A
Dissertation On

"An Impact of Social Media via Twitter Analytics"

Submitted in Partial Fulfilment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Vikesh Kumar Singh
University Roll No. 2K14/SWT/517

Under the Esteemed Guidance of
Dr. Rajni Jindal
HOD, Computer Science & Engineering, DTU

COMPUTER SCIENCE & ENGINEERING DEPARTMENT
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, New Delhi-42

This is to certify that the thesis entitled **"Impact of Social media via Twitter Analytics"** done by me for the Major project for the award of degree of **Master of Technology** Degree in **Software Engineering** in the **Department of Computer Science & Engineering**, Delhi Technological University, New Delhi is an authentic work carried out by me under the guidance of Dr. Rajni Jindal.

Signature:
Student Name
Vikesh Kumar Singh
2K14/SWT/517

Above Statement given by Student is Correct.

Project Guide:
Dr. Rajni Jindal
Head of Department,
Department of Computer Science & Engineering
Delhi Technological University, Delhi

ACKNOWLEDGEMENT

I take this opportunity to express my deep sense of gratitude and respect towards my guide **Dr. Rajni Jindal, Head of Department, Department of Computer Science & Engineering.**

I am very much indebted to her for her generosity, expertise and guidance i have received from her while working on this project. Without her support and timely guidance the completion of the project would have seemed a far-fetched dream. In this respect I find myself lucky to have my guide. She have guided not only with the subject matter, but also taught the proper style and techniques of documentation and presentation. I would also like to take this opportunity to present my sincere regards **Dr. Ruchika Malhotra**, Assistant Professor, DTU for extending their support and valuable Guidance.

Besides my guide, I would like to thank entire teaching and non-teaching staff in the Department of Computer Engineering, DTU for all their help during my tenure at DTU. Kudos to all my friends at DTU for thought provoking discussion and making stay very pleasant.

Vikesh Kumar Singh
M.Tech, Software Engineering
2K14/SWT/517

ABSTRACT

In today's world micro blogging has become emerging connection medium for Internet users [1]. Many users share their views on different semblance in famous websites such as Facebook, Qzone, LinkedIn, Twitter and Tumblr. With increase in the user on social networking sites, many big giants and media organization are trying to achieve different ways to get these social media information so that they can know what people think about their quality, product and companies.

Many firms, Big Organization, Political parties as always keen in knowing if the people will sustain with their event, program or not. Many social NGO's and Organization can ask people's views on current topics, challenge for open debate etc. All such kind of information's can be collected from such plenty of micro blogging websites. Here we represent a function which performs which will do classification based on tweets/retweets and calculate the impact of specific keyword/#Hashtag in Twitter.

Currently twitter network is dazzled with huge no of tweets tweeted by its users. For an productive categorization and searing of tweets, user need to use suitable meaningful sentence and hashtags in their tweets. Twitter has a huge number of users which may varies from Politian's, Celebrities, Actors, company representatives, an even country president uses twitter to express their views on social platform. By this ways we can collect the all possible text posts of users from different organization, companies, interest groups and different social groups [2].

In this project I propose generic functions/recommendation method of the tweets of the individuals/popular personalities that tweet which will create an impact on user mind after the tweets. If we have different groups of users and these tweets, we can easily create our methods so that we can find out the top most familiars users and top most familiars tweets from collected data. Hashtags/Keywords are then used to select to select most familiar tweets and user and then we can assign them some ranking values/scores to them.

In future I will explore a more on more types of different categories by after-peak value, before-peak value, and during-peak value popularity. It will be arousing to propose different recommendation methods.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 General Concepts

In today's world micro blogging has become emerging connection medium for Internet users [1]. Many users share their views on different semblance in famous websites such as Facebook, Qzone, LinkedIn, Twitter and Tumblr. With increase in the user on social networking sites, many big giants and media organization are trying to achieve different ways to get these social media information so that they can know what people think about their quality, product and companies.

Many firms, Big Organization, Political parties as always keen in knowing if the people will sustain with their event, program or not. Many social NGO's and Organization can ask people's views on current topics, challenge for open debate etc. All such kind of information's can be collected from such plenty of micro blogging websites. Here we represent a function which performs which will do classification based on tweets/retweets and calculate the impact of specific keyword/#Hashtag in Twitter [2].

Currently twitter network is dazzled with huge no of tweets tweeted by its users. For an productive categorization and searing of tweets, user need to use suitable meaningful sentence and hashtags in their tweets .Twitter has a huge number of users which may varies from Politian's, Celebrities, Actors, company representatives, an even country president uses twitter to express their views on social platform. By this ways we can collect the all possible text posts of users from different organization, companies, interest groups and different social groups. However, we knew Analyzing Twitter data is not an easy cup of tea for anyone.

We know twitter has huge number of text posts data and the rate of posts is increasing every by day as lot many people are joining social media these days. Tweets very from person to person and some are small in length, thus very ambiguous in nature.

Tweets very from person to person and may be short in length, so they are very ambivalent in nature. The casual way of thinking, writing, a unique custom of acronymization, orthography, and a usage of different set of elements in hash tags, similarly user mentions can be used by various user which requires different vision to solve any kind of problem.

As social networking sites are tremendously growing, massive number of people wants to express their feelings and opinions are enormously increasing day by day these days. Information collected from these can be benevolent for Companies, governments, politicians, business, organizations and individuals as well. With 600+ tweets each day, Twitter is enormous information hub for all of us. Twitter is mainly a micro-blogging website which is known of its short text message which is so called tweet. Each tweet has only 140 character limit which is unique feature of it [3]. Twitter has 300+ million active users in this websites, so it will be very useful source of data to all of us today. Users mainly discuss political aspects, treading topics of the week and share personal opinion of those topics via their personal tweets.

On Twitter, users always write their tweets as short sentences having upto 140 characters which may be alphabets, numeric, special character etc [3]. A hashtag is a word prefixed by a # symbol and one or more hashtag can be inserted into any tweet. Empirical research indicates that hashtags had been used for multiple occasions and it can be created based on the thinking of the user mind. Many people need hashtags to classify their tweets.

Many people need hashtags to post content related to any special events, occasion, elections, campaign or any disaster etc. Main usage of hashtags is to promote branding of items, things, and events etc. Tweets having hashtags can be easily find by any user and this will help in finding any useful discussion, any trending topic so that can join on the same topic in same platform. Hashtag is not centrally controlled or managed by any organization, any group or user, so it may be bit difficult for lots of users to select suitable hashtags for any tweets.

With the increasing sensation of social media, such as: Twitter, Facebook, Instagram and LinkedIn, with everyone having its special priorities and usage techniques which are infecting social life these days. Networking website called Facebook, where everyone in the network is connected to each other and has relationships in the same network. Inversely happens in Twitter all people in the network does not basically a requite relationship with some others.

1.2. MOTIVATION

In previous years, different type of social information channels, such as micro-blogging and text messaging has been come into light and come in our daily routine life. Although there is never any limit to scope of facts fetch from tweets and messages, these are small and short messages which is used to express people views, feeling, sentiments and opinion and this is how the world is going currently.

Tweets and message are preferably very short, a sentence or headline rather than complete document on any topic. The language which is used in text is very basic, contains misspellings with innovative spelling and punctuation, catchy words, slang, URL's, short abbreviations i.e. RT for "RE-Tweet" and lots of Hashtag used which are types of tagging any person on Twitter messages [3].

Another important facet of social network media data such as Tweets which includes rich structured information of many people related to communication medium [4]. Twitter always maintains information of who all follows whom and re-tweets count and tags inside of any tweets provide which various facts and figures to concerned people/Organizations.

It's the instance and personal nature of social message that makes it's unlike any different form of data, which can show its own provocations of user perspective on consumer etiquette.

Text analytics when applied to social media data or other unstructured data surfaces conversations that reflects and in a more natural manner a consumer's unprompted and unsolicited perception of your company's brand, service or offering.

Surface unprompted insights & details:

- Movie Preferences , Brand references
- Viewing Intentions , Lifestyle references
- Personal Preferences , choice reference

1.3 RELATED WORK

Recommendation systems [1] provides information filtering systems with the help of it we can forecast the proclivity of different users for their different choices, items (like movies, serials, TV shows, songs, books, companies) or social component's that he has to recommended before any tweet to be published.

Two types of recommendation systems are – personalized and non-personalized [1]. A non-personalized recommendation system always has priority for distinct user's preference. Any user can recommend the mast watch top five popular movies of the current year. But this may considered as the personalized preference of the individuals and this list may vary from person to person. Our focus is on personalized recommendation. The two major ways to perform personalized recommendation are collaborative filtering and content- based recommendation.

Collaborative Filtering Approach [1] is based on surmise of the collaborative filtering approach is that if a person A has selected few same things as similar selected by other person B earlier, A is more likely will pick B 's things also rather than the items of a unknown person. In reference to rating recommendation, collaborative filtering is used to know the rating of the target user assigns to an item using the rating on that item assigned by other user who will share the similar rating surmise as the target audience. This filtering is called "User-to-User" collaborative filtering.

Content Based Approach target familiarity between items by comparing their features and characteristics. Recommendation system has made to select user things which are familiar to other user they have used earlier. The content based approach assures to use of things to find out the

Feature Extraction

In micro-blogging world with major insights on Twitter, a bigram model outperforms both unigram and trigram using Multinomial Naïve Bayes classifiers. Although, the inverse was true in case of SVM and MaxEnt classifier [5]. Introduction of a combination of unigram and bigram in feature extraction promised better results in MaxEnt as well as NB classifiers.

Unigram+Bigram

Both unigrams and bigrams are used as features

• In movie-review a decrease observed for Naive Bayes and SVM, but an increase for MaxEnt

• Currently bed found that as compared to unigram features, accuracy improved for Naive Bayes & MaxEnt.

Naive Bayesian Classifier

• Straightforward and frequently used method of supervised learning [6].

• Provides a edible way for dealing with any number of attributes or classes, and is based on probability theory.

• Maximum entropy classifiers are commonly used as alternatives to Naive Bayesian classifier because they do not require statistical independence of the features that serve as predictors.

• Provides around 79% accuracy for tweets

1.4 PROBLEM STATEMENT

Features Considered I plan to make use of following additional features apart from the ones mentioned till now:

• If a tweet contains more than one sentence, I will give more weightage to sentences coming afterwards

• This is due to the likelihood of most tweets to be convincing in nature.

• I am planning to use the hash tags to get idea about tweets of different users.

• Hashtags may be: #IndiaVsPak #IPL2017, #YOGIADITYANATH, #YOGIINUP

• These hashtags should be more structured although they are not complete sentences, so we would need to parse these tweets before processing.

• Hashtags like #feelingsad, #InlovewithSong, #Enjoying, etc. give sufficient info about the polarity of different emotions of tweets [15].

Popular KeyWords

I need to filter out the popular keywords that are frequently used in twitter as a trend.

Popular keywords that are not directly hashtags but used many times in tweets and retweets

Popular keywords in trends are NextPresidentofIndia, UPCM, and IPL2017 etc. which can also be used to categories and helpful in finding the impact of tweets of the person.

• Try to incorporate the effect of modifiers like "very", "too", etc.

• Consider this tweet: "Such a great knock. Team scored this at the loss of just one wicket." Now the ` problem is that it contains one word "great" and the other "loss", and so we would get the overall sentiment as neutral. But it is indeed positive. It is important to capture the idea, as to why it is so. The reason is that they say 'loss of "only" one', meaning at a minimal loss. So, if we capture this notion as well, we will get a pretty increase in accuracy [14].

1.5 SCOPE OF THIS THESIS

In this Project, I have investigated the possibilities of analyzing social media with Machine Learning and Sentiment Analysis, Tweets/Retweets, Hashtag and popular keywords trending on twitter.

Previous research in the field of IoT(Internet of things) which is mainly focused on the synthetic monitoring of the Application/web browser actually running on the real runtime environment.

The purpose of this thesis is to analyze if social media analysis can be used to predict the influence of the tweet of the person. By using the proposed model for extracting the tweets and applying the classification based on the thesis, we can easily calculate the impact of the person's tweets.

2. LITERATURE REVIEW

We have to record and supervise different data values and sets, mostly studies starts by gathering required data values taken from twitter, after this we need to perform different techniques to eliminate redundant data or unnecessary tweets from structured format data. Lastly we have several types of data to be analyzed and those are used by main researchers.

2.1 Data Sets

Analyzing structured data [4] have been universally used these days. For those cases, traditional Relational Database Management System (RDBMS) is used to deal. These days unstructured data is increased day by day from various sources (like Internet, Social Media, Blogging websites etc.) which are also known as Big Data. Any Single computer processor can't process this amount of enormous data. So RDMS is not sufficient to deal with such huge unstructured data. For this a nontraditional database is required to process all data, which is also called NoSQL database.

2.2. Data Retrieval

For data retrieval we need to find answer of few queries: Important characteristics of data retrieved?

Is data is static, user information such as "user id, name, gender"; or dynamic data such as user's tweet and its network? Which kind of data is important for analysis? How it will be processed? What is actual size of data collected? It is very easier to keep finding of certain keyword associated with any hashtag rather than keyword which is not associated.

2.3 Ranking and Classifying Twitter Users

User network has different types; a network of twitter users within any group, network of specific event (Hashtag), a network within particular user's account and many more such kind of network is created within network. Lists are also used to group certain set of different users to better organize the coming tweets.

For ranking users of twitter, we need to study important characteristics of twitter. This can be achieved by studying its topology and its all user details from dataset.

There are many people working with many methodologies to find the ranking analysis in better way. Twitter users are always ranked by knowing the number of follower for which we need to study the PageRank & tweet rate. A new technique is introduced by rank twitter users so that they can easily classify list of users into different users such as Politicians, Sports persons, Celebrities, bloggers etc.

2.4 Machine learning

Machine learning (ML) [5] is subpart of AI associated with the implementation of programs and algorithms that can learn autonomously. Machine learning has strong connections with statistical and mathematical optimization, whereas all of these areas targets to find out related concepts, regularities, and patterns from empirical data. Therefore, statistics and mathematical optimization provide methods and applications to the area of machine learning.

2.5 Natural language Processing

Natural language processing (NLP) is a field in AI (Artificial intelligence) and it is related to interaction across computers & human natural language [9]. As a part of Human-Computer Interaction NLP is concerned with enabling computers to derive and interpret human natural language. Recent work in NLP is algorithms based on ML and more specifically statistical machine learning.

2.6 Data Mining

Data mining [5] is the method of gathering required information from large data sets (commonly known as Big Data) in order to predict trends, behavior and other types of information that serve as a foundation for the organizations capability to make data-driven decisions. Knowledge discovery is fetching previous useful and unknown information from existing databases is an effective way.

3. PROPOSED WORK

Clustering can also be known as most valuable unsupervised learning problem [4] so, for every this kind of problem, it deals with providing a structure in collection of unlabeled data.

Other definition of clustering can be “the process of organizing objects into groups whose members are similar in some way”.

Bayes classifier

Given:

$X = [x_1, x_2, \dots, x_n]$, Classes $C_1, C_2, \dots, C_m$ and training set T

Define an algorithm which computes:

$P(C_i|X)$ using Bayes Rule

Return $C_{map} = \text{argmax}_{C_i} P(X|C_i)$

Naïve Bayes Classifier

Compute [29]:

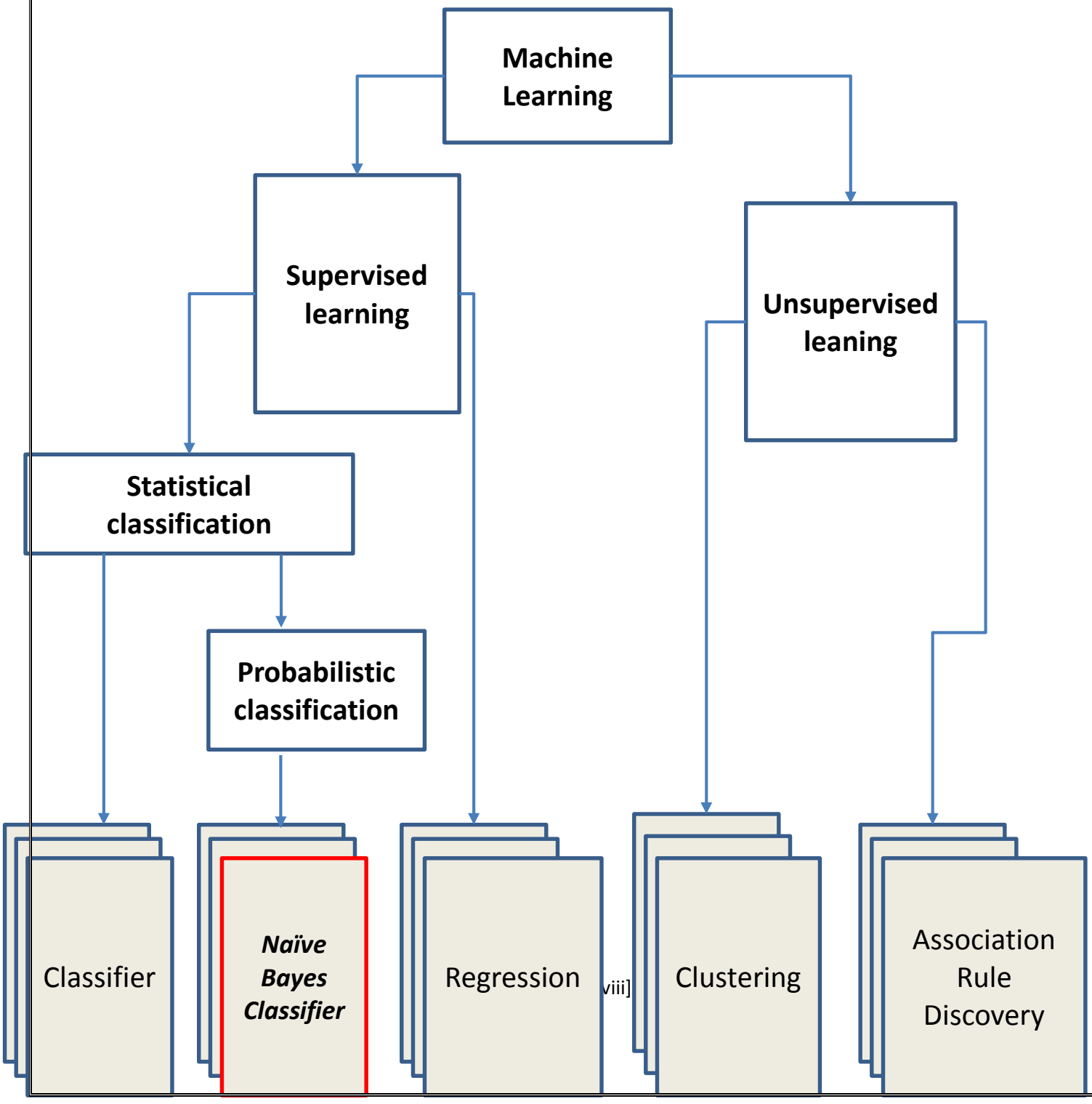
$K(A|C_i) = K(a_1|C_i)K(a_2|C_i, a_1) \dots K(a_n|C_i, a_1, a_2, \dots, a_{n-1})$ Which is known as a difficult problem?

Solution:

Naïve (Idiot) assumption – consider that $a, a_2, \dots, a_n$ attributes are independent so that

$$K(A|C_i) = K(a_1|C_i)K(a_2|C_i) \dots K(a_n|C_i)$$

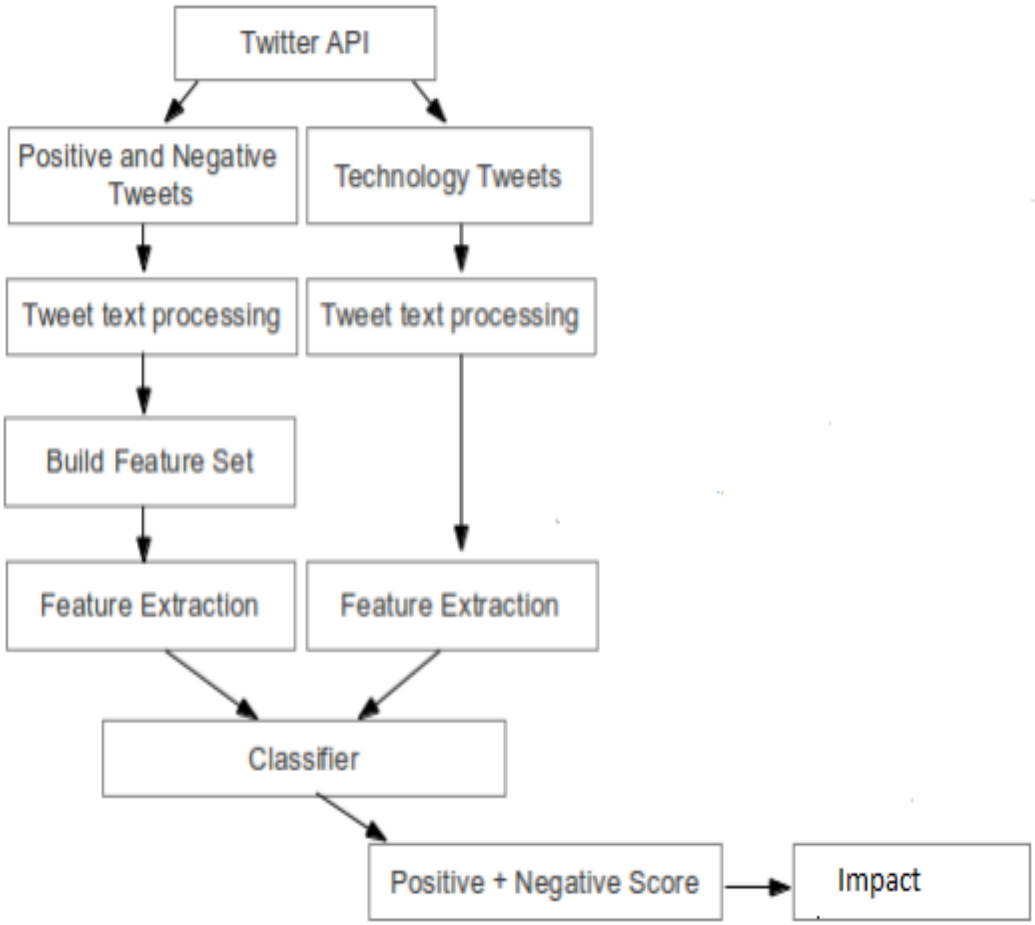
Figure 1: Showing Machine learning categorization



[viii]

3.2 Proposed Model

Figure 2: Showing proposed model



Twitter provides resilient API for tweet stack. Two potential ways to collect Tweets are: the search API or the streaming API. With the help of streaming API user can access the real-time tweets as a input query [14]. Whenever we send first request for connection to stream of twitter is required from the server.

Whenever streaming of server is opened, streaming connection and their tweets are synced as they come in to the user of the twitter. Although, it has few shortcomings, first language is not specifiable, resulting in stream which has tweets in all types of languages, including few non-Latin based alphabets as well. Moreover, a Stream tweet has only small part of the original tweet body. A polluted training set has proved much more difficult to get real dataset of tweets.

As this persist these kinds of issues in real life, we find better solution to use Twitter Search API [7]. Search API's are in REST format (or Representational State Transfer). By this way user request required queries of tweets which they want to get. It simply used HTTP methods (GET, POST, DELETE, and PUT) to perform operations. This API search has many filters as per requirement of the user like based on language, region and time. Rate limit is attached to these queries, but those are handled via code only, that's why rate limit is not an issue as of now. For getting the tweets we send our queries to search API and it revert us back with required response with small delay which is due to rate limit.

The query generated is made by strings of separate keywords coupled together with an OR used in between them. Even it may be not the complete output; it will always separate set of tweets that will be helpful for sentiment analyzer.

We need to train our sentiment analyzer, so that we can fetch data and make some sort of system to collect and organize dataset of tweets. So we first started working on collection of tweets that will server for us as training data set.

To train a sentiment analyzer and obtain data, we needed a system to collect tweets. Therefore, we first worked to collect a corpus of tweets that would serve as training data for our sentiment analyzer.

We have to categorize these tweets as positive or negative. Firstly, we have to decide whether to add "neutral" category sentiment label or not, then we will decide to check our analyzer to verify accuracy of all positive & negative labels used in this method.

Lastly we have to take decision to propose a methodology to make a collection of positive and negative sentiments with zero manual effort required in classifications. For achieving this task we require Twitter's search API as our base.

4. Methodology

A. Twitter Search/Streaming API

Twitter has provided a vigorous API which is used for tweet collection. Two way be which we can collect the Tweets are Streaming API and Search API. Real time access of tweets from queries is done by streaming API's. Initially User request a connections to twitter stream from server, once connection is stablished with server, tweets are then synced as per request.

B. Training Set Collection [1]

With the help of twitter search API, two separate datasets having collections of tweets as "Positive' and "Negative" [8]

Datasets are always formed programmatically and on the basis of this positive and negative queries on Keywords are formed:

- Positive sentiment: ":-) OR =) OR: D OR :) OR <3 OR love OR life"
- Negative sentiment: ":(OR = (OR dislike OR Ignore OR hate"

C. Tweet text processing

The text of each tweet has many irrelevant words in it which may not give proper meaning to sentiment. Users also user URL's, tag or they tag other people in their tweet also, which may not having any sentimental feeling or value.

To precisely calculate tweet's sentiment, we need to remove noise from the actual tweets. To accomplish this task we need to use various methodologies.

Many user tweets words and character simply add bit noise to sentimental analysis.

Table 1. Showing User and its Text message

User	Text
Apple Official	Iphone 6 coming soon!!
Johnbattelle	Big Data is Opening Doors, but Maybe too many
ladyGaga	Wow 21,000,000 monsters. Can we build a twitter Country now and all go live there. #NoSpeakingJusttweeting
Laurnes_deGroot	RT @shadowvieworg: @sustainablebizz

D. Tokenization

Initial step used in filtering the text by spaces, forming a list of each word per text. This is also referring as bag of words. These words can be used in features to train our classifier created.

E. Twitter Symbols

Maximum tweets have different symbols associated with it such as "@" or "#" and many more URL's [10]. Word just after "@" symbol is a username of the twitter user. This can be removed completely as it will not add any value to the text. Hashtag is also user in twitter very frequently to categorize information about any tweet, so we need to filter it out this also as it has no sentimental values.

F. TRAINING THE CLASSIFIERS

After we collect a huge amount of tweets with "positive" and "negative" sentiment, we can generate and train any classifier now. There are three types of classifiers: Naïve Bayes, Maximum Entropy and support vector machine [6].

For each tweet, we filter the same features from the Tweets to classify those words or sentence.

A. Feature Extraction

Classifiers & features are always selected based on easy usage and previous work in particular area [7]. Firstly we decide feature set whether it is useful to us or not, then we select that it is for our train feature.

1. Unigrams

A Unigram [5] is simply and N-gram of one size or a single word. Each and every word in tweet is created with the unigram classifier. Let make any positive tweet which contains “market”, first we check whether the word contains the word “market” or not. As the word “market” come from a positive tweet, our classifier will classify this word as a positive tweet.

2. N-grams

We filter bigram & trigram from our tweets as features to train our generated classifier. Bigrams are pairs of consecutive words, which add to the reliability of our classifier. In our N-gram, missing of gaps are not allowed and words must follow each other. Whenever we want to add bigram and trigram, we have increased features set by n-1 for bigrams and 9n-2 for trigrams, where n is number unique word among all the tweets.

3. External Lexicon

The list of worlds which are associated with particular sentiment which may be positive or negative, this list can be feed by an external lexicon called SentiStrength. With the use of SentiStrength lexicon we can easily broader coverage of twitter words which may be missed in collection from tweets.

B. Feature Filtering

Chi-Squared Information Gain

A chi-squared test is used to check the score of each word, bigram and trigram if used in our training set. In python’s Natural language toolkit allows us to calculate chi squared scores with the conditional frequency of each feature used.

C. Classifiers

Our classifier will do the labeling of future tweets as either "positive" or negative" based on attributes created by us. Here we are using three common classifiers used for text classifications:

1. Naive Bayes : The classifier is an application of Bayes Rule:
   $P(c|F) = P(F|c)P(c)/ P(F))$

2. Maximum Entropy

   For classifiers one should use the best possible model that will fulfill user given requirement that's why we use Maximum Entropy in this case. We use feature based models like MaxEnt. MaxEnt makes no independence premise for its features set, unlike Naive Bayes. We can add features like bigrams and phrases to MaxEnt without worrying about features overlapping. The model is represented by the following:

   $PME(c|d, \lambda) = \exp[\Sigma i \lambda i fi(c, d)]/ \Sigma c\ 0 \exp[\Sigma i \lambda i fi(c, d)]$

   In this formula, c is the class, d is the tweet, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas so as to maximize the conditional probability.

3. **Support Vector Machines**

   Support Vector Machines is another popular classification technique [10]. We use the SV Might [4] software with a linear kernel. Our input data are two sets of vectors of size m. Each entry in the vector corresponds to the presence a feature. For example, with a unigram feature extractor, each feature is a single word found in a tweet.

5. Experiment and Results

I have identified the filters/Classifiers that are needed to capture the real-time/Original value from the twitter user's datasets.

On the basis of the below classifiers when collectively used for large set of users, then we can get below results.

Table 2. Showing different classifiers and its value

Classifiers	Value
Tweet ID	7.7746E+17
User Name	Vikesh Singh
Profile	Software Professional
Tweet content	@Rajnathsingh: Evry tym we can't sit back just condemn it. If they have chosen their death. Go 4 it dis tym !!
Latitude	29.44702
Longitude	75.67180999999999
Place	Gorakhpur
Country	India
Followers	40
Following	230
Tweet language	English
Hashtag	1
Retweets	2
Total audience	40

The impact of our tweets is directly proportional to the state/Counter you belongs .

I have used a Index value to calcutate the the Population impact

Index key = **Population of State/Population of India**

Table 3. Showing Population and Index key of state

Population of	Index Key(Population of State/Population of India)
Uttar Pradesh	0.150
Maharashtra	0.120
Bihar	0.110
West Bengal	0.090
Madhya Pradesh	0.085
Tamil Nadu	0.084
Rajasthan	0.083
Karnataka	0.082
Gujarat	0.082
Andhra Pradesh	0.081
Odisha	0.077
Telangana	0.076
Kerala	0.075
Jharkhand	0.075
Assam	0.072
Punjab	0.071
Chhattisgarh	0.069
Haryana	0.067
Jammu and Kashmir	0.061
Uttarakhand	0.052
Himachal Pradesh	0.050
Tripura	0.035
Meghalaya	0.031
Manipur	0.030
Nagaland	0.028
Goa	0.027
Arunachal Pradesh	0.023
Mizoram	0.100
Sikkim	0.009

Below the are List Used to Create Test value Set. I have assigned a Weight to each polarity.

1 is least value used and 10 is max.

We use different different queries from Twitter API which are randomly choosen from many other domains also. Queries may consist of having information about compaines(Facebook & TCS etc), consumer products(Mobile phones,TV's) and people(Modi,Sachin,Virat). Below table shows the different categories of these queries which are listed

Table 4. Showing different Category with their assigned scores

Query String	Negative	Positive	Total	Category
at&t	3		3	Company
Google	4	9	13	Company
Obama	3	8	11	Person
UP Election		5	5	Event
Modi	2	9	11	Person
Insects		2	2	Misc.
IPhone	2	6	8	Product
IPL 2017	4	8	12	Event
Samsung	3	7	10	Company
Sachin	2	6	8	Person
Viral market		2	2	Misc.
Delhi	5		5	Location
Bahubali-2	1	7	8	Movie
Titanic		9	9	Movie

List of Queries which I have used to Create Test data Set

Below the are list of Emoticons with Ploarity. I have assigned a Weight to each polarity.

I have mainly catergoried Emoticons on the basis of Polarity which is

- Positive , Extremely-Positive
- Extremely-Negative, Neutral

1 is least value used and 10 is max is used by me .

Table 5. Showing Emoticons and their assigned weight

Emoticons	Polarity	Weight
:-) :) :o) :] :3 :c)	Positive	8
:D C:	Extremely-Positive	9
:-(:(:c :[	Negative	3
D8 D; D= DX v.v	Extremely-Negative	1
: |	Neutral	5

Part of the dictionary of emoticons and their weightage.

Below the are list of Tweets by the person having what profession he/she has.

Total Number of Audience and total tweets of the person will decide the impact of the user tweets on the audience weighted under scale of 1 to 10.

Table 6. Showing people tweets, total audience and weightage

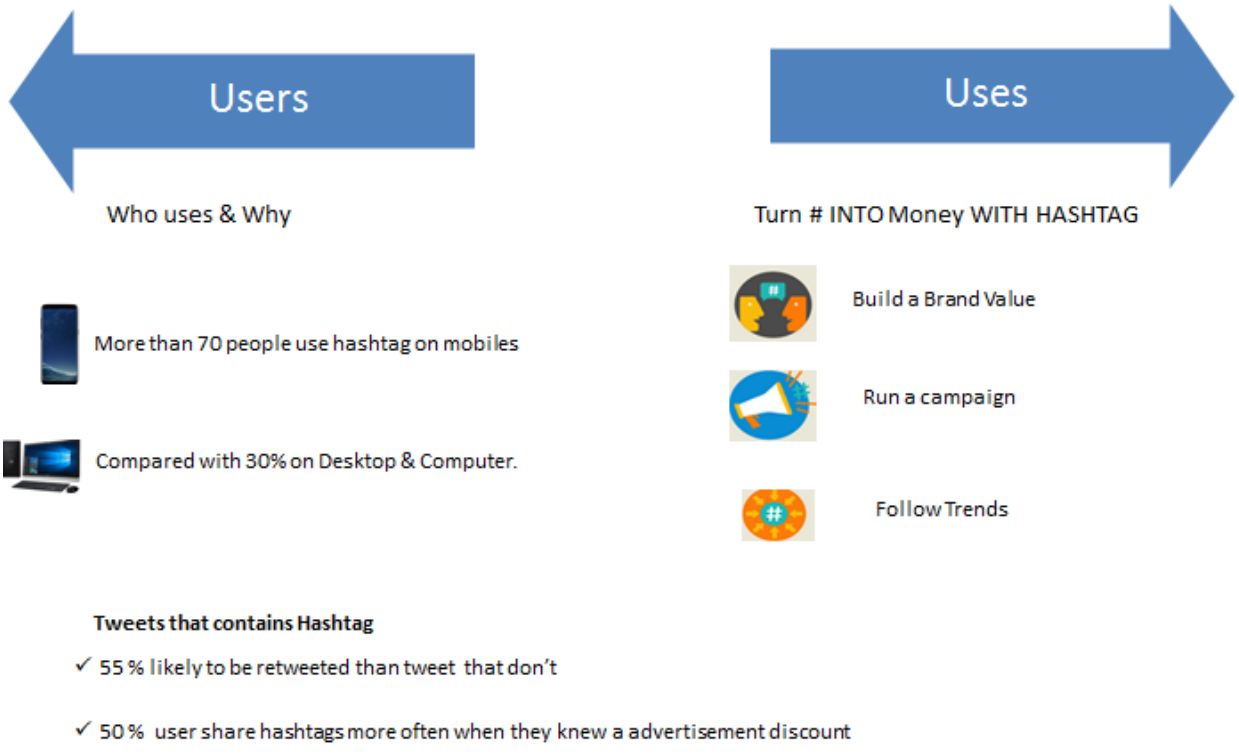
Tweet	Profession	Total audience	Total tweets	Weight
Narendra Modi	Politician	29900000	15100	7
Mark Zuckerberg	CEO	156000	3	2
Justin Bieber	Musician	94200000	30600	8
warren buffet	Investors	1200000	9	2
Tom Cruise	Actor/Model	6200000	11000	6
Sachin Tendulkar	Sports Personal	16000000	1554	5
Vikesh Singh	Software Professional	40	210	0.01

Hashtags

With the help of hashtags can organize Tweets by keyword [11]. Many people use the hashtag symbol (#) before a specific keyword or sentence in their Tweet which they are writing to categorize those Tweets and help them show more easily in Twitter search. Tapping on any hashtag created will listup all the tweets related to particular hashtags.

Although when you created any hashtag it will automatically become a clickable link. If someone want to check the hashtag, they can click on that hashtag and check all the most latest tweets that uses those particular hashtags [12]. This is done by the users to make them orgaize in such particular way that will make ease for other users to find and follow those tweets on the same theme or related topic.

Figure 3: Hashtags Usage and monitization

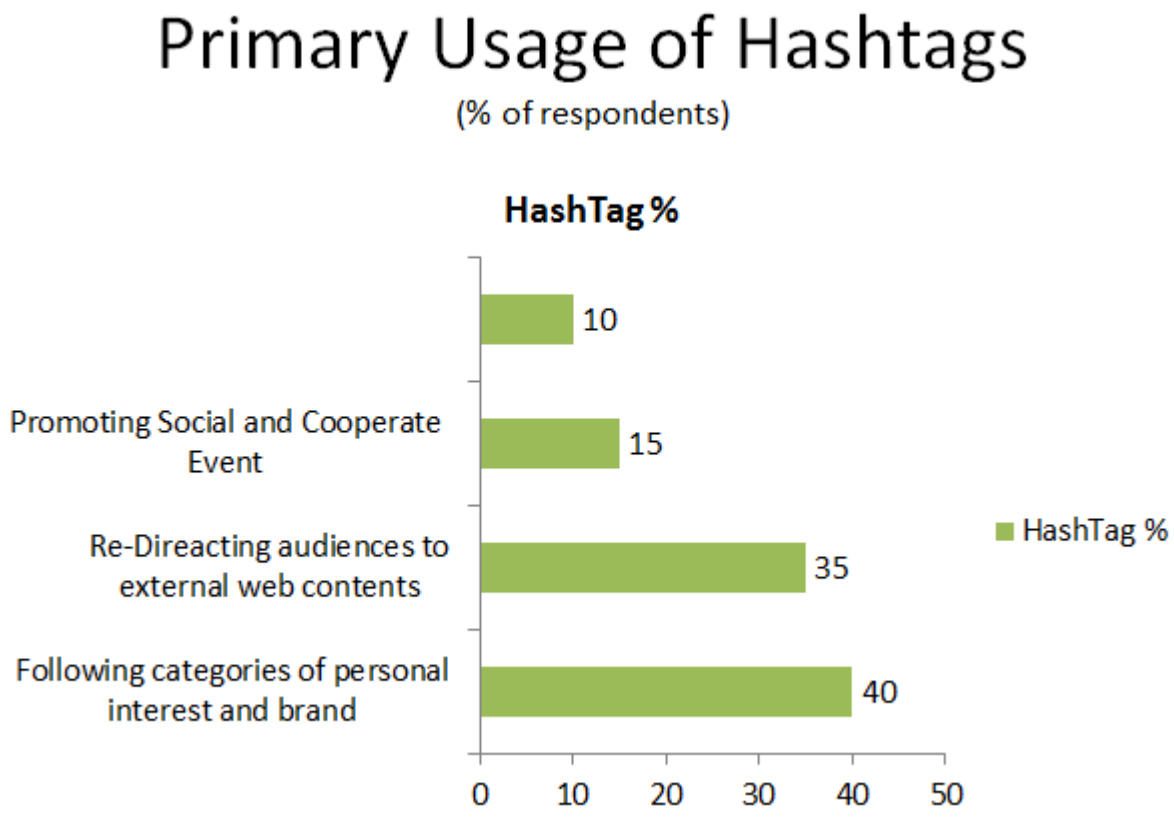


Above graphs showing the primary uses of hash tags. People mainly uses this for commutating personal ideas and feelings, searching brands and telling personal interest.

For sharing the videos and link and promoting corporate brands etc.

A hashtag is simply a way to categorize a tweet's topic, which then makes it easier for people to search for other tweets about those topics.

Figure 4: Primary Usage of Hashtags



Influence of hashtag categorized by me and putting a weight to the types of hashtag used in the tweets. I have categorized the Tweets as

- User: These hashtags are initiated by Users randomly, based on the feeling or current activities around the environment of the user.
- Event: These hashtags are used for the Particular event and time.
- Product: To promote Product and brand of the company.
- Always: These kinds of hashtags have long livelily and they exit for larger duration may be 6 month or 1-2 years or more.
- Misc.: Other all types of the hashtag used.

Below table shows the hashtags and their respective weight I used.

Table 7: Hahtag and respective weightage

Hashtag	Types	Weight
#3SaalGolmal	User	8
#YogiUPCM	User	8
#IPL2017	Event	7
#INDIAPAK	Always	6
#Bharatkveer	Always	6
#Ambedkarjayanti	Event	7
#Oscars2017	Event	7
#SamsungS8	Product	5
#ILovePizza	Misc	5

Retweets

A tweet is called a retweet when it is posted again with addition of some content in it. This features helps us to quickly share the tweet with your all the followers on twitter [12]. We can retweet anyone's tweet as well as our own tweet also. People also use RT at the start of any tweet to let you know that they are re-posting other people tweet.

A "retweet" is also community- driven phenomenon on twitter like hashtags which will helps us to make this service far better to it users and it allow users to increase his discussions in better suitable way.

As you have seen in Facebook , there is one feature which is known as reshare, peoples use to share their friends post or from any public walls on their own walls which is not originally posted by them, this facebook feature is typically the same as Twitter retweeting.

If you're familiar with Facebook, then you may have already seen a friend reshare a post that was originally posted by one of their own friends or one of the public pages they've liked. Facebook resharing is basically the same as twitter retweeting.

Impressions

Exposure [11] is total number of times tweets about the search term were delivered to twitter streams or the number of overall potential impression generated by any tweet created.

Impression means that tweet which is created has to deliver to main twitter stream by any particular account of the user.

In Union Metrics Twitter reporting [11], we define **reach** as the total number of estimated unique Twitter users that tweets about the search term were delivered to.

Result Analysis

My result analysis is based on the last 90 days tweeter data of any User.

The classifiers that I used in my result s are Tweet ID, Profile, total audience etc.

Below is the list of things which I can filter out easily using above research work

In my project:-

1. Top Followers
2. Gender Percentage of followers
3. Tweets per second/min/hour
4. Name & No of countries of the followers
5. Name & No of users state wise /City wise
6. Empirical Functional used to calculate Impact of a tweet of any person influencing others.
7. Tweets and retweets
8. Impressions

With these details I can calculate /Find out the areas where any political party/leader needs to work. Which sections of the society they need to focus for the popularity and their leads in Election.

With this research we can easily find out the area/Locality where the political leader has less influence. Surely he/she will start working on those areas to increase its influence in that area.

This research is very helpful in Election time by any Politian's. They can start working on their week areas before the election and heads towards Win- Win situation before Election starts.

I have used Yogi Aditya Nath last 90 days tweets for my research work.

Table 8: Top Followers & Gender Percentage of follower's analysis

Classifiers	Value
Tweet ID	Myogiadityanath
User Name	Yogi Aditya Nath
Profile	UP CM
Total audience	1940244
Toper follower	Mitalishah121 ETVUPLIVE Uppolice teawithdev akashbanerjee
Gender	Male : 66 % Female: 34 %
Tweets per second	0.833419288
Tweets per minute	50.0051573
Tweets per hour	3000.309438
Tweets per day	72007.42651

Table 9: Name & No of countries of the followers

Country	No of peoples
India	635
Pakistan	16
Argentina	4
United Kingdom	4
Singapore	3
United States	2
Australia	2
Oman	2
United Arab Emirates	1
Japan	1
China	1

Table 10: Name & No of user's state wise /City wise

India	No of users
New Delhi, India	68
Noida, India	24
Lucknow, India	20
Bhongaon Mainpuri U.P	16
Lucknow	14
GujRAT	13
Shahjahanpur, India	13
New Delhi	12
ALLAHABAD	12
Kanpur, India	10
Hathras, U.P. India	8
Mumbai, India	8
Agra, India	8
Ghaziabad, India	8
Allahabad, India	7
Varanasi, India	7
India	7
Kota, India	6

Orai, India	6
Mumbai	5
India	5
Bangalore	5
Gorakhpur, India	5
Pune, India	4
Delhi 84 india	4
Argentina	4
Varanasi	4
LUCKNOW	4
Rampur, Uttar Pradesh	4
Noida	4
faridabad,india	4
Amravati, India	4
Delhi	4
Utraula, India	4
Singapore	3
Sultanpur, India	3
KANPUR	3
Delhi, India	3
Noida	3
Agra U.P India	3
Azamgarh	2
Allahabad	2
Moradabad, India	2
India	2
Pune	2
B B Nagar Bullandshar up	2
New Delhi/Sonipat/Gurgaon	2
Sitamarhi Bihar, New Delhi	2
MUMBAI	2
Dhanbad, India	2
Jodhpur, India	2
Aligarh, India	2
Shimla	2
Mumbai, Maharashtra	2
Gujrat & Rajsthan	2
Shamli, India	2
INDIA ,UP	2

Kannauj, India	2
Mathura, India	2
Ghaziabad	2
Jaipur	2
Faridabad	2
Mumbai	2
Delhi NCR	2
Chennai	2
Bhiwani To Sirsa Train	2
Lucknow	2
Indore, India	2
Ghazipur, India	2
Ranchi	2
Jaunpur, India	2
Navi Mumbai, India	2
Bhiwani Haryana	2
DGP HQs LUCKNOW	2
Gonda, India	2
Anand, India	2
Kanpur	2
Shimla, India	2
Budaun, India	1
Ghaziabad	1
Gurgaon	1
Kanpur,Uttar Pradesh	1
Azamgarh, India	1
Ajmer, India	1
Bhilwara, India	1
Kanpur	1
Pihani, India	1
Andover, MA	1
Bikaner	1
DELHI, BHOPAL	1
Bareilly	1
agra,india	1
Kasganj, India	1
New Delhi (भारत)	1
New Delhi,India	1
Betul, India	1
Aliganj, India	1

Kanpur.india	1

Influence factor on Factor

Social influence [13] happens only when his/her view/thoughts or actions are influenced by any other people. We need to check and calculate influential users which are done by the message propagation by replying the below queries:

1. Know the real owner of the tweets.
2. Number of audiences of users.
3. Retweet rate of the initial tweet.

Table 11: User, Name, Followers and tweets

S.No	Name	Followers	Followings	Tweets
1	Katy Perry	108295028	205	8806
2	Justin Bieber	105199935	317217	30639
3	Barack Obama	98922592	625508	15492
4	Rihanna	85879643	1126	10062
5	Taylor Swift	85818552	0	83
6	Ellen DeGeneres	76898252	35957	15768
7	lady Gaga	76319460	128260	8655
8	You Tube	70948462	1030	21565
9	Cristiano Ronaldo	68000335	96	3088
10	Justin Timberlake	64920375	258	3770

I have determined a **Final Key Index value** (FKI) which has capabilities of determining user performance on twitter. It will evaluate the user's overall online influence with a score and ranged it from 1 to 100; with 100 is the maximum amount of possible generated influence.
It can combine 10 ~ 12 parameters which a have calculated earlier.
FKI has a "ability to drive people to action", doing replies and retweets are significant factor as well.

Table 12: Top Ten Twitter anonymous users in India

Twitter User	Followers IND	Total Followers
Rahul	8145	19587
Rohit	2625	5499
Vivek	2237	3277
Pinki	2837	3569
Pooja	2161	13524
Aditya	1935	4038
Abhishek	1847	3226
Neha	1822	2773
Anjali	1799	3226
Deepak	1993	3489

Table 13. Twitter users in India: No of Tweets, Following, No of list

Twitter User	# No of Tweets	Following	# No of list
Rahul	5459	149	788
Rohit	1316	2442	367
Vivek	868	112	256
Pinki	4081	1826	338
Pooja	2783	15456	312
Aditya	12638	1038	367
Abhishek	2863	354	286
Neha	6562	323	349
Anjali	744	147	246
Deepak	11401	743	354

Table 14. Follower/Following Ratio, Retweet and Mention Ratio, Interaction Ratio and RKI values.

Twitter User	F/F Ratio	RT Ratio	Interaction Ratio	RKI
Rahul	130.1	16.50%	3.40%	60
Rohit	2.3	4%	0.80%	44
Vivek	3.5	9.30%	1.80%	46
Pinki	2	15%	9.90%	57
Pooja	1	5%	0.70%	49
Aditya	3.9	5.90%	8.90%	49
Abhishek	9.4	3.50%	1.70%	58
Neha	9.8	8.70%	13.20%	57
Anjali	21.2	4.80%	1%	41
Deepak	4.8	8.10%	12%	58

F/F Ratio = Total followers/following
RT Ratio = Total no of tweets/(no of retweets+ mention)

Interaction Ratio= Followers/No of tweets

Rahul who is leading with almost 3 times as his runner up has its followers.

When we talk about most often used factor for calculating success on twitter, followers of Rahul, Pooja and Rohit leads the list (check table no 12). But when we check amount of interaction with other person in the user table i.e. retweets or mentions, Rahul is leading and Pooja and Rohit were on rank 7 and 9 only.

Pinki and Vivek has occupied ranks 2 and 3 respectively, which means their tweets are more interactions than the tweets of other users. When we calculate interaction ratio, Neha, Deepak and Pinki has Maximum number of individual users in there network.

TWITTER NETWORKING POTENTIAL

The two ratios (RT & Interaction ratio) is added up and divided and taking their average to calculate Twitter Network Potential value. 100 % TNP means all the tweets of any user which are acted upon and all followers of that twitter have interacted with the user via that tweet. It is always possible to have more than 100% TNP if a tweet is retweeted or mention more than its followers count.

Table 15. Ranking score of Users based on Twitter networking potential.

Twitter User	TNP
Pinki	12.45%
Neha	10.95%
Deepak	10.95%
Rahul	9.95%
Aditya	7.40%
Vivek	5.55%
Anjali	2.90%
Pooja	2.85%
Abhishek	2.60%
Rohit	2.55%

Pinki has topped this chart as she has maximum influential person, whereas Neha is ranked 2 in this chart.
Multiplying the TNP with the daily amount of tweets by the any user results in daily Twitter Networking Potential.

6. CONCLUSION

With this research we can easily find out the area/Locality where the political leader has less influence. Surely he/she will start working on those areas to increase its influence in that area.

This research is very helpful in Election time by any Politian's. They can start working on their week areas before the election and heads towards Win- Win situation before Election starts.

Many researchers already using this kind of twitter analytics and helping the political leaders to know the most and least influence area and work according before the election start to increase this confidence level in between people.

Below is the list of things which I can filter out easily using above research work

In my project:-

1. Top Followers
2. Gender Percentage of followers
3. Tweets per second/min/hour
4. Name & No of countries of the followers
5. Name & No of users state wise /City wise
6. Empirical Functional used to calculate Impact of a tweet of any person influencing others.

Future Scope

Based the functions generated, Classifiers and the weightage I assigned to the different functions, I generate a cumulative function that will calculate the Impact of the functions.

I will add servals more functions like use of Impression, Followers" versus "following" ratio.

1 – Number of followers the user has
2 – Number of tweets they've made
3 – The number of people that user is following

Moreover I will work on the different categories of the Hashtags used by the twitter users now a days and their life Cycle.

REFERENCES

- [1] Su Mon Kywe, Ee peng LIN, Feida ZHU "A Survey of Recommender Systems in Twitter" Institutional Knowledge at Singapore Management University, 2012
- [2] Hana Anber, Akram Salah, A. A. Abd El-Aziz "A Literature Review on Twitter Data Analysis", International Journal of Computer and Electrical Engineering, 2016.
- [3] Fred Morstatter, Huan Liu, Shamanth Kumar, "Twitter Data Analytics" Published in Springer Press, 2013.
- [4] Muqtar Unnisa, Ayesha Ameen," Opinion Mining on Twitter Data using Unsupervised Learning Technique " Published in 2000
- [5] N. Cristianini & J. Shawe-Taylor, "An Introduction to Support Vector Machines and other Kernal-based Learning Methods" Cambridge University Press, 2000.
- [6] C. D. Manning and H. Schutze, "Foundations of statistical natural Language processing" MIT Press, 1999.
- [7] Alec Go, Richa Bhayani, Lei Huang "Twitter Sentiment Classification using Distant Supervision" Stanford University, CA 2010
- [8] B. Pang, L. Lee, "Opinion mining and sentiment analysis" Published by Foundation and Trends in Information Filtering, 2008
- [9] Rong Lu and Qing Yang "Trend Analysis of News Topics on Twitter", International Journal of Machine Learning and Computing, June 2012
- [10] Pankaj Kumar, Kashika Manocha, Harshita Gupta. "Enterprise analysis through opinion mining", 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), 2016
- [11] Apoorv Agarwal, Boyi Xie, Owen Rambow, Rebecca Passonneau , " Sentimental Analysis of Twitter Data" , CSE, Columbia University, New York , 2013
- [12] Allison Shapp. "Variation in the Use of Twitter Hashtags", Qualifying Paper in Sociolinguistics, New York University, Spring 2014

- [13] Chang, H. C, “A new perspective on Twitter hashtag use: Diffusion of innovation theory ” , New York, 2011
- [14] Kacprzyk, Janusz,” Advances in Intelligent Systems and Computing”, Springer Press, 2015
- [15] Ayushi Dalmiya, Manish Gupta, Vasudeva Verma, “Twitter Sentimental Analysis : The Good, the Bad, the Neutral”, IIIT-H at SemEvam, 2015