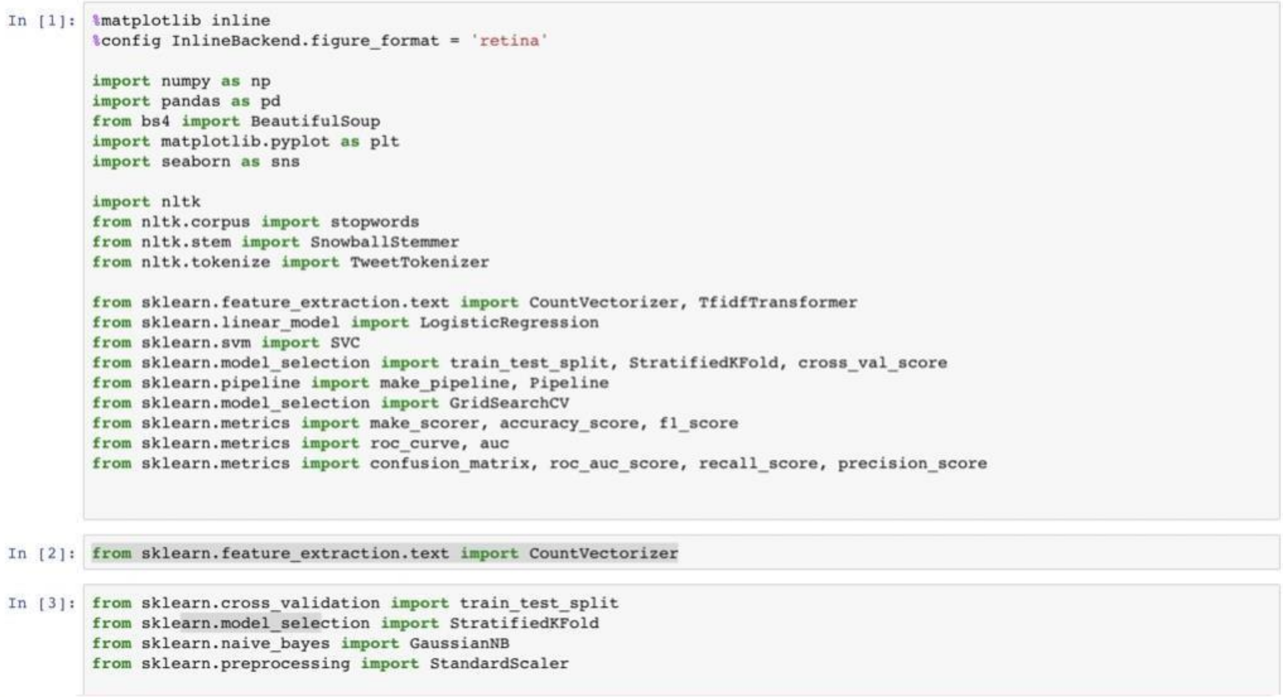
4.3 Python for verification of Sentimental Score using SVM

We take only the tweets we are very confident with. We use the Beautiful Soup library to process html encoding present in some tweets because scrapping.

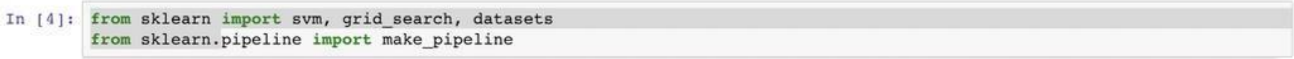
We are going to distinguish two cases: tweets with negative sentiment and tweets with non- negative sentiment

4.3.1 Python Code

Data loading and cleaning

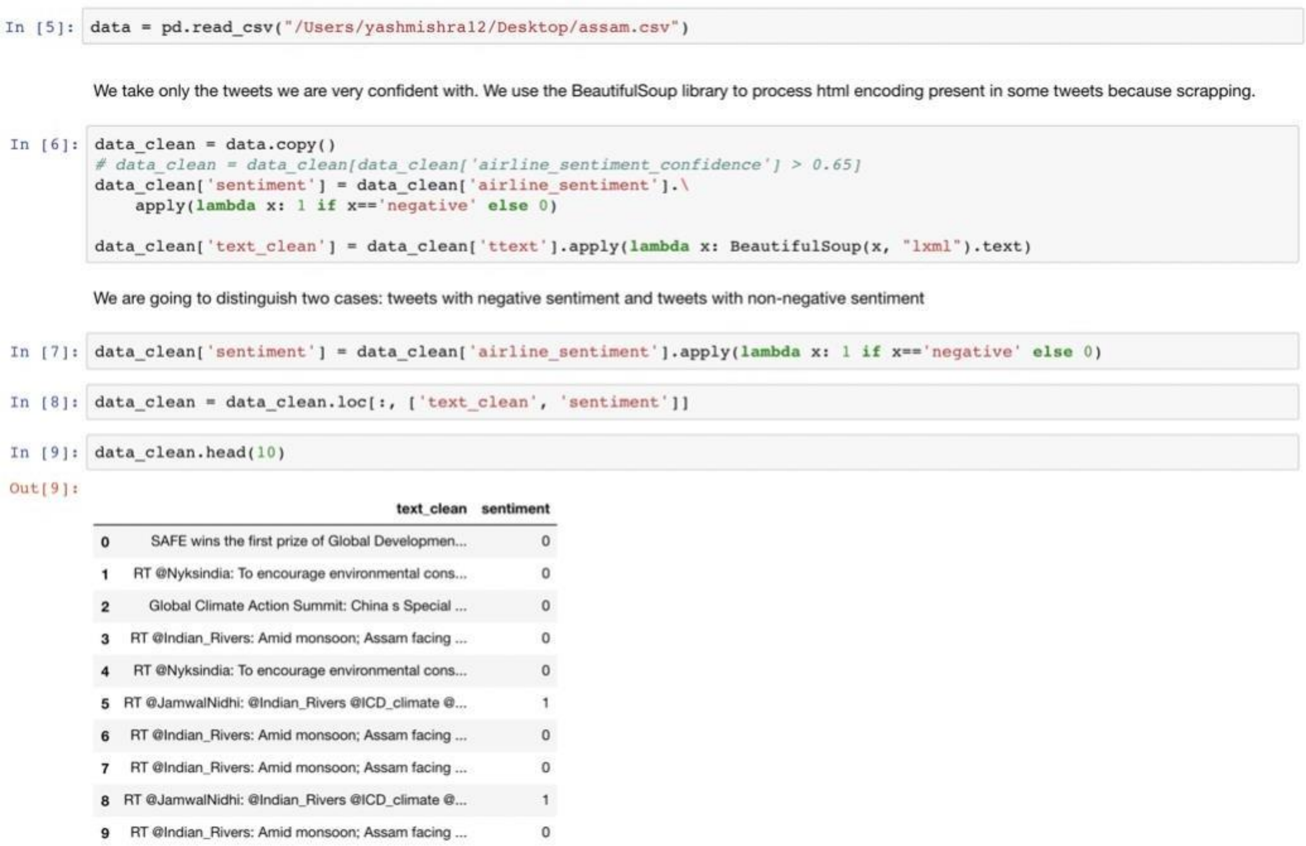


```
In [1]: %matplotlib inline
        %config InlineBackend.figure_format = 'retina'

        import numpy as np
        import pandas as pd
        from bs4 import BeautifulSoup
        import matplotlib.pyplot as plt
        import seaborn as sns

        import nltk
        from nltk.corpus import stopwords
        from nltk.stem import SnowballStemmer
        from nltk.tokenize import TweetTokenizer

        from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
        from sklearn.linear_model import LogisticRegression
        from sklearn.svm import SVC
        from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
        from sklearn.pipeline import make_pipeline, Pipeline
        from sklearn.model_selection import GridSearchCV
        from sklearn.metrics import make_scorer, accuracy_score, f1_score
        from sklearn.metrics import roc_curve, auc
        from sklearn.metrics import confusion_matrix, roc_auc_score, recall_score, precision_score

In [2]: from sklearn.feature_extraction.text import CountVectorizer

In [3]: from sklearn.cross_validation import train_test_split
        from sklearn.model_selection import StratifiedKFold
        from sklearn.naive_bayes import GaussianNB
        from sklearn.preprocessing import StandardScaler
```



```
In [4]: from sklearn import svm, grid_search, datasets
        from sklearn.pipeline import make_pipeline
```

```
In [5]: data = pd.read_csv("/Users,
```



```
In [5]: data = pd.read_csv("/Users/yashmishra12/Desktop/assam.csv")
```

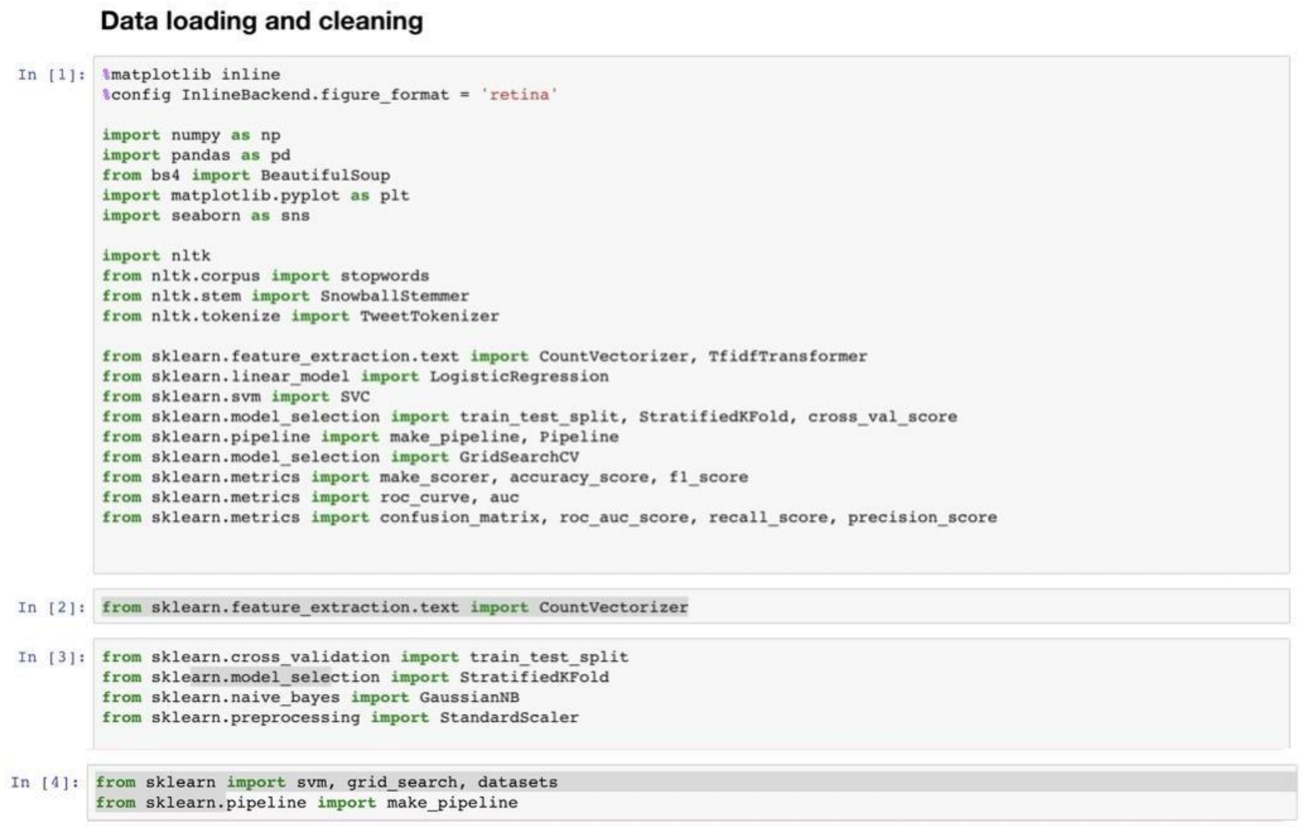
We take only the tweets we are very confident with. We use the BeautifulSoup library to process html encoding present in some tweets because scrapping.

```
In [6]: data_clean = data.copy()
        # data_clean = data_clean[data_clean['airline_sentiment_confidence'] > 0.65]
        data_clean['sentiment'] = data_clean['airline_sentiment'].\
            apply(lambda x: 1 if x=='negative' else 0)

        data_clean['text_clean'] = data_clean['ttext'].apply(lambda x: BeautifulSoup(x, "lxml").text)
```

We are going to distinguish two cases: tweets with negative sentiment and tweets with non-negative sentiment

```
In [7]: data_clean['sentiment'] = data_clean['airline_sentiment'].apply(lambda x: 1 if x=='negative' else 0)

In [8]: data_clean = data_clean.loc[:, ['text_clean', 'sentiment']]

In [9]: data_clean.head(10)

Out[9]:
```

	text_clean	sentiment
0	SAFE wins the first prize of Global Developmen...	0
1	RT @Nyksindia: To encourage environmental cons...	0
2	Global Climate Action Summit: China s Special ...	0
3	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
4	RT @Nyksindia: To encourage environmental cons...	0
5	RT @JamwalNidhi: @Indian_Rivers @ICD_climate @...	1
6	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
7	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
8	RT @JamwalNidhi: @Indian_Rivers @ICD_climate @...	1
9	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0



Data loading and cleaning

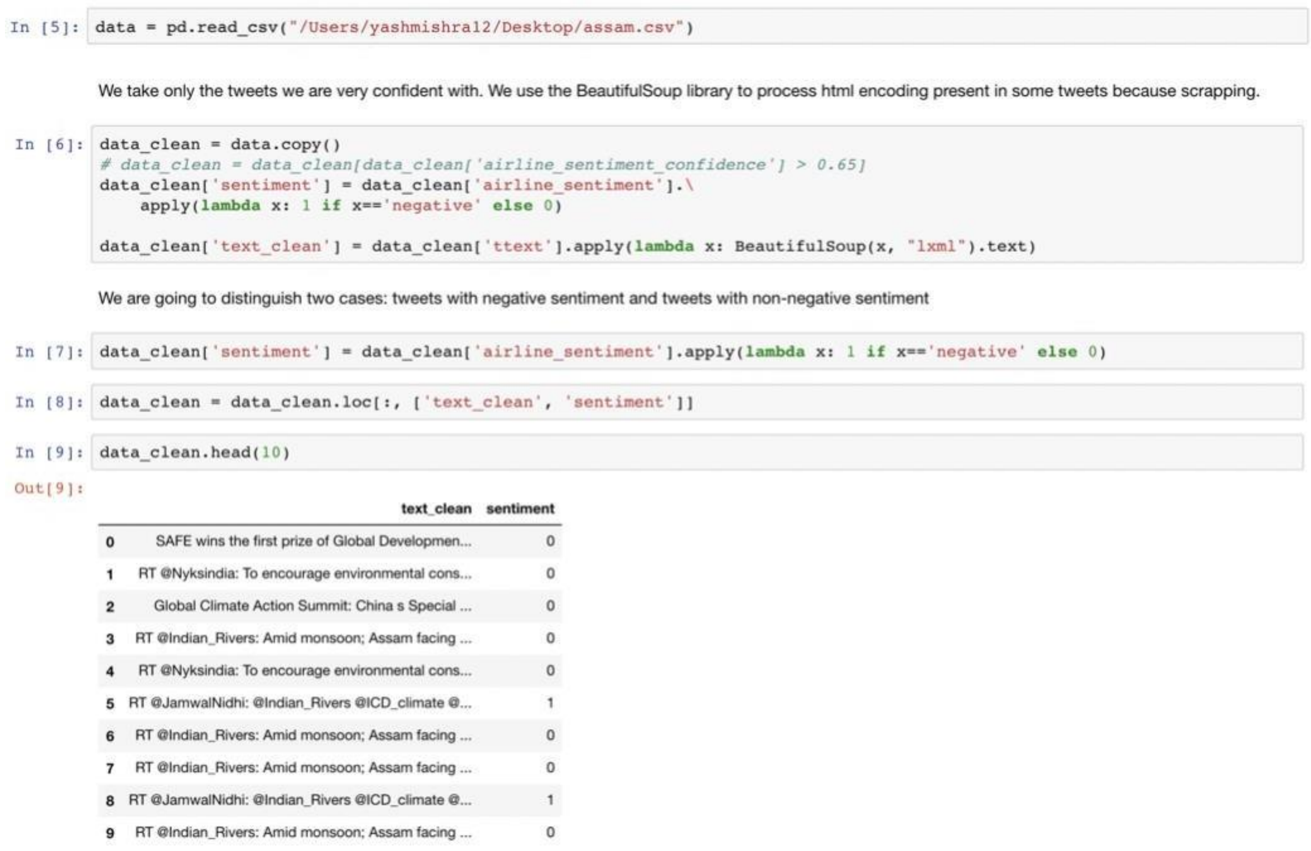
```
In [1]: %matplotlib inline
        %config InlineBackend.figure_format = 'retina'

        import numpy as np
        import pandas as pd
        from bs4 import BeautifulSoup
        import matplotlib.pyplot as plt
        import seaborn as sns

        import nltk
        from nltk.corpus import stopwords
        from nltk.stem import SnowballStemmer
        from nltk.tokenize import TweetTokenizer

        from sklearn.feature_extraction.text import CountVectorizer, TfidfTransformer
        from sklearn.linear_model import LogisticRegression
        from sklearn.svm import SVC
        from sklearn.model_selection import train_test_split, StratifiedKFold, cross_val_score
        from sklearn.pipeline import make_pipeline, Pipeline
        from sklearn.model_selection import GridSearchCV
        from sklearn.metrics import make_scorer, accuracy_score, f1_score
        from sklearn.metrics import roc_curve, auc
        from sklearn.metrics import confusion_matrix, roc_auc_score, recall_score, precision_score
```

```
In [2]: from sklearn.feature_extraction.text import CountVectorizer
```

```
In [3]: from sklearn.cross_validation import train_test_split
        from sklearn.model_selection import StratifiedKFold
        from sklearn.naive_bayes import GaussianNB
        from sklearn.preprocessing import StandardScaler
```

```
In [4]: from sklearn import svm, grid_search, datasets
        from sklearn.pipeline import make_pipeline
```

```
In [5]: data = pd.read_csv("/Users
```



```
In [5]: data = pd.read_csv("/Users/yashmishra12/Desktop/assam.csv")
```

We take only the tweets we are very confident with. We use the BeautifulSoup library to process html encoding present in some tweets because scrapping.

```
In [6]: data_clean = data.copy()
        # data_clean = data_clean[data_clean['airline_sentiment_confidence'] > 0.65]
        data_clean['sentiment'] = data_clean['airline_sentiment'].\
            apply(lambda x: 1 if x=='negative' else 0)

        data_clean['text_clean'] = data_clean['ttext'].apply(lambda x: BeautifulSoup(x, "lxml").text)
```

We are going to distinguish two cases: tweets with negative sentiment and tweets with non-negative sentiment

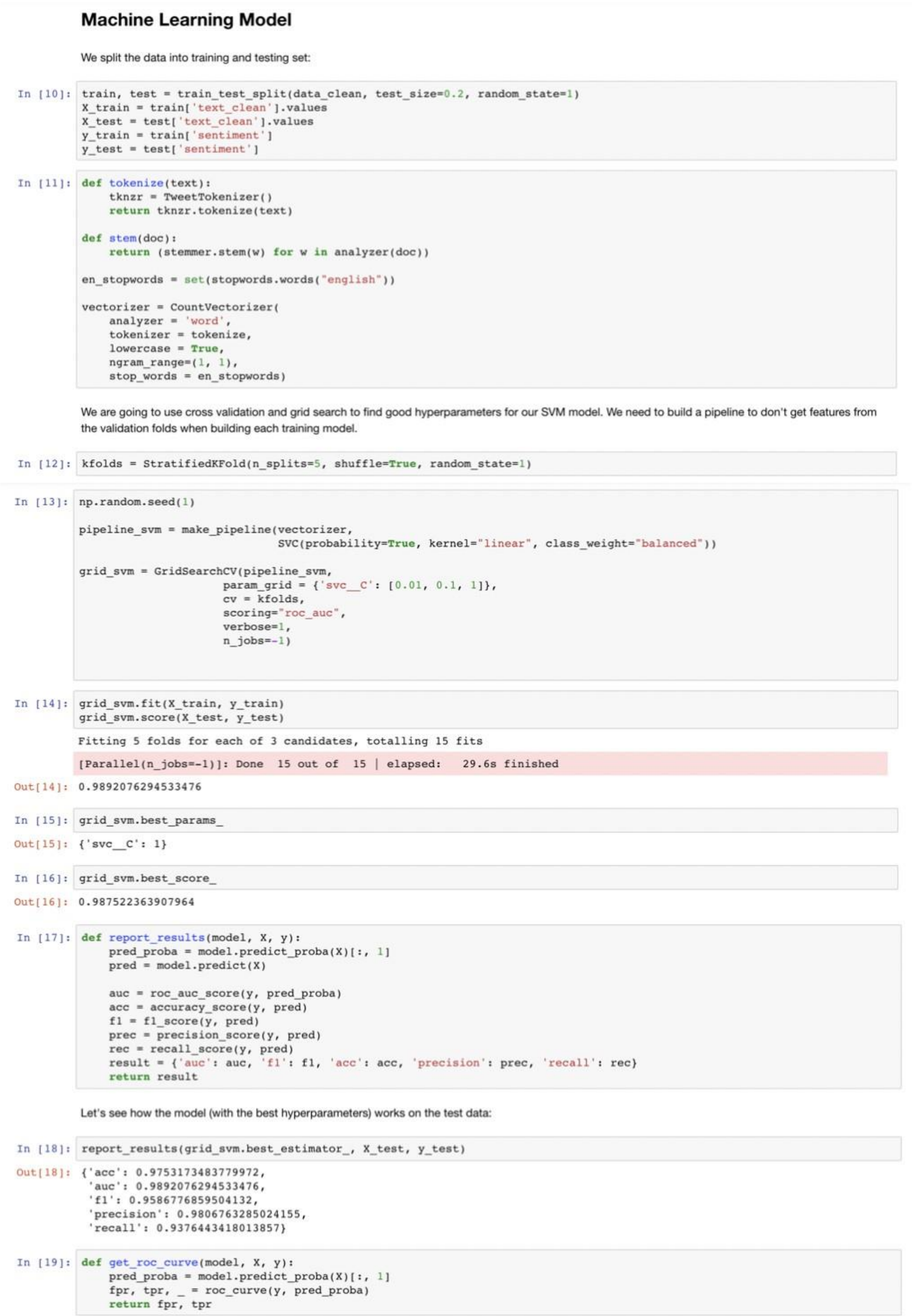
```
In [7]: data_clean['sentiment'] = data_clean['airline_sentiment'].apply(lambda x: 1 if x=='negative' else 0)
```

```
In [8]: data_clean = data_clean.loc[:, ['text_clean', 'sentiment']]
```

```
In [9]: data_clean.head(10)
```

Out[9]:

	text_clean	sentiment
0	SAFE wins the first prize of Global Developmen...	0
1	RT @Nyksindia: To encourage environmental cons...	0
2	Global Climate Action Summit: China s Special ...	0
3	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
4	RT @Nyksindia: To encourage environmental cons...	0
5	RT @JamwalNidhi: @Indian_Rivers @ICD_climate @...	1
6	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
7	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0
8	RT @JamwalNidhi: @Indian_Rivers @ICD_climate @...	1
9	RT @Indian_Rivers: Amid monsoon; Assam facing ...	0



Machine Learning Model

We split the data into training and testing set:

```
In [10]: train, test = train_test_split(data_clean, test_size=0.2, random_state=1)
         X_train = train['text_clean'].values
         X_test = test['text_clean'].values
         y_train = train['sentiment']
         y_test = test['sentiment']
```

```
In [11]: def tokenize(text):
             tknzr = TweetTokenizer()
             return tknzr.tokenize(text)

         def stem(doc):
             return (stemmer.stem(w) for w in analyzer(doc))

         en_stopwords = set(stopwords.words("english"))

         vectorizer = CountVectorizer(
             analyzer = 'word',
             tokenizer = tokenize,
             lowercase = True,
             ngram_range=(1, 1),
             stop_words = en_stopwords)
```

We are going to use cross validation and grid search to find good hyperparameters for our SVM model. We need to build a pipeline to don't get features from the validation folds when building each training model.

```
In [12]: kfolds = StratifiedKFold(n_splits=5, shuffle=True, random_state=1)
```

```
In [13]: np.random.seed(1)

         pipeline_svm = make_pipeline(vectorizer,
                                     SVC(probability=True, kernel="linear", class_weight="balanced"))

         grid_svm = GridSearchCV(pipeline_svm,
                             param_grid = {'svc__C': [0.01, 0.1, 1]},
                             cv = kfolds,
                             scoring="roc_auc",
                             verbose=1,
                             n_jobs=-1)
```

```
In [14]: grid_svm.fit(X_train, y_train)
         grid_svm.score(X_test, y_test)

         Fitting 5 folds for each of 3 candidates, totalling 15 fits

         [Parallel(n_jobs=-1)]: Done  15 out of  15 | elapsed:   29.6s finished

Out[14]: 0.9892076294533476
```

```
In [15]: grid_svm.best_params_

Out[15]: {'svc__C': 1}
```

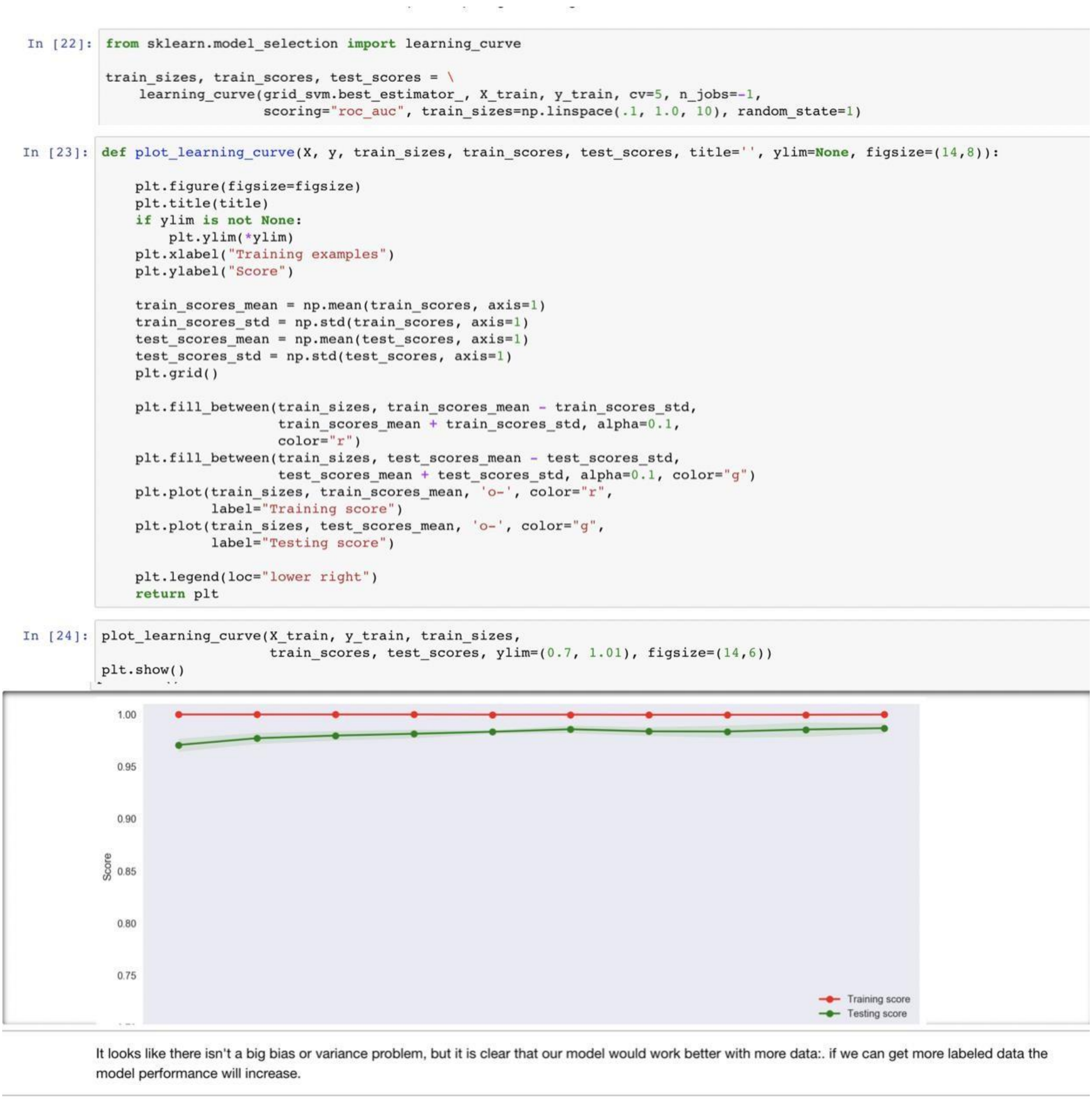
```
In [16]: grid_svm.best_score_

Out[16]: 0.987522363907964
```

```
In [17]: def report_results(model, X, y):
             pred_proba = model.predict_proba(X)[:, 1]
             pred = model.predict(X)

             auc = roc_auc_score(y, pred_proba)
             acc = accuracy_score(y, pred)
             f1 = f1_score(y, pred)
             prec = precision_score(y, pred)
             rec = recall_score(y, pred)
             result = {'auc': auc, 'f1': f1, 'acc': acc, 'precision': prec, 'recall': rec}
             return result
```

Let's see how the model (with the best hyperparameters) works on the test data:

```
In [18]: report_results(grid_svm.best_estimator_, X_test, y_test)

Out[18]: {'acc': 0.9753173483779972,
          'auc': 0.9892076294533476,
          'f1': 0.9586776859504132,
          'precision': 0.9806763285024155,
          'recall': 0.9376443418013857}
```

```
In [19]: def get_roc_curve(model, X, y):
             pred_proba = model.predict_proba(X)[:, 1]
             fpr, tpr, _ = roc_curve(y, pred_proba)
             return fpr, tpr
```



```
In [22]: from sklearn.model_selection import learning_curve

         train_sizes, train_scores, test_scores = \
             learning_curve(grid_svm.best_estimator_, X_train, y_train, cv=5, n_jobs=-1,
                            scoring="roc_auc", train_sizes=np.linspace(.1, 1.0, 10), random_state=1)
```

```
In [23]: def plot_learning_curve(X, y, train_sizes, train_scores, test_scores, title='', ylim=None, figsize=(14,8)):

             plt.figure(figsize=figsize)
             plt.title(title)
             if ylim is not None:
                 plt.ylim(*ylim)
             plt.xlabel("Training examples")
             plt.ylabel("Score")

             train_scores_mean = np.mean(train_scores, axis=1)
             train_scores_std = np.std(train_scores, axis=1)
             test_scores_mean = np.mean(test_scores, axis=1)
             test_scores_std = np.std(test_scores, axis=1)
             plt.grid()

             plt.fill_between(train_sizes, train_scores_mean - train_scores_std,
                              train_scores_mean + train_scores_std, alpha=0.1,
                              color="r")
             plt.fill_between(train_sizes, test_scores_mean - test_scores_std,
                              test_scores_mean + test_scores_std, alpha=0.1, color="g")
             plt.plot(train_sizes, train_scores_mean, 'o-', color="r",
                      label="Training score")
             plt.plot(train_sizes, test_scores_mean, 'o-', color="g",
                      label="Testing score")

             plt.legend(loc="lower right")
             return plt
```

```
In [24]: plot_learning_curve(X_train, y_train, train_sizes,
                             train_scores, test_scores, ylim=(0.7, 1.01), figsize=(14,6))
         plt.show()
```

It looks like there isn't a big bias or variance problem, but it is clear that our model would work better with more data:. if we can get more labeled data the model performance will increase.

Figure 4.3.1

4.4HTML, CSS and JavaScript for Front-end

HTML, CSS and JavaScript were used for dynamism of the website and make it interactive for the users.

Figure 4.4.1

5. Categories

The Ease of Living Index is structured according to 4 pillars- Institutional, Social, Economic and Physical
that represent the broad conceptual elements that define ease of living.
Further these pillars are classified into categories. In our project we some of the categories to decide about the living standard of the people.

5.1 Education

Education is one of the most important aspects of human development and therefore, the degree of access and quality of education is critical to building inclusive cities. The indicators under this category reflect, both the ease of access to educational institutions for the children and also the quality of education in the same. The category also places emphasis on measuring progress with respect to eliminating gender disparities in access to education.

5.2 Health

Healthy cities lead to happy and productive residents. This category includes indicators that measure the capability and capacity of health care infrastructure and services in cities e.g., number of hospital beds, number of healthcare professionals, response time of medical emergencies. Other indicators reflect the incidence of communicable diseases in cities which is not only a reflection of the health of their residents but is also closely linked to pollution levels in cities and the state of sanitation services.

5.3 Transportation

The presence of safe, convenient, affordable and accessible alternatives to driving in a city has a huge impact on a city's development and hence, ease of living. This category includes indicators that reflect on how cities encourage the use of public transport and non-motorized transport, by assessing the existing infrastructure in a city on the basis of availability and safety. Measures taken to improve facilities for pedestrians are also assessed. Inclusiveness of public infrastructure is examined by the extent to which new and redeveloped government buildings, malls, public toilets, footpaths, subways and foot-bridges are built as per universal design principles.

5.4 Culture

This category captures the degree to which a city embraces and maintains its cultural and natural heritage, and promotes sustainable tourism. It is a strong indicator of the vibrancy of a city, which has a bearing on the quality of urban life. It is also a reflection of a city's performance in the upkeep of the business environment for tourism (hotel infrastructure) and the availability of opportunities to explore local identity and culture (restoration of historical and ecological sites).

5.5 Crime

Safety and security have a tremendous impact on the ease of living in a city as people highly value feeling safe inside and outside of their homes. The level of safety in a city can be captured quantitatively by examining at the number of crimes recorded in the city, especially against vulnerable groups (women, children, and elderly people). The number of streets and public places in a city that are covered by surveillance systems are a way to assess the efforts of the city to prevent all forms of violence.

5.6 Energy

Availability of good quality (low voltage fluctuations) and reliable power is a basic necessity for the industries and services in a city to function well. This category includes indicators related to the number of electrical connections and power interruptions as a reflection of the quality of power supply. Indicators related to the percentage of energy derived from non- conventional sources, and energy consumption by other services such as water supply, sewerage, and street lighting aim to track the progress of a city in terms of sustainably managing its natural resources and increasing the use of renewable sources of energy.

5.7 Healthcare Facilities

Healthcare facilities are important at any stage in life. Easy access to good healthcare can increase your quality of life exponentially, so be on the lookout for towns and cities with good hospitals and medical schools. Often, there will be a correlation between cities and the quality of the healthcare.

5.8 Food

If you're a foodie, you may want to try to find a place to live near the ocean or near a metropolitan city center. Grocery store fare, while plentiful, doesn't replace the quality of fresh food from the ocean or fresh produce from the farmers' market. If eating locally and sustainably is important to you, consider whether you can pursue this lifestyle in your new home.

5.9 Employment

Employment opportunities vary from state to state and city to city, so spend some time researching the job markets in different areas of the country. Start by analysing quality employment opportunities within your industry, then determine where the highest concentration of these jobs is located.

5.10 Estate

Since buying a home is the single largest investment you will probably ever make, you need to seriously consider this factor. With real estate in a constant state of flux, it's important to research current home prices, the length of time homes is for sale, the resale values of homes, and probable long-term value estimates.

6. Software level Design

6.1 High Level Design

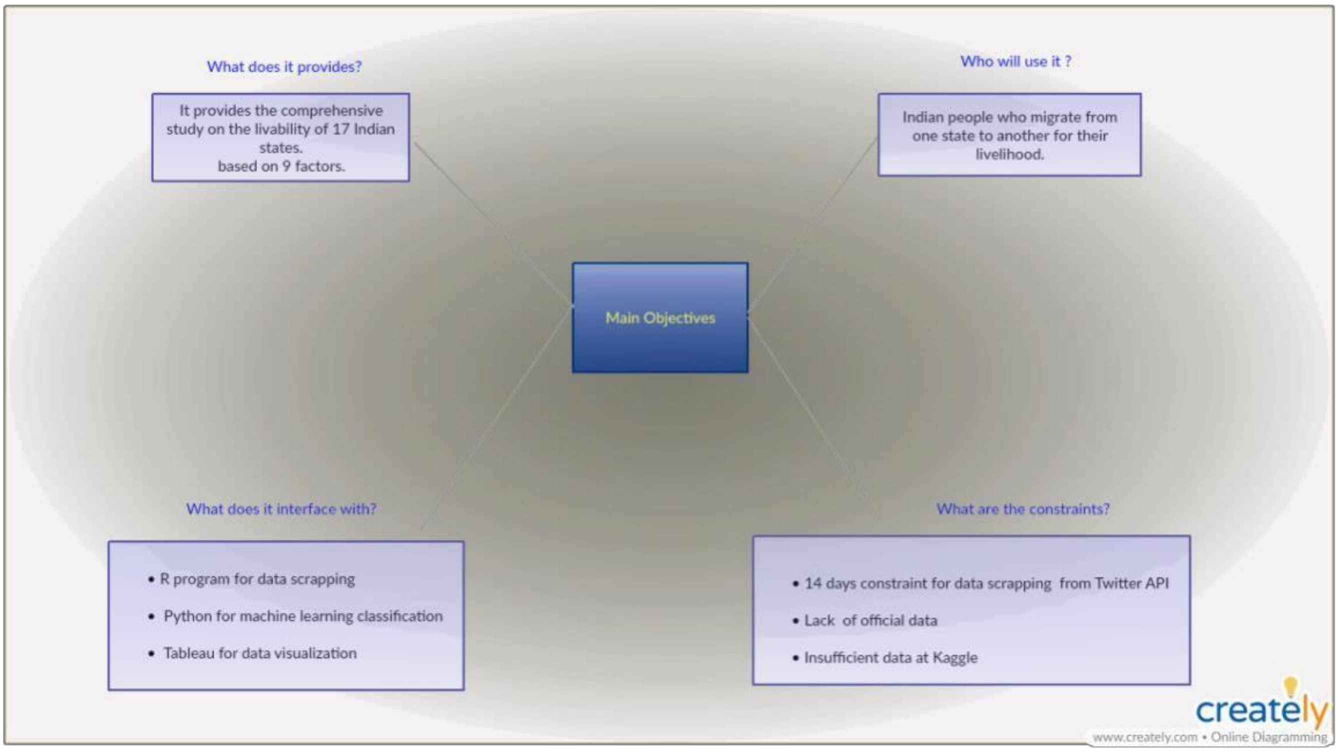

What does it provides?
It provides the comprehensive study on the livability of 17 Indian states. based on 9 factors.
Who will use it ?
Indian people who migrate from one state to another for their livelihood.
Main Objectives
What does it interface with?
• R program for data scrapping
• Python for machine learning classification
• Tableau for data visualization
What are the constraints?
• 14 days constraint for data scrapping from Twitter API
• Lack of official data
• Insufficient data at Kaggle
creately
www.creately.com • Online Diagramming

Figure 6.1.1

6.2 Low level design

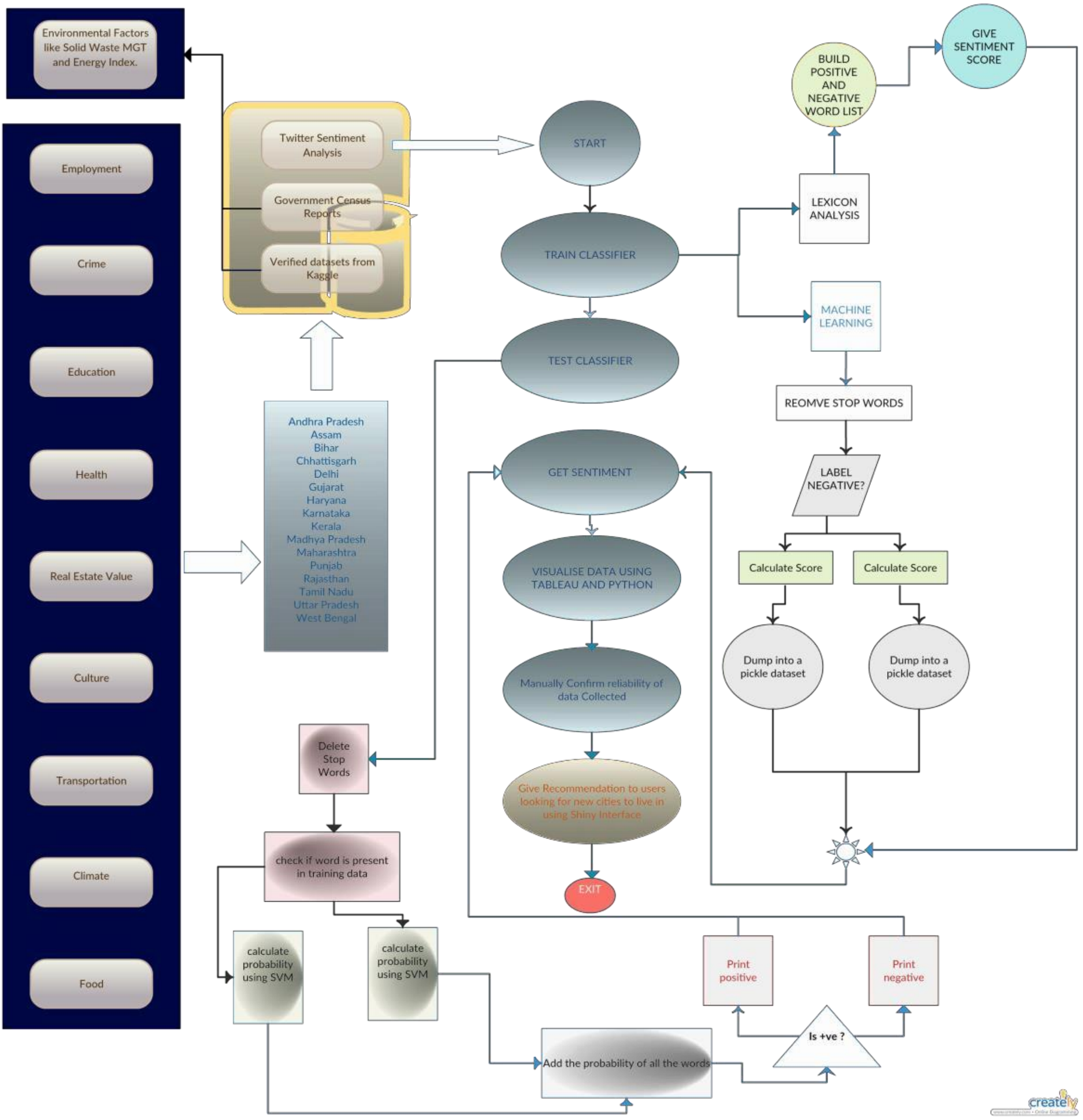

Environmental Factors like Solid Waste MGT and Energy Index.
Employment
Crime
Education
Health
Real Estate Value
Culture
Transportation
Climate
Food
Twitter Sentiment Analysis
Government Census Reports
Verified datasets from Kaggle
Andhra Pradesh
Assam
Bihar
Chhattisgarh
Delhi
Gujarat
Haryana
Karnataka
Kerala
Madhya Pradesh
Maharashtra
Punjab
Rajasthan
Tamil Nadu
Uttar Pradesh
West Bengal
START
TRAIN CLASSIFIER
TEST CLASSIFIER
GET SENTIMENT
VISUALISE DATA USING TABLEAU AND PYTHON
Manually Confirm reliability of data Collected
Give Recommendation to users looking for new cities to live in using Shiny Interface
EXIT
BUILD POSITIVE AND NEGATIVE WORD LIST
GIVE SENTIMENT SCORE
LEXICON ANALYSIS
MACHINE LEARNING
REOMVE STOP WORDS
LABEL NEGATIVE?
Calculate Score
Calculate Score
Dump into a pickle dataset
Dump into a pickle dataset
Delete Stop Words
check if word is present in training data
calculate probability using SVM
calculate probability using SVM
Add the probability of all the words
Is +ve ?
Print positive
Print negative

Figure 6.2.1

7. Interface

7.1 Sentimental Analysis

Figure 7.1.1

7.1.1 Twitter Score

Sentimental Analysis score was collected from Twitter. Twitter allows us to scrape tweets as old as two weeks: 14 days. We also used the dictionary of positive and negative words created by the following scientists: Minqing Hu and Bing Liu and merged their dictionary with other set of scientists named: Bing Liu, Minqing Hu and Junsheng Cheng

7.1.2 Official Score

Official Score of each category for all the states were collected through Official Websites of Indian Government and Kaggle Datasets which had cited their references explicitly.

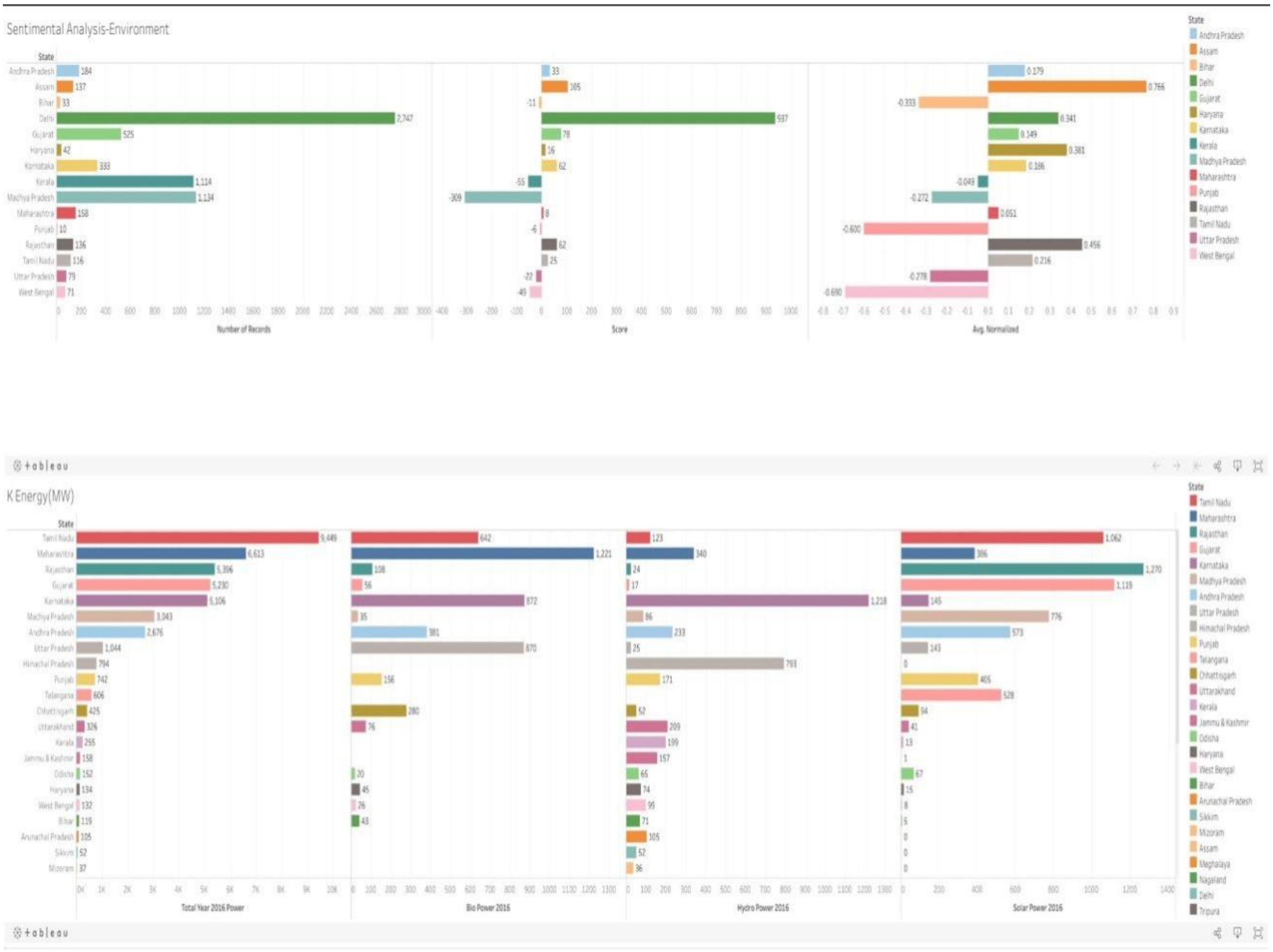
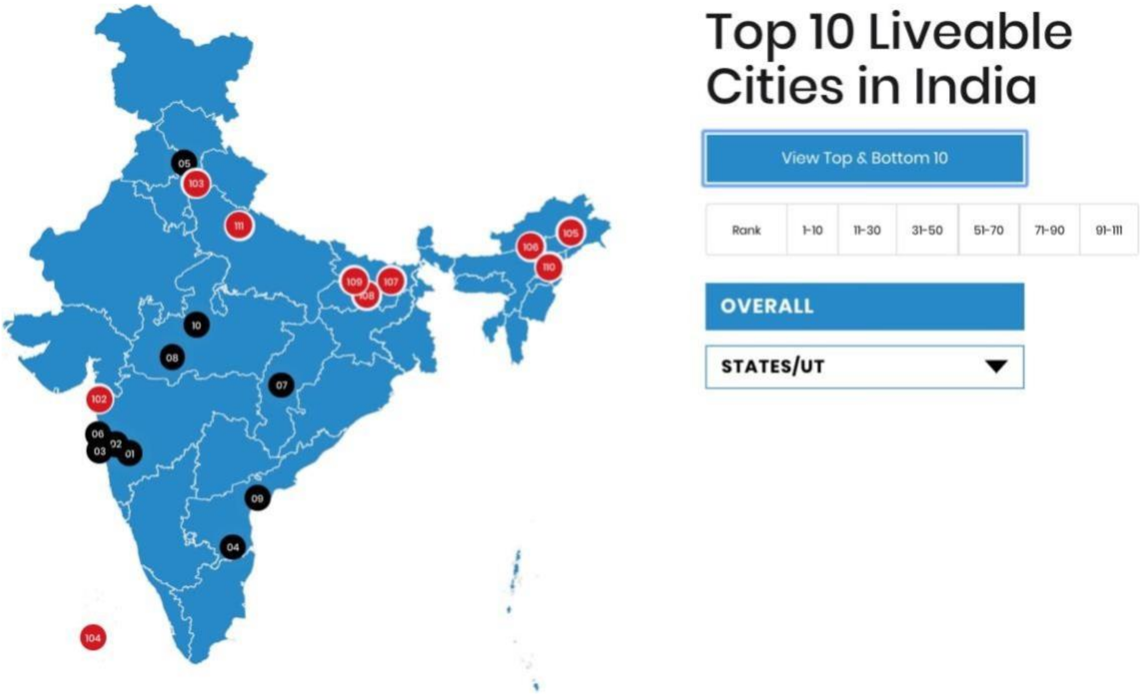


Figure 7.1.2.1

7.2 Index and Ranking

Indexing and Ranking of each state and the cities of each state was referred from the survey done by the ministry of Urban Development in 2017 which ranked all the states except West Bengal and 111 cities on various parameters.



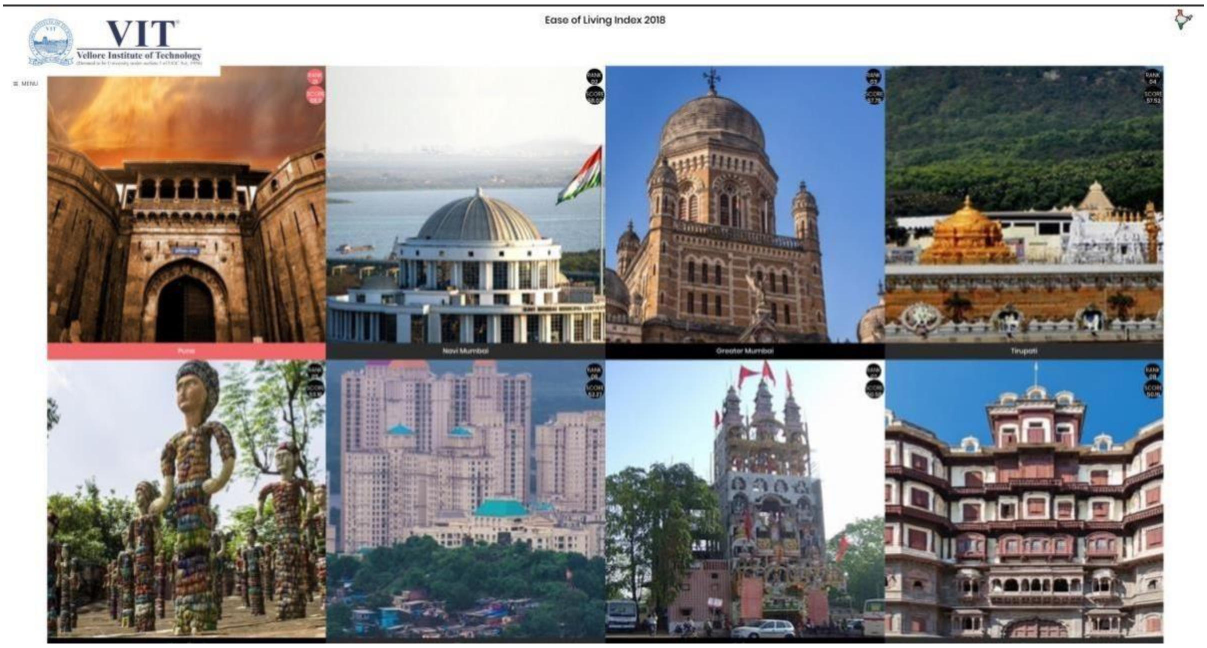
Top 10 Liveable Cities in India

View Top & Bottom 10

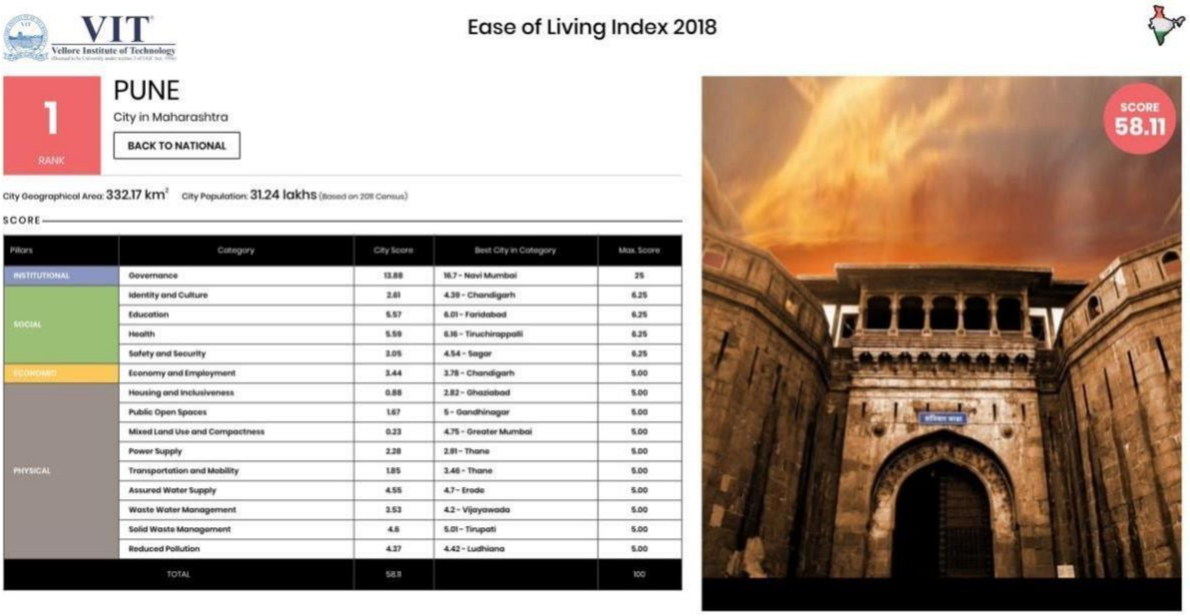
Rank	1-10	11-30	31-50	51-70	71-90	91-111

OVERALL

STATES/UT ▼



Ease of Living Index 2018



VIT
Vellore Institute of Technology

Ease of Living Index 2018

1
RANK

PUNE

City in Maharashtra

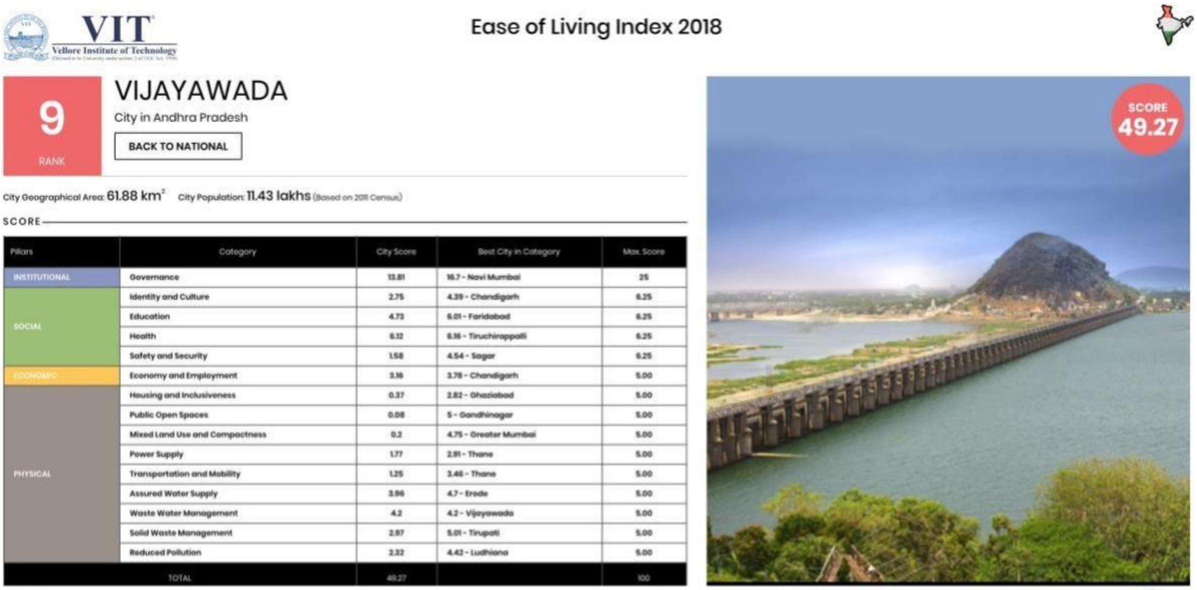
BACK TO NATIONAL

City Geographical Area: 332.17 km^2 City Population: 31.24 lakhs (Based on 2011 Census)

SCORE

Pillars	Category	City Score	Best City in Category	Max. Score
INSTITUTIONAL	Governance	13.88	16.7 - Navi Mumbai	25
SOCIAL	Identity and Culture	2.61	4.39 - Chandigarh	6.25
	Education	5.57	6.01 - Faridabad	6.25
	Health	5.59	6.16 - Tiruchirappalli	6.25
	Safety and Security	3.05	4.54 - Sagar	6.25
ECONOMIC	Economy and Employment	3.44	3.78 - Chandigarh	5.00
PHYSICAL	Housing and Inclusiveness	0.88	2.82 - Ghaziabad	5.00
	Public Open Spaces	1.67	5 - Gandhinagar	5.00
	Mixed Land Use and Compactness	0.23	4.75 - Greater Mumbai	5.00
	Power Supply	2.28	2.91 - Thane	5.00
	Transportation and Mobility	1.85	3.46 - Thane	5.00
	Assured Water Supply	4.55	4.7 - Erode	5.00
	Waste Water Management	3.53	4.2 - Vijayawada	5.00
	Solid Waste Management	4.6	5.01 - Tirupati	5.00
	Reduced Pollution	4.37	4.42 - Ludhiana	5.00
TOTAL		58.11		100

SCORE
58.11



VIT
Vellore Institute of Technology

Ease of Living Index 2018

9
RANK

VIJAYAWADA

City in Andhra Pradesh

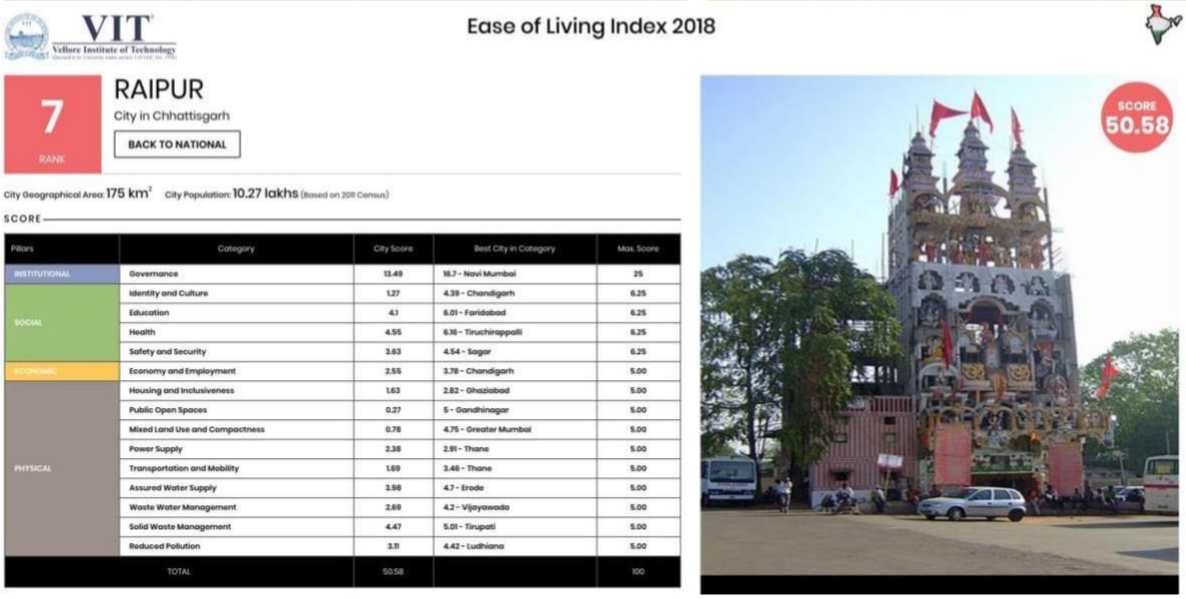
BACK TO NATIONAL

City Geographical Area: 61.88 km^2 City Population: 11.43 lakhs (Based on 2011 Census)

SCORE

Pillars	Category	City Score	Best City in Category	Max. Score
INSTITUTIONAL	Governance	13.81	16.7 - Navi Mumbai	25
SOCIAL	Identity and Culture	2.75	4.39 - Chandigarh	6.25
	Education	4.73	6.01 - Faridabad	6.25
	Health	6.12	6.16 - Tiruchirappalli	6.25
	Safety and Security	1.58	4.54 - Sagar	6.25
ECONOMIC	Economy and Employment	3.16	3.78 - Chandigarh	5.00
PHYSICAL	Housing and Inclusiveness	0.37	2.82 - Ghaziabad	5.00
	Public Open Spaces	0.08	5 - Gandhinagar	5.00
	Mixed Land Use and Compactness	0.2	4.75 - Greater Mumbai	5.00
	Power Supply	1.77	2.91 - Thane	5.00
	Transportation and Mobility	1.25	3.46 - Thane	5.00
	Assured Water Supply	3.96	4.7 - Erode	5.00
	Waste Water Management	4.2	4.2 - Vijayawada	5.00
	Solid Waste Management	2.87	5.01 - Tirupati	5.00
	Reduced Pollution	2.32	4.42 - Ludhiana	5.00
TOTAL		49.27		100

SCORE
49.27



VIT
Vellore Institute of Technology

Ease of Living Index 2018

7
RANK

RAIPUR

City in Chhattisgarh

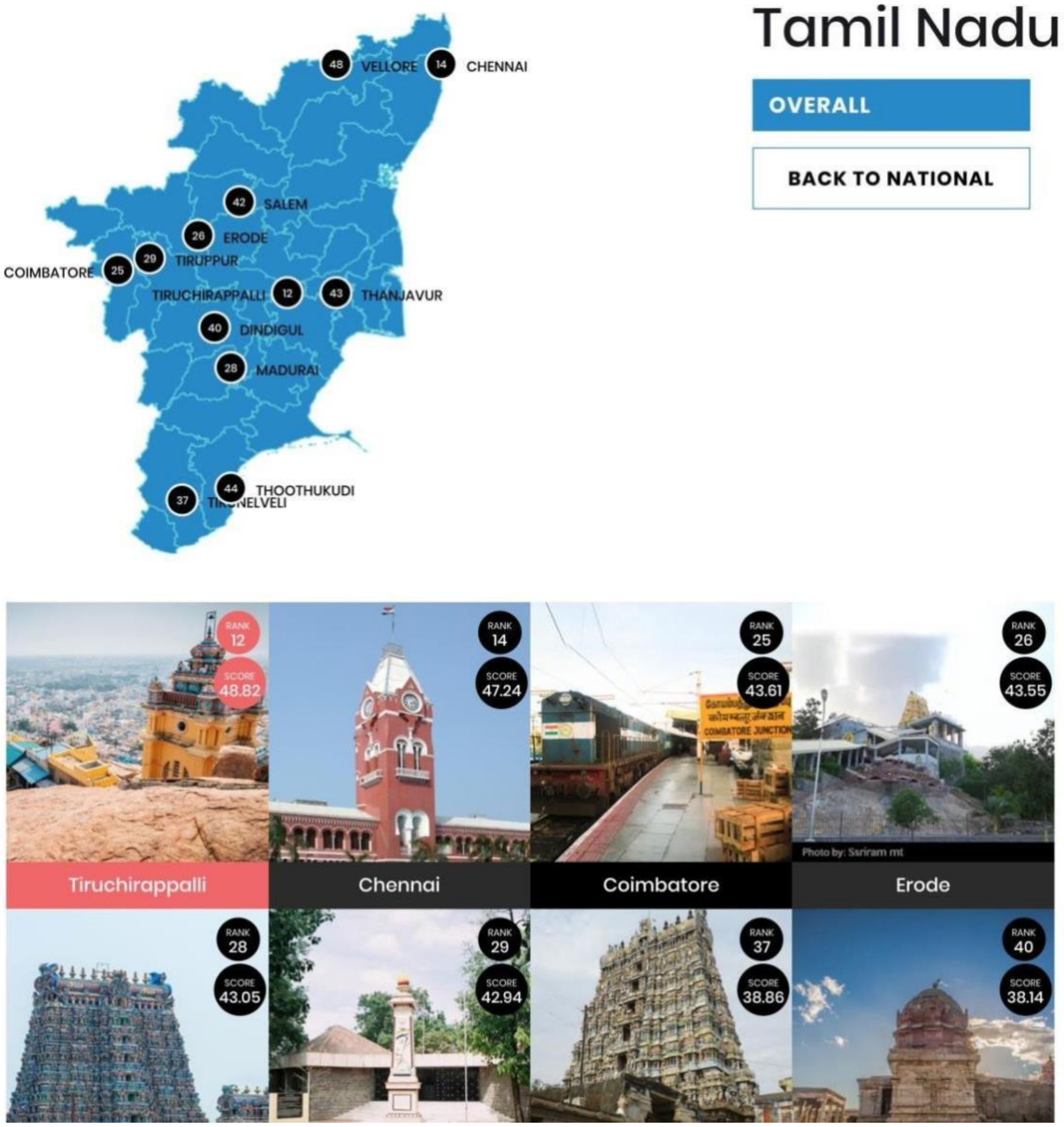
BACK TO NATIONAL

City Geographical Area: 175 km^2 City Population: 10.27 lakhs (Based on 2011 Census)

SCORE

Pillars	Category	City Score	Best City in Category	Max. Score
INSTITUTIONAL	Governance	13.49	16.7 - Navi Mumbai	25
SOCIAL	Identity and Culture	1.27	4.39 - Chandigarh	6.25
	Education	4.1	6.01 - Faridabad	6.25
	Health	4.55	6.16 - Tiruchirappalli	6.25
	Safety and Security	3.63	4.54 - Sagar	6.25
ECONOMIC	Economy and Employment	2.55	3.78 - Chandigarh	5.00
PHYSICAL	Housing and Inclusiveness	1.63	2.82 - Ghaziabad	5.00
	Public Open Spaces	0.27	5 - Gandhinagar	5.00
	Mixed Land Use and Compactness	0.78	4.75 - Greater Mumbai	5.00
	Power Supply	2.38	2.91 - Thane	5.00
	Transportation and Mobility	1.69	3.46 - Thane	5.00
	Assured Water Supply	3.98	4.7 - Erode	5.00
	Waste Water Management	2.69	4.2 - Vijayawada	5.00
	Solid Waste Management	4.47	5.01 - Tirupati	5.00
	Reduced Pollution	3.11	4.42 - Ludhiana	5.00
TOTAL		50.58		100

SCORE
50.58



Support Vector Machine

Figure 7.2.1

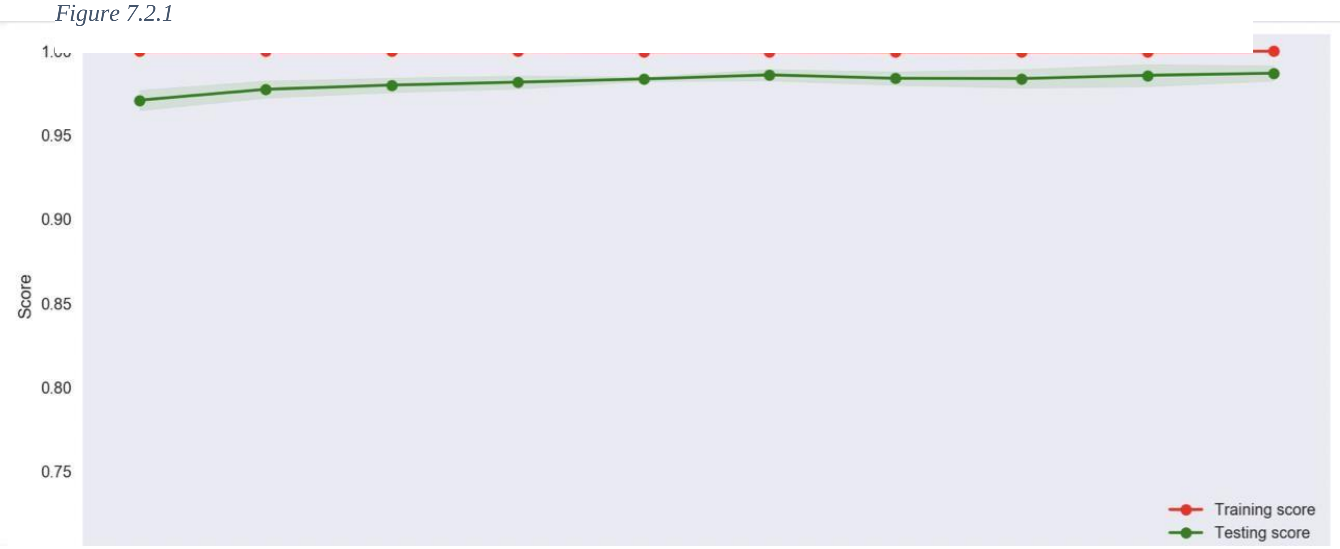


Figure 7.2.2

7.3 Graph Visualisation

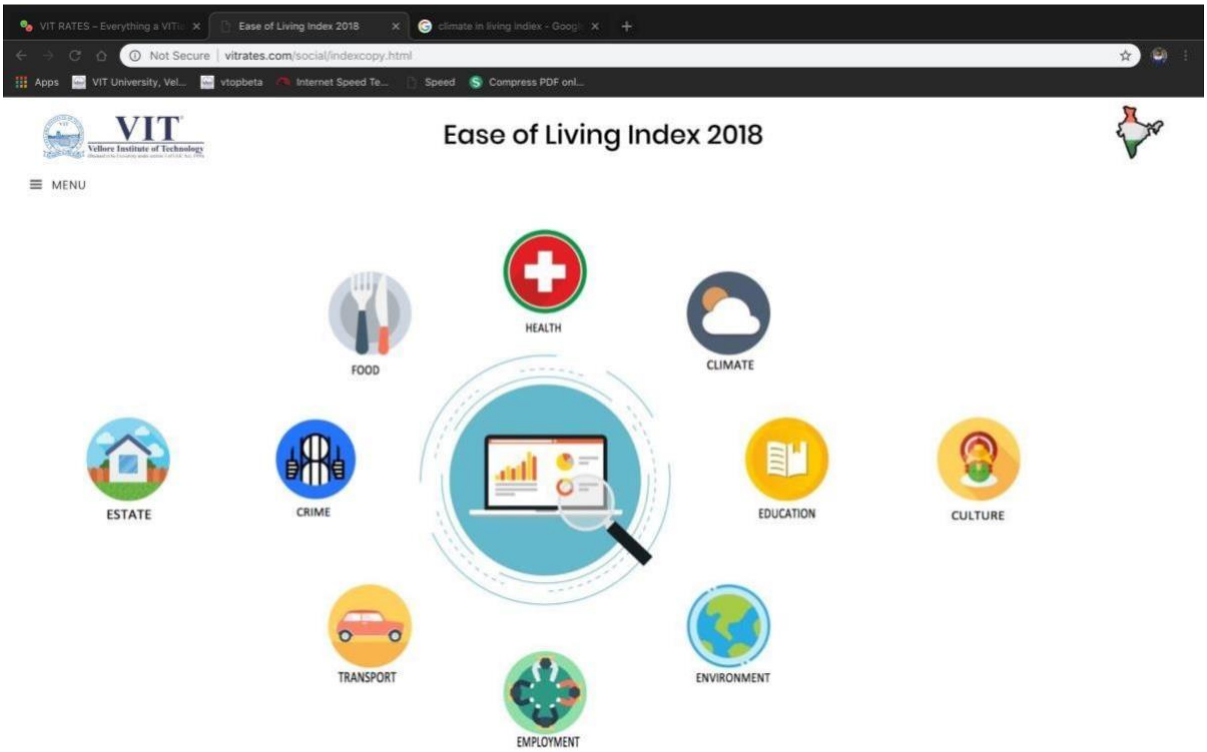

VIT RATES – Everything a VIT
Ease of Living Index 2018
climate in living index - Googl
Not Secure
vitrates.com/social/indexcopy.html
Apps
VIT University, Vel...
vtopbeta
Internet Speed Te...
Speed
Compress PDF onl...
VIT
Vellore Institute of Technology
MENU
Ease of Living Index 2018
HEALTH
FOOD
CLIMATE
ESTATE
CRIME
EDUCATION
CULTURE
TRANSPORT
ENVIRONMENT
EMPLOYMENT

Figure 7.3.1

All the categories of different states are integrated on the website page to analyse. Visualisation is done with the help of the Tableau software. Sentimental data obtained from the twitter and the official data obtained from the Kaggle and official records are compared and visualised.

Crime Graph

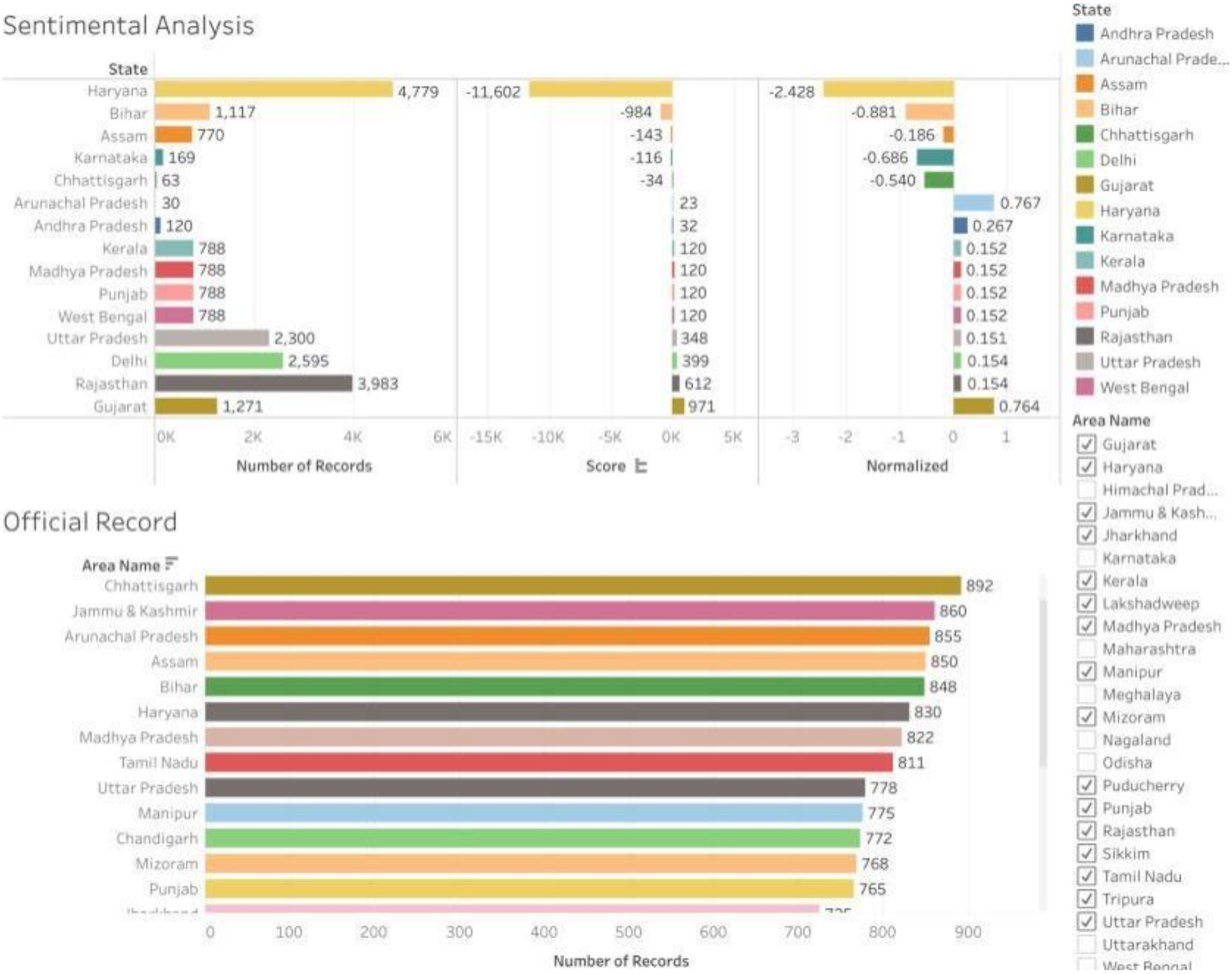


Figure 7.3.2

These are the graph of the Crime data of each state where the first graph is of Sentiment data and the second is of Official data.

In the sentiment graph of crime, Haryana has highest negative score which means the number of crimes is more in Haryana. Higher the negative score, higher the crime rate. In the official record graph, number of records of crime is higher for the state Chhattisgarh, Jammu & Kashmir, Bihar, etc. The result of the sentiment analysis is not up to the marks but matches the official record of some of the states. As the sentimental data is of only a short period of 1 month only. One reason could be also that number of tweet user is less active in the particular state.

The crime rates depend on the number of population and area of the state. The number of crime record of each state is normalized by the number of populations of that state.

Then also Haryana is on top for the number of crimes.

Similarly, other graphs can be checked for reference and their sentimental score has been visualized by Tableau 2018b student version.

8. Conclusion

Hence we can safely conclude that public opinion is highly volatile and cannot be used to blindly be considered for crucial and key inferences like indexing the liveability of a place. Sentiments or the score collected is also heavily dependent on various other factors like the dictionary of positive and negative words if we are choosing for lexical analysis and the tweets or the statements collected after cleaning.

Cleaning and post processing of the tweet also leads to heavy change in the context of the tweet. Moreover, one cannot ignore the fact that India is a developing nation and its citizens are still not heavy users or accustomed to sharing their views and opinions regularly on social media to be recorded and used to statistical inference. States like Uttar Pradesh, Bihar and North-Eastern states like Tripura, Sikkim etc along with union territories like Lakshadweep are prime examples which show case the lack of users in the front of social media for scientific data collection.

Therefore, we conclude from our study that official data which involves onsite study and inspection is the best procedure to judge a state's condition and its limit to comfort its citizens and social media, though a boon in many cases, fails to make for a reliable and concrete source of inference in this scenario.

9. References

[1]. Dunbar, R. I. M., Arnaboldi, V., Conti, M., & Passarella, A. (2015). The structure of online social networks mirrors those in the offline world. *Social Networks*, *43*, 39–47. https://doi.org/10.1016/j.socnet.2015.04.005

[2]. Patil, H. P., & Atique, M. (2015). Sentiment Analysis for Social Media: A Survey. *2015 2nd International Conference on Information Science and Security (ICISS)*, 1–4.

https://doi.org/10.1109/ICISSEC.2015.7371033

[3]. Chen, X., Vorvoreanu, M., & Madhavan, K. P. C. (2014). Mining social media data for understanding students' learning experiences. *IEEE Transactions on Learning Technologies*, *7*(3), 246–259. https://doi.org/10.1109/TLT.2013.2296520

[4]. Onnom, W., Tripathi, N., Nitivattananon, V., & Ninsawat, S. (2018). Development of a Liveable City Index (Lci) Using Multi Criteria Geospatial Modelling for Medium Class Cities in Developing Countries. *Sustainability (Switzerland)*, *10*(2). https://doi.org/10.3390/su10020520

[5]. Ediger, D., Jiang, K., Riedy, J., Bader, D. A., Corley, C., Farber, R., & Reynolds, W. N. (2010). Massive social network analysis: Mining twitter for social good. *Proceedings of the International Conference on Parallel Processing*, 583–593. https://doi.org/10.1109/ICPP.2010.66

[6]. Ji X Chun S Geller J Monitoring public health concerns using twitter sentiment classifications Proceedings - 2013 IEEE International Conference on Healthcare Informatics,
ICHI 2013 https://doi.org/10.1109/ ICHI.2013.47

[7]. Wen, W., & Wei, W. (2016). Social Media as Research Instrument for Urban Planning and Design. *2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, 614–616. https://doi.org/10.1109/ICMTMA.2016.150

[8]. Lineman, M., Do, Y., Kim, J. Y., & Joo, G.J. (2015). Talking about Climate Change and Global Warming. *PLOS ONE*, *10*(9). https://doi.org/10.1371/journal.pone.0138996

[9]. Farooq, A., Joyia, G. J., Uzair, M., & Akram, U. (2018). Detection of influential nodes using social networks analysis based on network metrics. *2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, 1–6. https://doi.org/10.1109/ICOMET.2018.8346372

[10]. Zadeh, L. A., Abbasov, A. M., & Shahbazova, S. N. (2015). Analysis of Twitter hashtags: Fuzzy clustering approach. *2015 Annual Conference of the North American Fuzzy Information Processing Society (NAFIPS) Held Jointly with 2015 5th World Conference on Soft Computing (WConSC)*, 1–6. https://doi.org/10.1109/NAFIPS-WConSC.2015.7284196

[11].Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,

[12]. Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[13] Kaggle hospital count: https://www.kaggle.com/akshatuppal/all-india-health-centres- directory

[14] Transportation-Road: http://www.mospi.gov.in/statistical-year-book-india/2017/190

[15] Unemployment: https://unemploymentinindia.cmie.com/

[16] Cleanliness: https://www.financialexpress.com/india-news/full-list-of-swachh- survekshan 2017-rankings-know-where-your-city-ranks-in-swachh-bharat-survey/653949/

[17] Energy http://mospi.nic.in/statistical-year-book-india/2017/185

[18] Statistical Year Book India 2017
http://mospi.nic.in/statistical-year-book-india/2017

[19] State population data source:
https://fusiontables.google.com/DataSource?dsrcid=645331