

Sentiment Analysis on Political Speeches And It's Impact on Indian Politics

A Dissertation submitted in partial fulfillment of the requirements for the award of the degree of

Master of Technology

In

Information Systems

By

Vidyakamal

2016PIS1005

Under the guidance of

Dr. Swati Aggarwal

Assistant Professor, COE Department



**DIVISION OF COMPUTER ENGINEERING
NETAJI SUBHAS INSTITUTE OF TECHNOLOGY
NEW DELHI**

2018

DECLARATION

I Vidyakamal a bonafide student of **Master of Technology in Information System** in **Netaji Subhash Institute on Technology, Delhi** would like to declare that the dissertation entitled **Sentiment Analysis on Political Speeches And Its Impact on Indian Politics** submitted by me in partial fulfillment of the requirements for the award of the Degree of **Master of Technology in Information System** is my original work.

I carried out the research reported in this dissertation under the supervision of Dr. Swati Aggarwal (Assistant Professor) in the Division of Computer Engineering, Netaji Subhash Institute of Technology, New Delhi.

Date:

Place:

Vidyakamal

.....

Student Signature

CERTIFICATE

This is to certify that the Dissertation entitled **Sentiment Analysis of Political Speeches And Its Impact on Indian Politics** is a bonafide record of independent research work done by **Vidyakamal (Roll No.: 2016PIS1005)** under my supervision and submitted to **Netaji Subhash Institute of Technology, Delhi** in partial fulfillment for the award of the Degree of **Master of Technology in Information System**. This work is satisfactory for the award of the thesis credits.

Signature:

.....

Dr. Swati Aggarwal

Assistant Professor, COE Department

NSIT, Delhi

ACKNOWLEDGEMENT

I owe special debt of gratitude to my research supervisor **Dr. Swati Aggarwal** for her constant support and inspirational remarks throughout the course work. It was her mentorship that encouraged me to expedite my thesis work and could complete it in time. Her precious suggestions and constructive guidance has been indispensable in the completion of this thesis work. She has supported me in this endeavour and appreciated me in my efforts during the thesis work.

I would not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation. Last but not the least, I deeply acknowledge my faith in GOD almighty for his blessings, the love and support of my parents, classmates and friends who were a constant source of encouragement during the thesis work. Their moral was indispensable.

Vidyakamal

ABSTRACT

This thesis project addresses the problem of sentiment analysis of political speeches given by the leaders and we also find out the impact of speeches on elections. After last general election in India, the political landscape of the India has undergone through a significant changes which are speculated to stem mass adoption of social media. The political parties and there top leader are using social media network as a primary way to connect with the peoples. In this modern era of communication each and every one connected to the internet. And they tend to express their stands on social media website about products, movies, sports, social matter and even on government policies. With more and more people coming online, social media becoming one of the top medium for opinion sharing.

Many experts believed that social media is playing a crucial role on elections. In current scenario a huge population in India is expressing their view and opinion for government policies or for any social issue through social media. Politicians are frequently accused of changing their stance on issues or going along with the party's stance. We wanted to see whether politician's speech differed greatly depending on their party affiliation, or when or where they were speaking. Are these speeches have any significant effects on election results or not. In this thesis we study sentiment analysis on speech, tweets and the reaction of public as well, by extracting tweets from twitter with the help of twitter API, and speeches which are available in public domain. In this project I used machine learning algorithm naïve bayes and support vector machine algorithms for classification and NLP (Natural Language processing).

LIST OF FIGURE

Fig 1: Sentiment Analysis Approaches/Methods.....	16
Fig 2: Flow Diagram of analysis.....	20
Fig 3: Overall system architecture	22
Fig 4: Wordcloud Example	25
Fig 5: Create an application in Twitter API.....	29
Fig 6: Application credentials.....	30
Fig 7: Emotions words used in narendra modi public speeches	31
Fig 8: Word cloud of Narendra Modi speeches	32
Fig 9: Emotions words used in Rahul Gandhi public speeches	33
Fig 10: Word cloud of Rahul Gandhi speeches	34
Fig 11: Narendra Modi Total tweets sentiment %	35
Fig 12: Trend of positive negative or neutral on time	35
Fig 13: Trend of favorite and Retweet count	36
Fig14: Rahul Gandhi timeline tweets sentiment %	36
Fig 15: Trend of positive negative or neutral on time	37
Fig 16: Trend of favorite and Retweet count	37
Fig 17: Sentiment for @narendramodi used by people	38
Fig 18: Sentiment for @rahulgandhi used by people	38

LIST OF TABLE

Table 1: Techniques For Sentiment Analysis	9
Table 2: Top Hashtag used by BJP and INC	39
Table 3: Confusion Matrix fig. of Naïve bayes Algorithm	40
Table 4: Confusion Matrix fig. of SVM Algorithm	40

TABLE OF CONTENTS

Contents

DECLARATION	i
CERTIFICATE.....	ii
ACKNOWLEDGEMENT.....	iii
ABSTRACT.....	iv
LIST OF FIGURES	v
LIST OF TABLE	vi

CHAPTER 1

Introduction:	1
1.1 Motivation.....	1
1.2 Thesis Contribution.....	2
1.3 Thesis Organization	3

CHAPTER 2

Related Work:	4
2.1 General Sentiment Analysis.....	4
2.2 Political Sentiment Analysis	5
2.3 Related Work in Sentiment Analysis Challenges	8

CHAPTER 3

Background on Sentiment Analysis:	12
3.1 Introduction.....	12
3.2 Types of Sentiment Analysis	13
3.3 Related Work in Sentiment Analysis Methods	14
3.4. Sentiment Analysis Challenges.....	17

CHAPTER 4

Methodology & Implementation:	19
4.1.1 Political Speeches	19
4.1.2 Tweets	19
4.2 Proposed Technique (Overview)	19
4.3. The proposed Technique Methodology	21
4.4 Tools	27
4.5 Twitter.....	28

CHAPTER 5

Result and Analysis:	31
5.1 Sentiment Analysis on Speeches.....	31
5.2 Sentiment Analysis on Tweets.....	34
5.3 Model To Predict Tweet Sentiment	39

CHAPTER 6

Conclusion & Future Work:	41
6.1 Conclusion	41
6.2 Future Work.....	42
 References	 43

Chapter 1: Introduction

1.1 Motivation:

We can say that previous decades are golden period of time for the growth of the internet and it's growing each and every day. And by internet we are connected to each other by somehow or we can connect to anyone in the world just by a click and social media networks are best example of it. The best characteristic of the social media is the capability for large groups of people to interact or communicate instantaneously has completely changed the world. The growth of the internet has enabled all sorts of collaboration, large group communication, exchange of ideas, and the rapid dissemination of news to the population. Many domain experts believed in that social media is playing a effective role on elections and even the believe that its has change the election scenario. In current scenario a huge population of world is expressing their view and opinion for government policies or for any social issue through social media. People opinion based on what there like, dislike, interest etc. After last general election in 2014 in India, the political landscape of the India has undergone through a significant changes which are speculated to stem mass adoption of social media. As we know the NDA allies had won the elections and the political analysis expert believe that the social media campaigns play a crucial role to win the general election. And after defeat the UPA allies realize that they also need to strengthen their social media champagne. Now both political allies using the social media platform like facebook, twitter to express their thought to public what are their agenda of development and what they do if they come in to the power. And on other hand peoples are also expressing their opinion on social media openly what kind of polices, governance etc they want from ruling parties. Machine learning and artificial intelligence are emerging technology of current time and model to predict the results of important political elections or predict the sentiment on speeches as well as polls is also an emerging application to sentiment analysis. And this will help leaders and voter as well. Leader can get the people opinion and voter can know about their leaders.

1.2 Thesis Contribution:

This thesis aim to get the sentiment on speeches of political leader which were delivered during the election campaigning and also on tweets by the leader and their political affiliated party. And help in to find out which kind of the speech has been delivered and how the peoples are reacting. What's there opinion on the speech and tweets in terms of their negative positive or neutral opinion?

This project will group the speech words in **“Anger”**, **“anticipation”**, **“disgust”**, **“joy”**, **“fear”**, **“sadness”**, **“surprise”**, **“trust”**, **“positive”**, and **“negative”** categories by using NLP and display the percent emotions keywords which were used in speech. On tweets Dataset we are finding the tweet sentiment whether the tweets are negative positive or neutral. This project also contributes to find out the tweets sentiment in which these are tagged. Then we are find the trend of these tweets over time and using the tag extraction function we are making dataset of used keywords/Hashtags and then extracting those tweets. And then we are finding the trends of different sentiments.

In the project I created model to predict the sentiment of the tweets. And this model will help in to predict the sentiment and what is the accuracy by using two machine learning algorithms. By using word cloud we can also find out keyword by this project. And this will help use to understand what are the top priory of the leader.

1.3 - Thesis Organization:

This thesis work is organized in following way:

- **Chapter 2: Previous Interconnected work:** gives summary of the sentiment analysis and the research concluded in this field.
- **Chapter 3: sentiment analysis Background:** overview for sentiment analysis and the important definitions in this domain. It examines the differentiate sentiment analysis techniques architecture, the importance of sentiment analysis and the sentiment analysis challenges.
- **Chapter 4: Methodology and tools:** Explanation of methodology used in this project. How this methods work. There advantages and disadvantages and also brief introduction of tools which were used to do this project.
- **Chapter 5: Result of analysis:** Presentation of sentiment analysis result and showing the finding from analysis.
- **Chapter 6: Conclusion & Future Work:** It concludes the thesis and mentions the possible direction for future work.

Chapter 2: Previous Interconnected work:

There are various sentiment analysis had been done by researchers on political speeches and tweets. One of the more prominent ones that I found during my research, Presidential candidate (Trump & Clinton): Analysis and Visualization Tweets and Sentiments of Trump and Hillary. And there are various other works done on product review and movies review etc.

“Opinion mining or sentiment analysis refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials”. Sentiment analysis and opinion mining became one of the most important sources of decision making. But still several challenges need further attention.

The previous work on projects can be categories into two groups,

- Sentiment analysis research on general domain.
- Sentiment analysis research on political domain.

2.1 Sentiment Analysis on General Domain: Sentiment analysis research or in the area of opinion mining research started with product and movies reviews. And it can be categories in general sentiment analysis.

- In 2002 PMI(Point wise Mutual Information) were used to evaluate the sentiment orientation of phrases. And Pang et al as well Employed supervised learning with different sets of n-gram features in the same year, Given model by these author secured an accuracy of 83% with unigram presence features on the task of document level binary sentiment classification.
- An another author proposed a tool which evaluates the quality of text based on annotations on scientific papers. In this methodology they have used two approaches in sentiment of annotation to collects sentiments. This model evaluates total sentiment scores and counts all the annotation produces by the documents. The problem declares in a relationship between

annotations that is complex. The technique needs to have a big query knowledge base containing metadata.

- An author proposed a “Web Based Opinion Mining system” for hotel reviews. This project is an example of a general purpose sentiment analysis. The paper introduced an evaluation system for online user’s reviews and comments to give support to quality controls in hotel management system. The model is capable of detecting and retrieving reviews from the website and deals with German reviews. It has multitopic domains and is based on multipolarity classification. The system could recognize the neutral such that “don’t know” to “classify sentiment polarity that as neutral” and the multiple topic cases identified in their corpus.
- Product review for mobile device were analyzed by Zhang, et-al in. This research can help in evaluate accuracy. It is useful in a judgment of the product quality and status in the market. In this research author used mainly three algorithms K-nearest neighbor, Classifier, Random forest and Naïve Base, both supervised and unsupervised learning to evaluate the sentiments accuracy. The random forest is one of these three algorithms which enhance the performance of the classifier. There are some ways in analyzing sentiments and opinions. (Godbole, et-al) analyzed news sentiments and blogs. It splits prior work in the context of their specific task (sentiment analysis for news and blogs) into two categories. First category which - regards with techniques for automatically creating sentiment lexicon and the second one which relates to systems that analyze sentiment for entire documents.

2.2 Sentiment Analysis on Political Domain : Research in the sentiment analysis or in the area of sentiment mining which are centric to the politics or the sentiment analysis which is done on political leader’s tweets, speeches, or statements etc. In last few year, we observe that growing interest of political party and top political leader in online political opinion and sentiment in order to predict the result or outcome of elections.

- In the sentiment analysis of political domain Tumasjan et al presented most important paper for political sentiment analysis which focused on the 2009 federal election of Germany. The research focused to find out whether twitter can be used to predict election outcomes or not.

And one other important project in the sentiment analysis of political domain is Analysis of Presidential candidate (Trump & Clinton) 2016: it have Analysis and Visualization Tweets and Sentiments of Trump and Hillary and various others research are also available which show that sentiment analysis in politics have many miles to go in future.

- . Trump Vs, Clinton 2016: Analyzing and Visualizing The USA 2016 presidential election campaign has seen an unparalleled amount of media coverage, numerous presidential candidates, and various debates over wide-ranging topics from candidates of both the Republican and the Democratic parties. Micro-blogging website twitter is one of the top medium for people to to communicate, get the leader views, share their opinion, understand, relate and support the policies proposed by their favorite political leaders. In this project the author taken the tweets from twitterAPI and analyzed the sentiment of the tweets posted by two presidential candidate of democratic party and republican party (@realdonaldtrump and @hillaryclinton) on their Twitter feeds. Identified the most frequent policy- related keywords used by these `candidates, along with the twitter handles they most frequently mentioned in their tweets. Methodology used, extracted about 200,000 tweets accessing the live Twitter API The timeline for the analysis was from April 2016 to June 2016 concentrated on @realdonaldtrump and @hillaryclinton Twitter handles and also on trending hashtags like #trump2016 and #clinton2016 etc. e also collected information on the number of followers, retweets, and “favourited” tweets for both candidates.

Text mining and sentiment analysis were performed to focus on the following key points to better make the comparison between both candidates.

- Most used phrases on Twitter
- Which Twitter handles candidates most tweeted at
- Policy key words mentioned on Twitter
- Comparing sentiments on Clinton and Trump’s tweets

In the above sentiment analysis the SAS Sentiment Analysis Studio used to identify the sentiment distribution for tweets posted by Trump and Clinton. Different rules were specified for positivenegative and descriptors in accordance with the data collected.

- To find opinion poll correlation with political sentiment express in tweets O'Connor et al investigated to which extent they are related. To find out solution to this problem author used subjective lexicon which is proposed by Wilson et al., 2005. And by this they estimate the daily sentiment score for each entity. To define tweets sentiment, they proposed a tweet is positive if it have positive word and vice versa. Sentiment score for a particular day is calculated as the ratio of the positive over negative count. And by doing that results show that their sentiment score were co-related with opinion polls on presidential job approval but less strongly with polls on electoral outcome.
- In 2011 Choy et al discuss the “application of online sentiment detection” to predict the percentage of vote for all candidates which were contested in the presidential election of Singapore in 2011. They proposed a formula to evaluate the percentage of vote that each candidate will receive using census information on variables such as location, age group, sex, etc. Author combines this with a lexicon-based(Sentiment) sentiment analysis engine which process to calculates the sentiment in each tweet and aggregates the positive tweets and negative tweets sentiment for each candidate. Their application was successfully able to predict the narrow margin between the top two candidates but failed to predict the correct winner
- In 2012 Wang given sentiment analysis model which is a real-time sentiment analysis model for political tweets analysis which was actually analyzed the U.S.A presidential election of 2012. For this model wang extracted over 36 million tweets from twitter API and using Amazon Mechanical Turk they collected the sentiment annotations. For this model author used Naive Bayes model (Supervised learning algorithm)with unigram features, and Model achieved 60% accuracy on the four category classification.
- Bermingham and Smeaton proposed a model in 2011 for Irish General Election of 2011. Author analyses political sentiment in tweets by using supervised classification with unigram features, in this project author also used an annotated dataset different to and larger than the one we present, achieving 65.1% accuracy on the task of positive/negative/neutral classification and author gives conclusion tha volume is a stronger indicator of election outcome than sentiment, but the result also suggest sentiment still has a crucial role to play.

- In 2012 question arise that the use of twitter for prediction of election result. And to resolve this question, Gayo avello publish research work. Few last works which report positive results on this task using data from TwitterAPI are surveyed and partial results in their methodology and/or proposed methods noted. The focus of the author in this peper, is not the nonpredictive nature of political tweets but rather the accurate identification of type of sentiment expressed in the tweets the sentiment can be any type. If the accuracy of sentiment analysis of political tweets can be improved then this will likely have a positive effect on its usefulness or if accuracy cant be improve then its limitations at least better understood as an alternative or complement to traditional opinion polling.

2.3 Related Work in Sentiment Analysis Challenges:

For the purpose of this thesis, recognize the “sentiment challenges” means to find the sentiment challenges in evaluation and detection polarities for text or tweets and find the effective solutions for improving accuracy for text analysis. We can minimize the key of sentiment challenges in ten sentiment challenges that face the evaluation process of Sentiment reviews.

- Spam & fake detection
- Implicit & Explicit Negation
- Bipolar sentiments
- World knowledge
- Domain dependence
- Huge lexicon
- Natural language processing (NLP) overheads
- Pragmatics
- Thwarted Expectations
- Anaphora/co-reference Resolution
- Ambiguity

TABLE I: TECHNIQUES FOR SENTIMENT ANALYSIS

R*	Approach	Tools/Techniques	Experiment	Language Dependency	M/c Learning/Lexicon Based(ML*/LB*)	Data Scope	Data Source
1	User-Topic opinion prediction (2013)	Social context and Topical context incorporated Matrix Factorization (ScTcMF)	To predict the unknown user-topic Opinions.	Yes	LB*	Twitter	Tweets
2	Polarity shift in sentiment classification(2015).	Dual sentiment analysis(DSA)	Polarity classification task	No	LB*	Multi-domain sentiment English dataset. two Chinese dataset	Amazon.com, ChnSentiCorp Corpus
3	Qualitative analysis and large-scale data mining techniques(2014)	Naïve-Bayes multi-label classification algorithm	Show how informal social media data can provide insights into students' experiences.	Yes	ML*	Twitter	Tweets
4	Cross-domain sentiment classification(2013)	Corpus based	To evaluate the benefit of using a sentiment sensitive thesaurus for cross-domain sentiment Classification	Yes	LB*	Product reviews	Amazon.com

5	To interpret sentiment variations (2014)	Latent Dirichlet Allocation (LDA) based model, Foreground and Background LDA (FB-LDA), generative model called Reason Candidate and Background LDA (RCB-LDA)	To mine possible reasons of public sentiment Variations.	Yes	ML*	Twitter	Tweets
6	Sentiment and topic detection (2012)	Weakly supervised joint sentiment-topic (JST) model based on latent Dirichlet allocation (LDA, Reverse-JST)	To detect sentiment and topic simultaneously from text	Yes	ML*	Product reviews, Movie reviews	Amazon.com, IMDB movie Archive
7	Hashtag-level sentiment classification(2011)	SVM classifier	To automatically generate the overall sentiment polarity for a given hashtag in a certain time period, which markedly differs from the conventional sentence-level and document-level sentiment analysis.	Yes	ML*	Self-annotation manner to label the dataset, Twitter	Tweets

8	Sentiment polarity classification and sentiment strength detection (2012)	Hybrid approach (lexicon based + M/c learning)	To classify polarity and detect sentiment strength	Yes	ML* and LB*	Software reviews and movie reviews	CNET, IMDB
9	Sales prediction (2012)	Sentiment PLSA (S-PLSA, ARSQA, an Autoregressive Sentiment and Quality Aware model)	To Predict Sales Performance	Yes	ML*	Movie reviews	IMDB
10	Predicting Stock Price Movements (2014)	NLP techniques	To determine if the price of a selection of 30 companies listed in NASDAQ and the New York Stock Exchange can actually be predicted by the given 15 million records of tweets	Yes	ML*	Twitter	Tweets

Chapter 3: Background on Sentiment Analysis

3.1 Introduction:

Sentiment is can be seen as opinion, a thought based on a feeling about a situation, or a idea based on some feeling, or a process of thinking about something, soft feelings such as well as hard feeling for example love, sadness, affection etc. Text, videos and audio platforms are use to express these sentiments and analyzing these sentiments comes under sentiment analysis a to Wikipedia, opinion mining is also called as sentiment analysis, its refers to the use of NLP, analysis of text computational linguistics, biometric which is used in systematic identification , extract, study affective states, quantification, and subjective information. Sentiment analysis is widely work on customer materials such as reviews and survey responses, online and social media, and healthcare materials for applications that range from marketing to customer service to clinical medicine.

In simple words, sentiment analysis goals to find out the speaker attitude, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation, affective state that is to say, the emotional state of the author or speaker, or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).So basically, Sentiment analysis is practice of applying natural language process and text analysis techniques to identify and extract subjective information from piece of text. Basically sentiment analysis helps to find out whether a text is expressing the sentiment that is pos, neg, or neutral. Pos resfer as positive and neg as negative. Sentiment analysis is a best way to discover how people, particularly consumers, feel about a particular topic, product, or idea.

Sentiment analysis refers to the application of NLP, computational linguistics, and text analytics to identify and extract subjective information in source materials. Sentiment or sentiment opinion mining are based upon NLP.

3.1.1 Natural Language: in simple word known as ordinal language and this is refer to language by which people can communicate such as Hindi English or Urdu etc and these language are used by people for their each and every day communication. Human also used different - different sing to express there feeling so that is also a part of natural language.

3.1.2 Computational Linguistics: Generally computational linguistics involves in with the rule-based modelling or statistical based modellin of NLP from a perspective of computational. . Computational linguistics is the branch in which the techniques of computer science are applied to the analysis and synthesis of language and speech

3.1.3 Text Analytics: In simple words text analytics is basically process of text mining text analytics also include computation, linguistics, machine learning technique that model and structure the information. Before going into detail about some of the techniques mentioned above, it is important to realize what the concept of emotion in written text is.

3.2 Types of Sentiment Analysis:

Various kind of techniques and strategies used for sentiment analysis to identify the sentiments contained in a text. There are two main types of sentiment analysis.

- Objectivity and/or subjective identification sentiment analysis.
- Feature and/or aspect based sentiment analysis.

3.2.1 Objectivity and/or subjective identification sentiment analysis: Objectivity and/or subjective identification define as classifying a sentence of text into one of two categories: subjective or objectivity. However, we need to keep in mind that there are challenges when it comes to performing this type of analysis. The main challenge which comes that, the meaning of the word or even a phrase is often contingent on its context.

3.2.2 Feature and/or aspect based sentiment analysis: Feature/aspect identification gives way for the determination of different opinions or sentiments in relation to different aspects

of an entity. Feature and/or aspect based sentiment analysis Unlike subjectivity/objectivity identification, feature/aspect based identification allows for a much more broader overview of opinions and feelings.

3.3 Related Work In Sentiment Analysis Methods:

Current Sentiment analysis Approaches can be categories in following three group.

- Statistical approaches
- Lexicon/Knowledge-based approaches
- Hybrid approaches

3.3.1 Statistical Approaches:

Statistical approaches based on very basic elements from machine learning techniques such as support vector machine, latent semantic analysis, BOW, and semantic orientation. There are a few model of experiment that have been performed using more advanced methods which gives a try to identify the person who has expressed the sentiment and the target of the opinionated data grammatical relationships among. The words present in the piece of the text have also been used to mine the sentiment by Statistical/Machine learning based approaches: Support vector machine and NB classification methods are part of statistical/machine learning based approaches of sentiment analysis.

- **Supervise learning:** Supervised machine learning process we have labeled data set attributes. In this learning we have correct answer and there is teacher as well. In this machine learning we train function that maps an input to an output based on example input-output pairs. By supervised learning algorithm we try to training data and model a function, which will be useful to map further examples. We called It supervised learning because algorithm learn from the training dataset. In this type of algorithms we already know the correct answers. The algorithm predicts iteratively on the training data and if there is any wron prediction then that is correct by teacher.

- **Unsupervised learning:** In unsupervised learning the algorithms work on unlabeled responses. Unlike in supervised learning in unsupervised learning they do not have correct answer and there is no teacher as well. In unsupervised learning we perform basically clustering and classification. Algorithms are responsible by their own devices to discover and present the interesting structure in the data.
- **Semi-Supervised Learning:** In this type of learning we used large data set and the data set attributes are partially labeled. Suppose we have data set in which we have X attribute data which have label and remaining Y data attribute are not labeled this type of learning is called as semi supervised learning.

3.3.2 Knowledge/Lexicon Based Approaches:

This category of sentiment analysis methods relies on NLP and lexical resources to extract the knowledge from the opinionated data. This knowledge is further processed to identify the sentiment of a particular text. This approach categorize the opinionated data on the occurrence of unequivocal affect phrases such as sad, joyful, bored, and scared etc. This task can be accomplished in two ways: first method is based on the corpus which uses a list of opinion phrases as seed and then discovers other opinion phrases in the corpus. The second method is dictionary based approach to find sentiment seed phrases, and then seeks the dictionary of their antonyms and synonyms.

3.3.3 Hybrid Approaches:

In this approach we use both knowledge based and structured base technique . Using both the approaches machine learning and knowledge based together improves the performance and accuracy of the sentiment analysis task. Hybrid approaches in sentiment analysis uses the both elements from knowledge/lexicon-based and machine learning such as semantic networks and ontology's for detecting semantics which are stated in a sophisticated manner

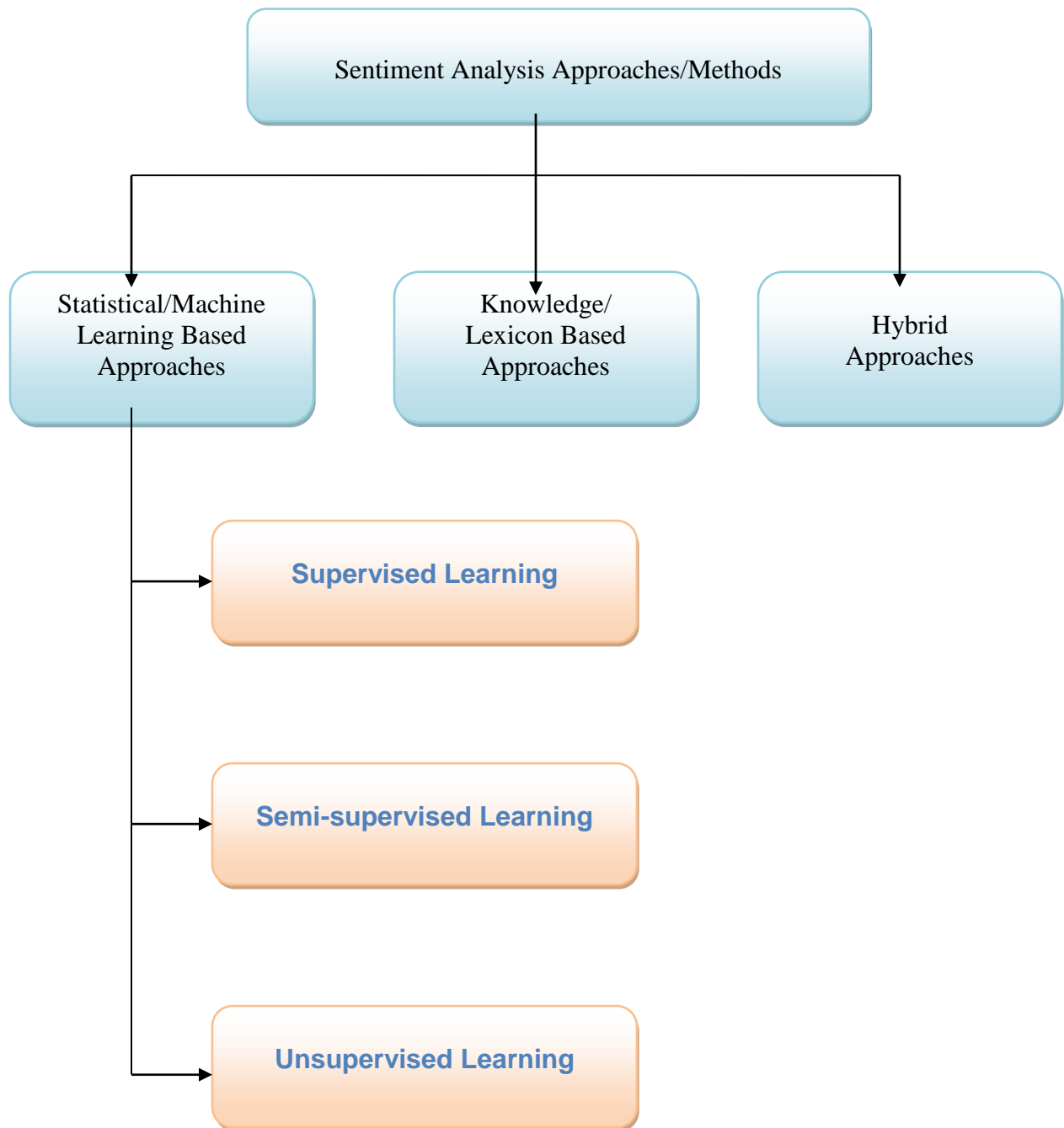


Fig: Sentiment Analysis Approaches/Methods

3.4. Challenges in Sentiment Analysis:

Following are some of the major challenges which were faced in sentiment analysis.

- Incremental Approach
- Parallel Computing For Massive Data
- Behavior/Homophily/ Credibility
- Sarcasm
- Author Segmentation review
- Grammatical Error
- Existing Lexicons refinement
- System designed for specific language
- Handling Noise & Dynamism

3.4.1 Incremental Approach: The major challenges of sentiment analysis arrived when we are using real time data because whenever new data is added we have to do analysis. Incremental approach gives liberty that we can use an existing result to be updated using only new individual data instances, without reprocessing previous results. This will very helpful when we don't have full dataset.

3.4.2 Parallel Computing For Massive Data: The major challenge in sentiment analysis is computing speed or computation time because in current time the social media are generating huge amount of data. And to get better computing we need to use parallel computing .

3.4.3. Credibility/Behavior/Homophily: Behaviors In sentiment analysis the credibility, behavior and homophile is hard define. Even if we mange to get behavior by data then it is difficult to validate that.

3.4.4. Sarcasm In text if author or speaker used sarcastic sentence then it is difficult to get the correct sentiment because for sarcasm the same sentence can be used to hurt someone or for offend, or in comic sense. This is major challenge which is hard to resolve.

3.4.5. Grammatically Error: We have various approaches that get sentiments but we have hardly an approach which can deal with grammatical error. If we get an effective approach which will be able to deal with grammatical error and map with correct words then results of sentiment analysis will improved

3.4.6. Review Author Segmentation Opinion for a target that is going to specified by various people who can be address as review authors. Total depending on the com style of comenting these authors, Tha should be grouped in to that so it becomes easy to get credibility. And this will help in decision making.

3.4.7. Refinement Of existing Lexicons: Many people comments, the Performance of sentiment analyzer depend on the correctness of the lexicon. Fine-tuning of existing lexicons is required to accommodate new words and destroy the words which are no more used for better results. Lexicon expansion through the use of synonyms has a drawback of the wording loosing it primary meaning after a few recapitulation.

3.4.8. Handling Noise and Dynamism Social media data are enormous, noisy, unstructured, and dynamic in nature, and thus novel challenges arise, introduces representative research problems of mining social media. Identifying and removal of noisy data is a challenging task.

3.4.9. Language Specific : This is the most major challenge which were faced in sentiment analys very often because most of the technique or tool are design to work on English or some specific language. This effect efficiency and accuracy of model we can improve technique to work on multi-languages in the scientific domain.

Chapter 4: Methodology & Implementation

As the name of thesis suggests it consist two parts, such as sentiment analysis on speech and the impact of the speech on election results. In this thesis project we perform sentiment analysis on political speeches which are available in public domain. And then we analyze the impact of speech on people through tweets. Manly political speeches are available in the video on youtube. So to make dataset of political speech we will convert these video to text. For analysis of impact of speech we will use tweets dataset.

4.1.1 Political Speeches: Elections empower citizens to choose their leaders. It gives all an opportunity for equal voice representation in our government Democracy is define as “for the people, and by the people”, that mean people choose government from the people. This means government leaders are determined by participation in elections. And to win the election candidate influence people by speech or by making promises to do when he gets power.

So the speeches which involve political interests are political speeches. Now in current time leader not only delivering their speeches through rallies but also they are using social media platforms like twitter facebook etc to convey their thoughts.

4.1.1 Tweets: In the project we are using leaders and their party tweets to extract the hashtag as well as to analysis of sentiment. By using these hashtag we extracting the public tweets which contain those keywords or hash tag. And then we are performing sentiment analysis on the dataset to get the respond trend over time. By the trend we can analyze the respond of public opinion.

4.2 Proposed Technique (Overview):

This thesis project contains two parts, One is sentiment analysis through the speeches and other one on tweets. For this thesis I have taken two top Indian leaders **Narendra Modi** and **Rahul Gandhi** public speeches for the election campaigning and tweets from their official twitter handle during the election. And also extract tweets from their party twitter handle. **Narendra Modi** represents the **Bharti Janta Party** and **Rahul Gandhi** to **Indian National Congress Party**.

I design relatively simple code in R for sentiment analysis of speech. In this project we have dataset of leader's speech and other data set we are creating by using the twitter API. In sentiment analysis of speech we are using natural language processing to extract the emotions or we can say that the different kind of emotions keyword and categories them into ten categories such as “Anger”, “anticipation”, “disgust”, “joy”, “fear”, “sadness”, “surprise”, “trust”, “positive”, and “negative”. And we are calculating the score value of these keywords and after that we are getting the overall sentiment or opinion of the speech using NLP and by using word cloud we also find out the most frequent words used in speeches. And for tweets analysis we use the NLP to predict the tweets sentiments whether they are negative positive or neutral. Using machine learning algorithms such as Naïve Bayes and SVM I have created a model to predict the tweets sentiment. To get the impact of the speeches and tweets, I extract keyword / Hashtag from the tweets of the official twitter handle used by leader and the party which he/she represent. And further extraction of tweets was performed using all those hashtag's. After making these tweets dataset, using the machine learning algorithms to get sentiment whether tweets are negative, positive or neutral and finding the trends and relation with the leader speeches or tweets.

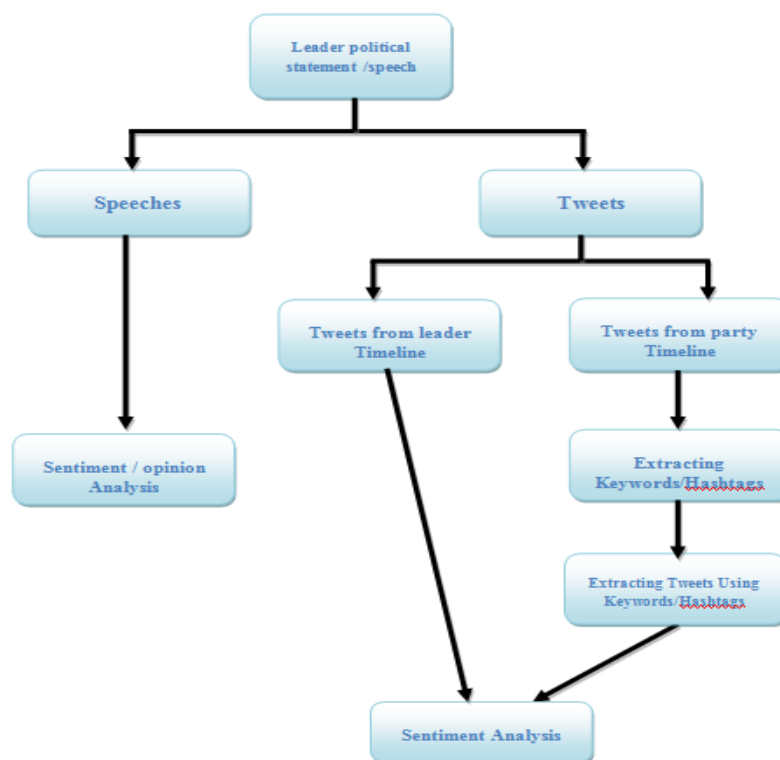


Fig 2: Flow Diagram of analysis

4.3. The proposed Technique Methodology

Sentiment analysis is the process of computation humans pinion, attitude, sentiment and emotions, and this sentiment expressed in written language as we know sentiment analysis is also known as opinion mining so we can say that all above sentiment can be compute to get opinion of peoples. Sentiment or opinion mining is on of the most active research areas in recent year. And the basic process of sentiment analysis is natural language process or NLP.

Models:-

4.3.1 Natural Language Processing: This is an area of computer science and artificial intelligence (AI) concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data. NLP is most reliable technique of sentiment analysis of opinion mining of text. For natural language processing we have many approaches and methods. When we perform analysis on a piece of unstructured text using natural language processing, each concept in the specified environment is given a score based on the way sentiment words relate to the concept and its associated score. This allows movement to a more sophisticated understanding of sentiment, because it is now possible to adjust the sentiment value of a concept relative to modifications that may surround it. Words, for example, that intensify, relax or negate the sentiment expressed by the concept can affect its score. Alternatively, texts can be given a positive and negative or neutral sentiment strength score if the goal is to determine the sentiment in a text rather than the overall polarity and strength of the text. Extracts sentiment and sentiment-derived plot arcs from text using three sentiment dictionaries

- AFINN
- Bing
- NRC

In the R language all above dictionaries are integrated in **syuzhet** package. In this package NLP is integrated as well. This package has three method for calculation of sentiment analysis AFINN, Bing, and NRC. All method use different dictionaries so the results also have differences. Syuzhet package use sentiment extraction tool developed in the NLP group at Stanford.

Stanford CoreNLP: This toolkit widely used, both in the research Natural language process community and also among commercial and government users of open source Natural language processing technology the inclusion of robust and good quality analysis components, and not requiring use of a large amount of associated baggage. It including the part-of-speech (POS) tagger, the named entity recognizer (NER), the parser, the coreference resolution system, sentiment analysis, bootstrapped pattern learning, and the open information extraction tools.

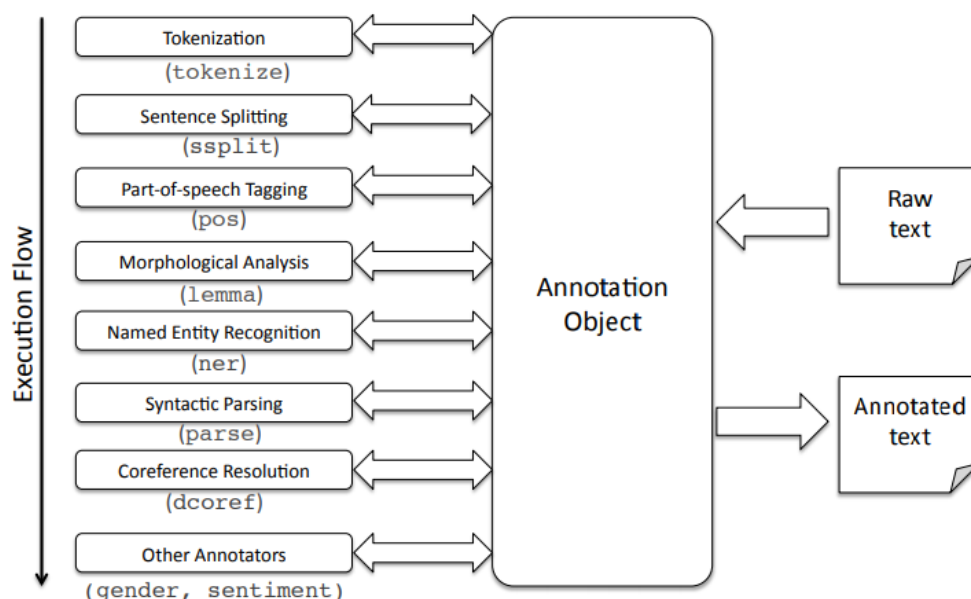


Fig 3: Overall system architecture StandfordCore NLP

4.3.2 Naïve bayes: A naïve bayes machine learning algorithm is a basically probability based algorithm. It uses the bayes theorem but assumes that the instances are independent of each other which is an unrealistic assumption in practical world naïve bayes classifier works well in complex real world situations. The naïve bayes classifier algorithm can be trained very efficiently in supervised learning for example an insurance company which intends to promote a new policy to reduce the promotion costs the company wants to target the most likely prospects the company can collect the historical data for its customers ,including income range ,number of current insurance policies ,number of vehicles owned ,money invested ,and information on whether a customer has recently switched insurance companies .

In this project we are training a model using the tweets dataset. And that dataset has two attributes: first is text and second one is class. Here class is basically sentiment positive or negative. To find class for the tweets we process the tweets on NLP to get sentiment. Using naïve Bayes classifier we can predict how likely a voter or public is to respond positively to a tweet posted on leader timeline. The naïve Bayes algorithm offers fast model building and scoring both binary and multiclass situations for relatively low volumes of data. This algorithm makes prediction using Bayes theorem which incorporates evidence or prior knowledge in its prediction. Bayes theorem relates the conditional and marginal probabilities of stochastic events H and X which is mathematically stated as

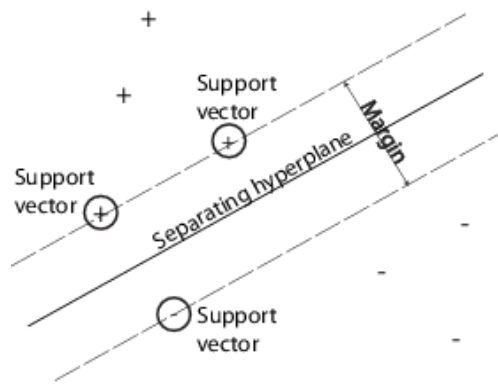
$$P(H|X) = \frac{P(X|H) P(H)}{P(X)}$$

- P : stands for the probability of the variables within parenthesis.
- $P(H)$: The prior probability or marginal probability of H ; it's prior in the sense that it has not yet accounted for the information available in X .
- $P(H/X)$: The conditional probability of H .
- X : posterior probability
- $P(X/H)$: conditional probability of X given H .
- $P(X)$: prior or marginal probability of X , which is normally the evidence. It can also be represented as

$$\text{Posterior} = \text{likelihood} * \text{prior} / \text{normalising constant}$$

The ratio of $P(X/H)/P(X)$ is also called as standardized likelihood.

4.3.3 Support vector machine: Support vector machine is a supervised learning algorithm which is used for classification as well as regression. SVM belongs to generalized linear classifier. In other word we can say that support vector machine is a prediction tool for classification and regression that uses ML theory to maximize accuracy for prediction while automatically avoiding over fitting. SVM can be defined as systems which use hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory. SVM becomes famous when, using pixel maps as input; it gives accuracy comparable to sophisticated neural networks with elaborated features in a handwriting recognition task It is also being used for many applications, such as hand writing analysis, face analysis and so forth, especially for pattern classification and regression based applications .The foundations of Support Vector Machines (SVM) has gained popularity due to many promising features such as better empirical performance .The formulation uses the Structural Risk Minimization (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle, used by conventional neural networks.SRM minimizes an upper bound on the expected risk, where as ERM minimizes the error on the training data .It is this difference which equips SVM with a greater ability to generalize, which is the goal in statistical learning .SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems. A support vector machine (SVM) is preferred when data has exactly two classes .An SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The best hyper plane for an SVM means the one with the largest margin between the two classes .Margin means the maximal width of the slab parallel to the hyper plane that has no interior data points .The support vectors are the data points that are closest to the separating hyper plane; these points are on the boundary of the slab



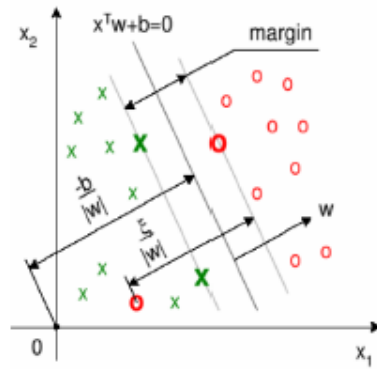


Figure 3: Geometrical Representation of the SVM Margin

4.3.4 Word Cloud:

Word Clouds are a way of showing which word was most significant towards the result and which word appeared frequently. Word clouds are a low-cost alternative for analyzing text from a large amount of textual data. The text is broken into component words and frequency for each is counted.



Fig 4: Wordcloud Example

In principle, the font size of a tag in a tag cloud is determined by its incidence. For a word cloud of categories like weblogs, frequency, for example, corresponds to the number of weblog entries that are assigned to a category. For smaller frequencies one can specify font sizes directly, from one to whatever the maximum font size. For larger values, a scaling should be made. In a linear normalization, the weight t_i of a descriptor is mapped to a size scale of 1 through f , where t_{\min} and t_{\max} are specifying the range of available weights.

$$s_i = \left\lceil \frac{f_{\max} \cdot (t_i - t_{\min})}{t_{\max} - t_{\min}} \right\rceil \text{ for } t_i > t_{\min}; \text{ else } s_i = 1$$

- s_i : display fontsize
- f_{\max} : max. fontsize
- t_i : count
- t_{\min} : min. count
- t_{\max} : max. count

4.3.5 Data Preprocessing

Before starting sentiment analysis the most important step is to pre process the data. As we are doing analysis on text and that is present in unstructured form. So if we perform analysis on row data the result will be far away from the reality. So before doing anything with text data we must first perform the data preprocessing or data cleaning. In this project we perform the same preprocessing method for both dataset it means for speech dataset and for tweets dataset as well.

For data preprocessing have perform following steps.

- convert text to lowercase
- Removal number
- Removal punctuation
- Remove English stop words
- Strip white space
- Special character removal

4.4 Tools

4.4.1 R : R is a scripting programming language and software environment for statistical analysis, graphics representation and reporting. R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. This programming language was named R, based on the first letter of first name of the two R authors Robert Gentleman and Ross Ihaka, and partly a play on the name of the Bell Labs Language S.

4.4.2 RStudio: RStudio integrated development environment for R and this software is open source and free for public use, RStudio was created by JJ Allaire, RStudio is available in two categories one is desktop and other is Rstudio server which allows accessing RStudio using a web browser. All the packages are available on major operating systems.

4.4.3 Tableau: Tableau is a very useful software which makes visualization and analysis very easy. In Tableau it is very easy for users to create and distribute an interesting and reliable dashboard, which is able to show the trends, variations, and density of the data in the form of visualization of graphs and charts. By Tableau we can easily connect to files. The software allows data processing and real-time collaboration, which makes it very different from other tools. Tableau is used by various kinds of professionals and students as well. In this project I used this software for visualization and analysis purpose

- Speed of Analysis
- Self-Reliant
- Visual Discovery
- Blend Diverse Data Sets
- Architecture Agnostic
- Real-Time Collaboration
- Centralized Data

4.4.4. Used R packages: In R function are the most important tool. Function have structure how to perform some particular task in the documentation all things are described how to use them, and sample data. And function can be download from <https://cran.r-project.org/> all function available for public use for free.

Following packages are used in this thesis project:

- **twitterR** : to connect twitter API and gets tweets.
- **ROAuth** : used to authenticate the API.
- **plyr** : used for common problem of read/write.
- **stringr** : used for simplify the string task or operation
- **ggplot2** : used for grammar of graphics
- **RColorBrewer** : used to draw nice graph according to variable.
- **tm** : Used to implement machine learning algorithms.
- **wordcloud** : this package is used to make word cloud of words.
- **Syuzhet**: This function is use to get sentiment using NLP

4.5 Twitter

Twitter is a amazing and free micro blogging service that provide platform to registered members to broadcast tweet and these posts short massage in length of 128 character. In addition,It can also be an amazing open source for text and social web analyses. Twitter members can broadcast tweets and follow other users' tweets by using multiple platforms and devices. Tweets and replies to tweets can be sent by cell phone text message, desktop client or by posting at the Twitter.com website.

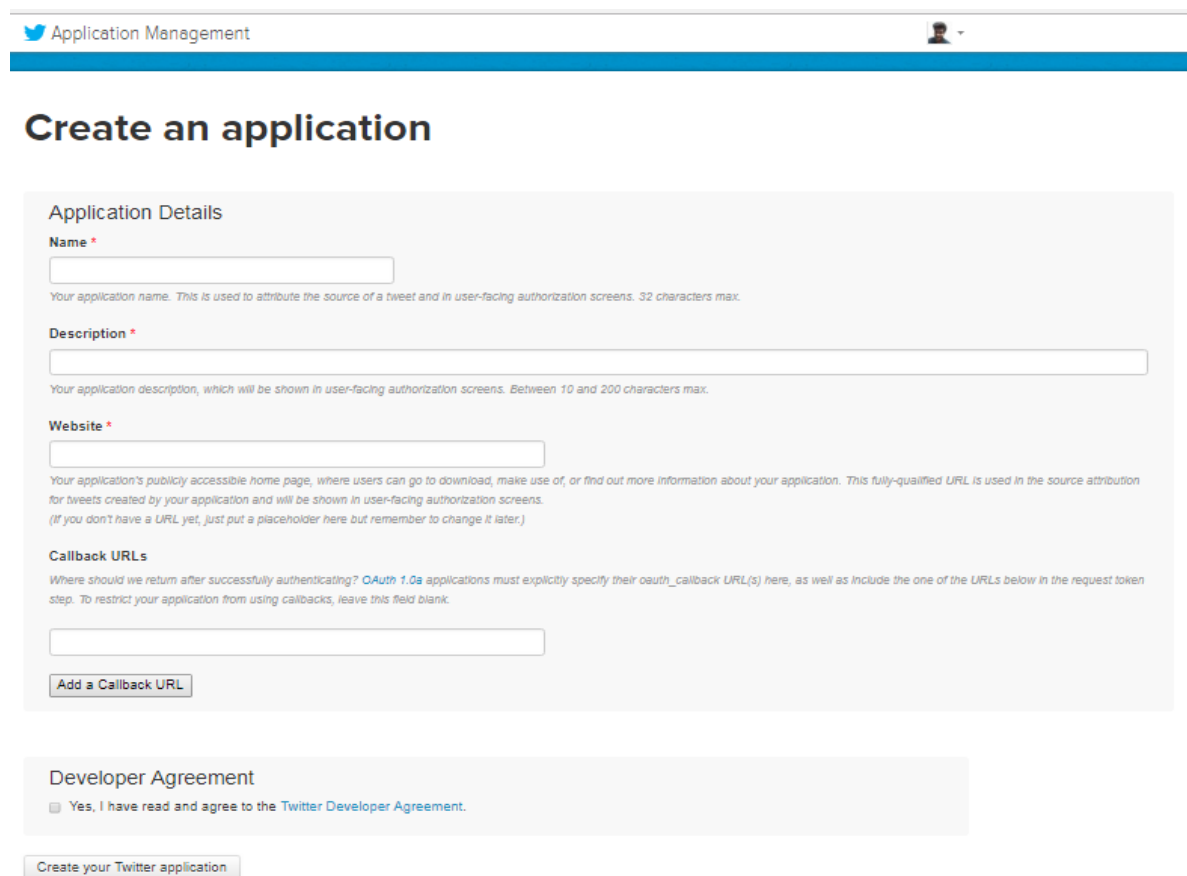
For sentiment analysis on tweets we first extract the tweets from twitter using Twitter API. Between all the different software's that can be used into analysis for twitter, R provide a wide type of options to do several of interesting and fun things. For this project I have used RStudio as its much easier working with scripts as compared to R. and to extract tweets from twitter we need to create Twitter App.

TwitterApp: Twitter has made extracting and analyzing tweets easier by developing an API. This is mainly used to extract tweets. The API helps us to extract data in a very structured format which can be then cleaned and processed for further analysis. To create a Twitter app, you first need to have a Twitter account. Once you have created, visit Twitter's app page and create an application.

Creation of Twitter Application

Starting step to perform sentiment analysis on tweets is to create a Twitter application. This application will give you access to perform analysis by connecting your R console to the Twitter using the API (Twitter). The steps for creating your Twitter applications are:

- Visit to <https://dev.twitter.com>.
- Use login details and once you are logged in your Twitter account.
- Click to My Applications -> Create a new application.



The screenshot shows the 'Create an application' page on the Twitter Developer Portal. At the top, there is a navigation bar with the Twitter logo and 'Application Management' text, and a user profile icon on the right. Below the navigation bar is a blue header with the text 'Create an application'. The main content area is a light gray box titled 'Application Details'. It contains four sections: 'Name' with a text input field and a note 'Your application name. This is used to attribute the source of a tweet and in user-facing authorization screens. 32 characters max.'; 'Description' with a text input field and a note 'Your application description, which will be shown in user-facing authorization screens. Between 10 and 200 characters max.'; 'Website' with a text input field and a note 'Your application's publicly accessible home page, where users can go to download, make use of, or find out more information about your application. This fully-qualified URL is used in the source attribution for tweets created by your application and will be shown in user-facing authorization screens. (If you don't have a URL yet, just put a placeholder here but remember to change it later.)'; and 'Callback URLs' with a text input field and a note 'Where should we return after successfully authenticating? OAuth 1.0a applications must explicitly specify their oauth_callback URL(s) here, as well as include the one of the URLs below in the request token step. To restrict your application from using callbacks, leave this field blank.' Below the 'Callback URLs' section is a button labeled 'Add a Callback URL'. At the bottom of the form is a 'Developer Agreement' section with a checkbox and the text 'Yes, I have read and agree to the Twitter Developer Agreement.' and a final button labeled 'Create your Twitter application'.

Fig 5: Create an application in Twitter API

Critosys

[Test OAuth](#)[Details](#) [Settings](#) [Keys and Access Tokens](#) [Permissions](#)

Information system

<https://en.wikipedia.org/wiki/Informatics>

Organization

Information about the organization or company associated with your application. This information is optional.

Organization	None
--------------	------

Organization website	None
----------------------	------

Application Settings

Your application's Consumer Key and Secret are used to **authenticate** requests to the Twitter Platform.

Access level	Read and write (modify app permissions)
--------------	---

Consumer Key (API Key)	UMjXC86wAdpUtkLLFXrqFJ9ig (manage keys and access tokens)
------------------------	---

Callback URL	None
--------------	------

Callback URL Locked	No
---------------------	----

Sign in with Twitter	Yes
----------------------	-----

App-only authentication	https://api.twitter.com/oauth2/token
-------------------------	---

Request token URL	https://api.twitter.com/oauth/request_token
-------------------	---

Authorize URL	https://api.twitter.com/oauth/authorize
---------------	---

Access token URL	https://api.twitter.com/oauth/access_token
------------------	---

Fig 5: Application in Twitter API

- Name your application and describe about your application in few words, provide your website's URL or your blog address (in case you don't have any website). Leave the Callback URL blank for now. After this process just do other formalities and create your twitter application. After doing all the steps the created application will show as below. Then after your have to note the Consumer key and Consumer Secret numbers as it will be used in RStudio later.

Creation of Twitter App has been done, the important credentials such as token and key are required to access the twitter API . Following keys and tokens will be required for access.

- API Key: - Consumer Key
- API Secret: - Consumer Secret
- API Token: - Access Token
- API Token Secret: - Access Token Secret

Chapter 5: Result and Analysis

The election commission of India has announced the date for Karnataka assembly elections on 29th April. Elections will be held in a single phase on May 12 and the results will be announced on May 15. In the Karnataka assembly election top two parties were BJP and INC. In this project we consider speeches of BJP top leader Narendra Modi and Rahul Gandhi from INC. The results of sentiment analysis of their tweets and speech are following.

5. Speech and Tweets Analysis

5.1.1 Narendra Modi Speech: During the Karnataka assembly election campaigning of BJP Narendra Modi done more than 16 public rallies and given public speech to influence the voter's. In this project we have taken 11 public speech which were delivered at Bellari, Belagavi, Chamarajanagar, Udupi, Jamakhandi, Bengaluru, Kalaburgi, Hubballi, Shivamonga, Tumkuru, Mangaluru. Following bar graph is showing the ten sentiment and there percentage in speech

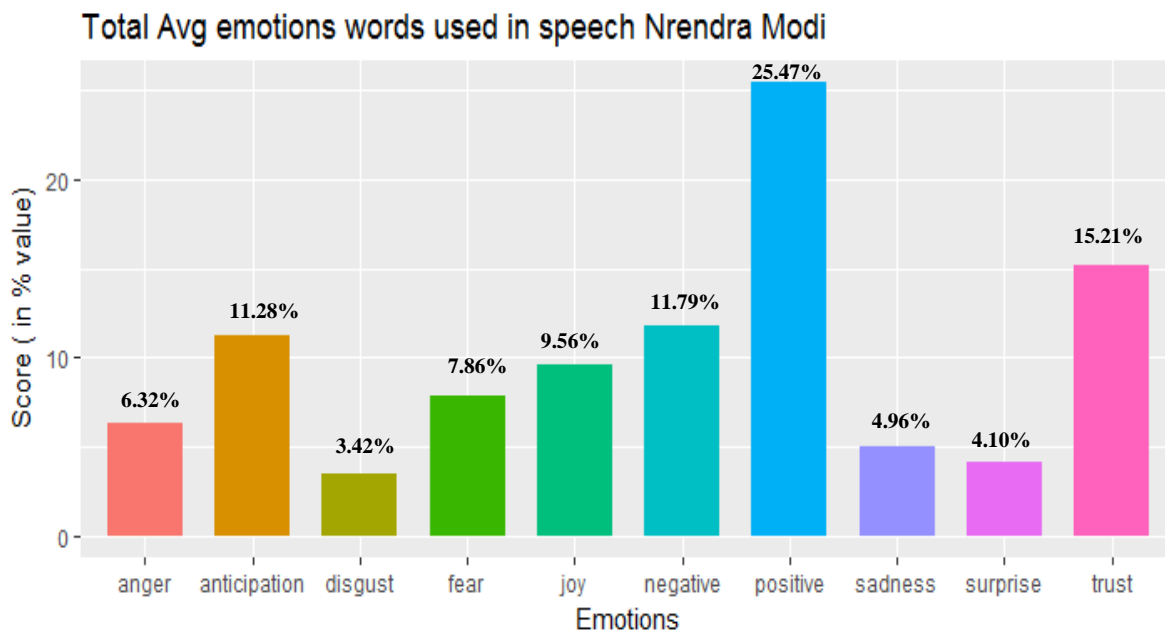


Fig 7: Emotions words used in Narendra Modi public speeches

The sentiment analysis show that the all the speech have 6.32% of anger, 11.28% of anticipation 3.42% of disgust 7.86% of fear 9.56% of joy 11.79% of negative words 25.47% of positive words used. There are also 4.96% of sadness 4.10% of surprise and trust 15.21%. Following word cloud graph show the most used word in the speeches.



Fig 8: Word cloud of Narendra Modi speeches

Now let's co-relate the sentiment bar graph and word cloud. Sentiment graph show that the sentiment words anticipation, negative, positive, and trust have present in speeches are greater than 10%. And by word cloud figure we can say that in the speeches by Narendra Modi he was mostly talking about his government achievement of centre, which currently running by his own party, attacking opposition because state government currently ruled by congress party, public concern etc. Now let's take the emotions which were greater than 10% in speeches. Positive words sentiment is around 26 and the most used words are congress, government, work, farmers, technology, electricity and you so we can assume that the Prime Minister Narendra Modi is speaking about his government achievement. And if we combined the trust factor with positive sentiment we can say in his speech he have trust on public that they will vote for him. And

negative sentiment is for congress, ministers, your, family, and Karnataka to criticize state government which is ruled by congress and by adding sadness, disgust, anger attributes we can find a bigger picture of his speech.

5.1.2 Rahul Gandhi Speech

Rahul Gandhi started Karnataka election campaigning in February and during the Karnataka assembly election campaigning of INC he covered 25 out of 30 districts of Karnataka and given around 23 public speeches. In this project we had taken 11 speeches for sentiment analysis which were delivered at Koppal, Yadgiri, Raichur, Bellari, Udupi, Bengaluru, Kalaburgi, Hubballi, Shivamonga, Tumkuru, Mangaluru, Ankola, Kumta, Honnavar, and Bhatkal. Following graph is showing the ten sentiment and their percentage in speech

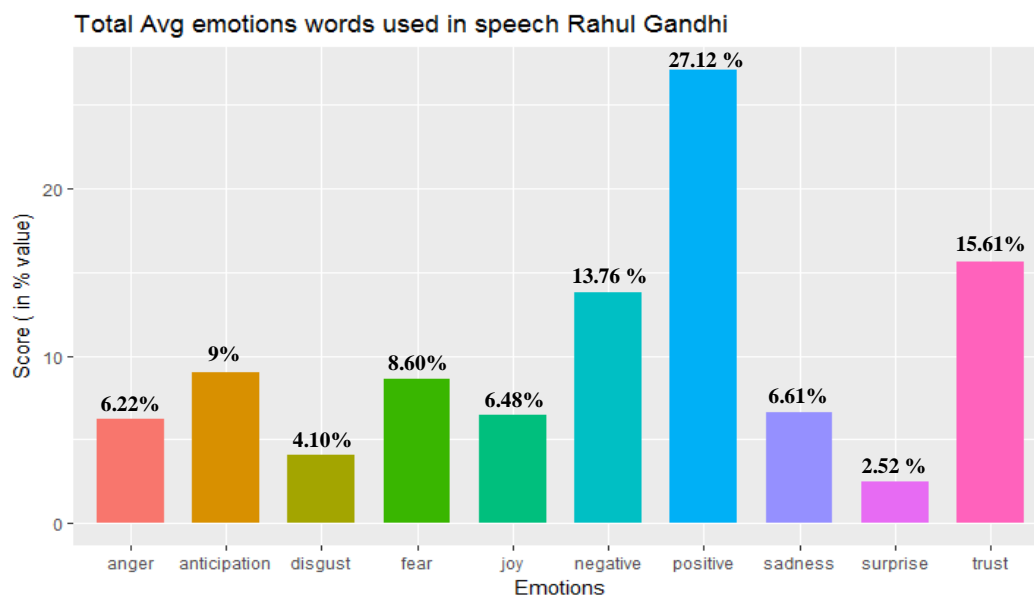


Fig 9: Emotions words used in Rahul Gandhi public speeches

The sentiment analysis shows that all the speeches have 6.22% of anger, 9% of anticipation, 4.10% of disgust, 8.60% of fear, 6.48% of joy, 13.76% of negative words, 27.12% of positive words used. There are also 6.61% of sadness, 2.52% of surprise and trust 15.61%.

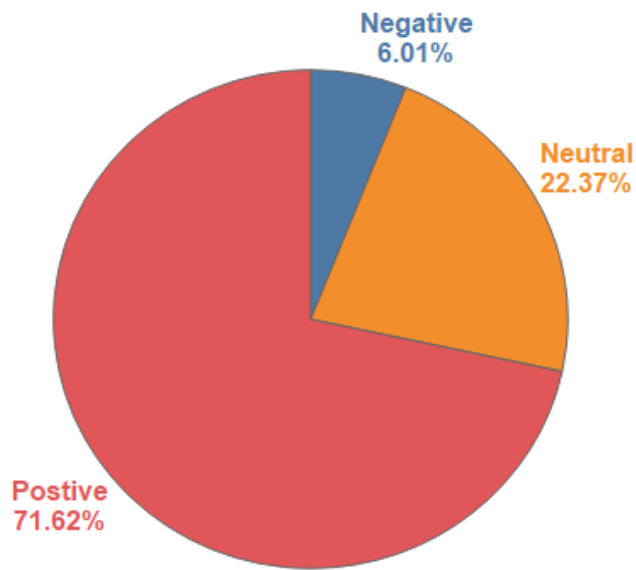


Fig 11: Pi chart of Total tweets sentiment in %

Above pi chart show that most of the tweets or retweets have positive sentiment or neutral only 6.01% of tweets are negative. And following trend graph showing tweets trend over time and that show during the election period the no of positive tweets has very drastic change and little bit down fall in neutral tweets. And litter up change in negative tweets and following favorite and retweets count show that during the time of election both favorite and retweets count increase so we can say during election time the leader supporters are actively like and share the post of the leader.

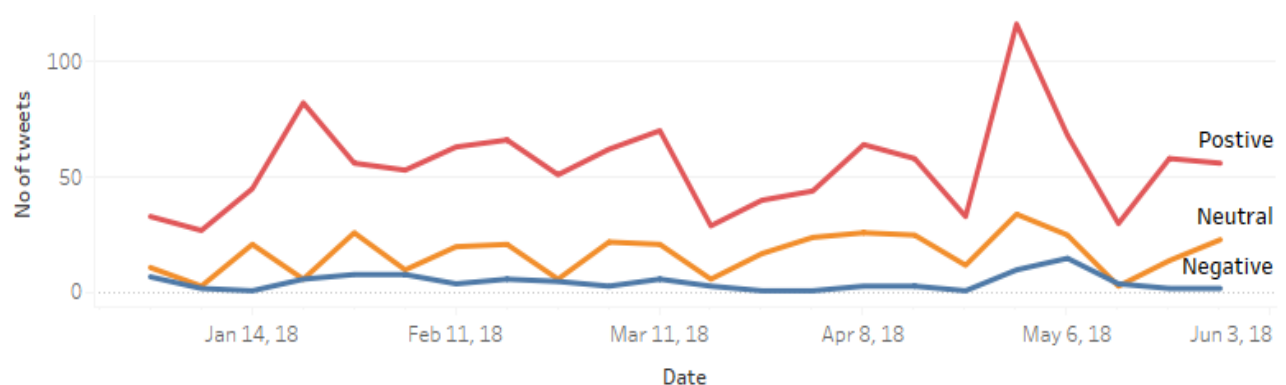


Fig 12: Trend of positive negative or neutral on time

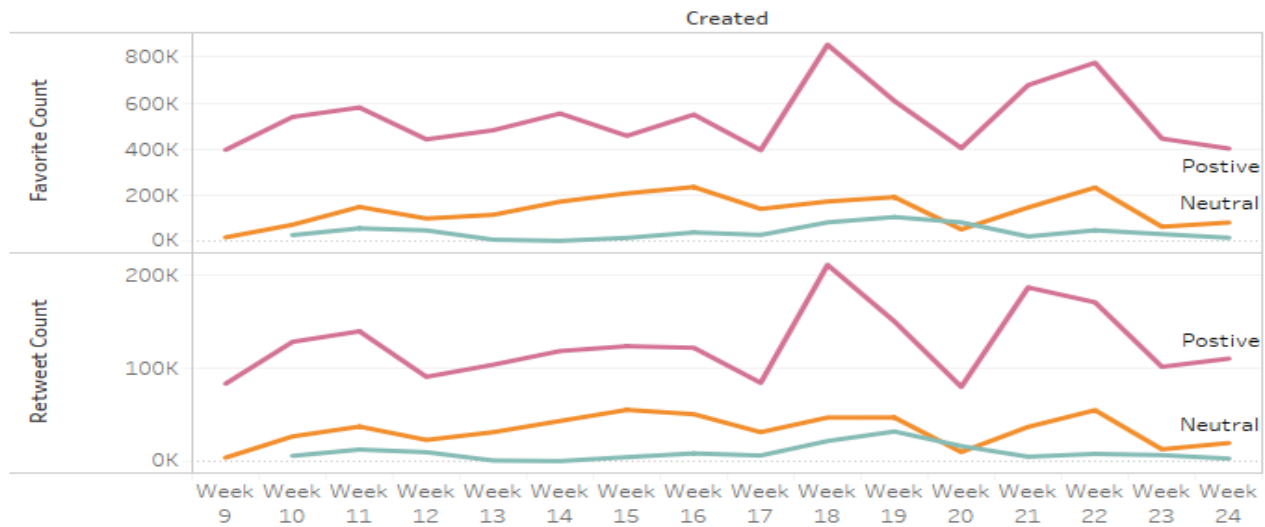


Fig 13: Trend of favorite and Retweets count

Narendra Modi twitter time line tweets trends are very predictable there are increment in positive tweets during the election time only.

5.2.2 Rahul Gandhi Twitter Timeline: After Rahul Gandhi twitter timeline analysis we have found that Rahul tweets much more negative sentiment tweets as compare to Narendra Modi Statistics show that in 3200 tweets and retweets there are 49.17% is positive 28.33% neutral and 22.50% are negative tweets.

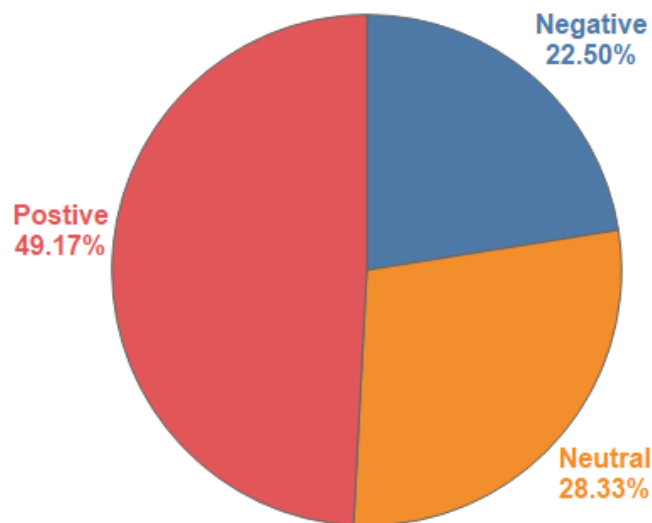


Fig14: Rahul Gandhi timeline tweets sentiment %

And the negative tweets are around four times of Narendra Modi negative tweets and so Rahul Gandhi have lesser positive sentiment tweets as compare to Narendra Modi.

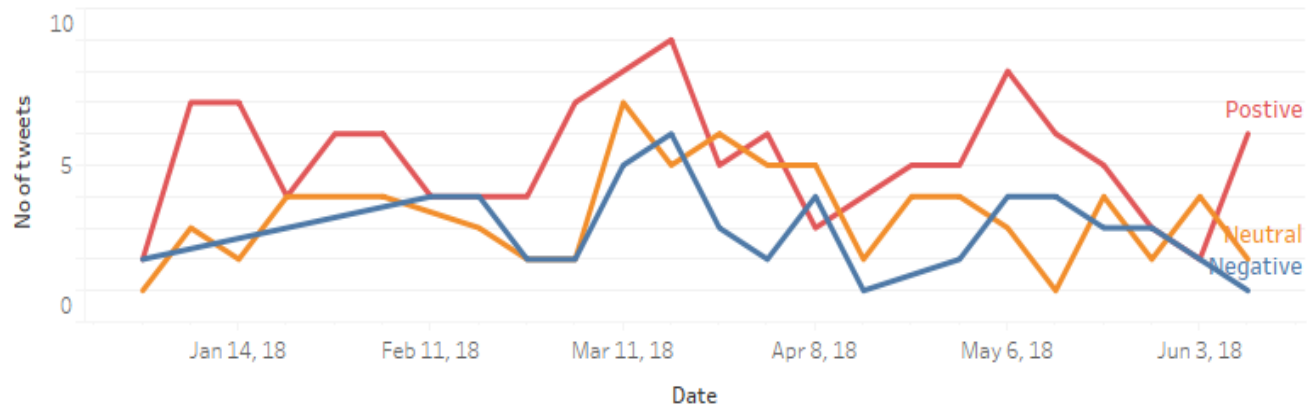


Fig 15: Trend of positive negative or neutral on time

And as we can see that the trend graph does not have any pattern so it is very difficult to predict why negative positive or neutral tweets have so much ups and down. But during the election of Karnataka which was held in may there are sharp increase in positive and negative tweets and very sharp down fall in neutral tweets.

And the next trend graph of favorite count and retweets count also showing the same trend. Which is also unpredictable but there are increment in favorite and retweets count.

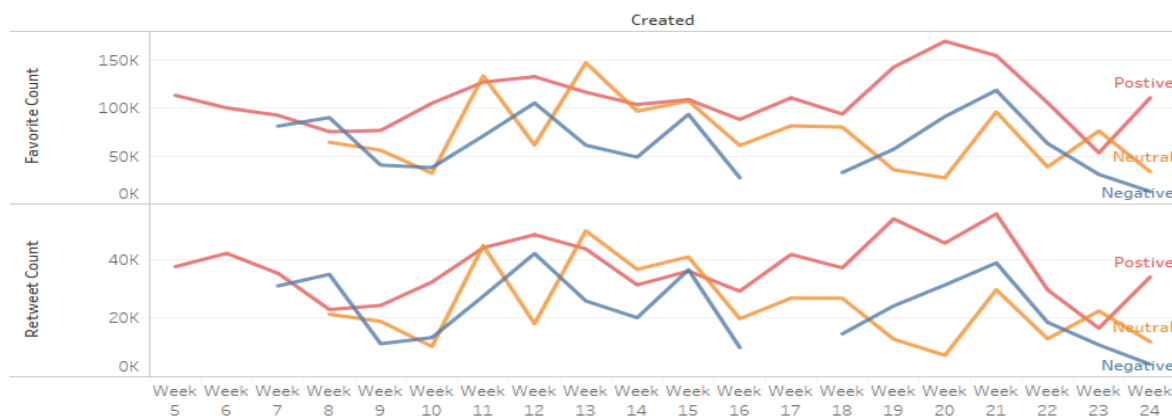


Fig16: Trend of favorite and Retweets count

5.2.3 Tweets in Which User Name Tagged: Now we are extracting tweets which have the string @narendramodi and @RahulGandhi. In tweets we use twitter use name to tag the person. So in this portion of analysis we are going to analyze those tweets in which people tag Narendra Modi and Rahul Gandhi. This part of analysis will help use to understand what are the people sentiment when they are tagging these leader. For this part of analysis we are using data set which have 10000 entry or we can say tweets for both the user name individually for @narendramodi and @RahulGandhi.



Fig 17: sentiment for @narendramodi used by people

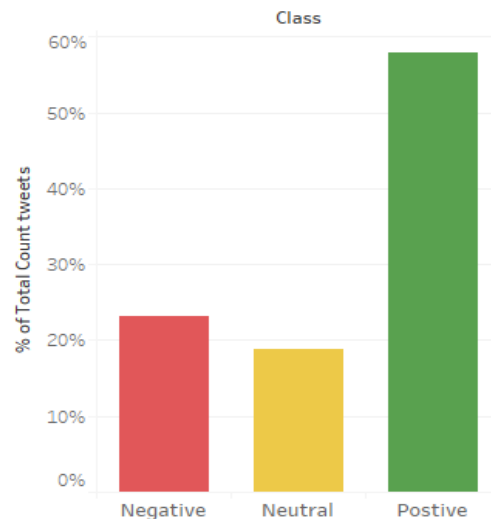


Fig 18: sentiment for @RahulGandhi used by people

Previous bar graph showing the sentiment of tweets in terms of positive negative and neutral for tweets which had @narendramodi and @RahulGandhi string. And all the tweets are created in between 1st April to 31st may 2018.

Now if we consider the first bar graph which is showing the tweets sentiment in which @narendramodi were tagged. This analysis showing that for Narendra Modi 71% of tweets have positive sentiment and around 7% of negative tweets as well as 22% of neural.

And second bar graph is for @RahulGandhi tagged tweets. And it is showing that 23% of tweets have negative sentiment for Rahul Gandhi and 58% of tweets have positive sentiment and 19% neutral tweets as well.

So by this bar plot and analysis we can say that Narendra Modi has positive impact on people that's why are less negative sentiment tweets are tweeted. But for Rahul Gandhi negative sentiment tweets are three times of Narendra Modi negative sentiment.

5.2.4 Party Tweets analysis: In this part of analysis we are using tweets from BJP and INC twitter timeline. We use tweets only to extract the #hashtag and then using these hashtag we are extracting tweets and doing sentiment analysis on it to get public opinion. Top 5 #hashtags used by BJP and INC are given following also trending some top 5 hashtags.

#hashtag Used by BJP	#hashtag Used by INC	Trending
#SarkaraBadalisiBJPGellisi	#KarnatakaWithCongress	#KarnatakaElection2018
#BJP4BetterKarnataka	#INC4Karnataka	#KarnatakaVerdict
#BJP4InclusiveVikasa	#CongressMathomme	#KarnatakaElection
#KarnatakaElection	#JanaAashirwadaYatre	##SarkaraBadalisiBJPGellisi
#KarnatakaTrustsModi	#congressfornavakarnataka	#KarnatakaElectionResult

By doing the analysis on the tweets which contain above given hashtag, we used almost one lacks tweets and by analysis we found that the tweets which contain hashtag generated by BJP or Narendra modi are have greater positive sentence and and the tweets which contains hashtag which were generated by congress or by Rahul Gandhi have greater negative words. And we also found that tweets for BJP have greater favorite count. So we can conclude that the BJP have good or positive image in public or public have greater positive sentiment for BJP.

5.3 Model to predict tweet sentiment: To predict the sentiment of the tweets I have used naïve bayes and support vector machine learning algorithms.

5.3.1. Naïve Bayes: In this model we have used 210 tweets for train case and for test case we used 90 tweets. The accuracy of this model is 87%. We used small data set because naïve bayes algorithm can also work well on small data set. If we increase the dataset size the accuracy will increase as well.


```

> system.time( NB_classifier <- naiveBayes(trainNB, df.train$class, laplace = 1) )
user system elapsed
0.42 0.00 0.53
> system.time( NB_pred <- predict(NB_classifier, newdata=testNB) )
user system elapsed
1.73 0.00 1.78
> #table("Predictions"= NB_pred, "Actual" = df.test$class )
> NB_conf.mat <- confusionMatrix(NB_pred, df.test$class)
>
> NB_conf.mat
Confusion Matrix and Statistics

          Reference
Prediction Neg Pos
Neg      16   4
Pos       0  11

              Accuracy : 0.871
              95% CI   : (0.7017, 0.9637)
              No Information Rate : 0.5161
              P-Value [Acc > NIR] : 3.545e-05

              Kappa : 0.7395
              Mcnemar's Test P-Value : 0.1336

              Sensitivity : 1.0000
              Specificity : 0.7333
              Pos Pred Value : 0.8000
              Neg Pred Value : 1.0000
              Prevalence : 0.5161
              Detection Rate : 0.5161
              Detection Prevalence : 0.6452
              Balanced Accuracy : 0.8667

              'Positive' Class : Neg

> #conf.mat$byClass
> #conf.mat$overall
> NB_conf.mat$overall['Accuracy']
Accuracy
0.8709677

```

5.3.2 Support Vector Machine: To train this model we used the same data set which were used in naïve bayes algorithm. And this algorithm produce 71% accuracy. SVM work well on large data set. So we can get better accuracy rate by increasing the data set size. As for now the Naïve bayes algorithms produce better predicitive model. This predictive model can work on any dataset which have two attribulte text and class(categorie).

```

> test_SVM<-as.data.frame(test_SVM)
> system.time( SVM_classifier <- svm(class~.,data = train_SVM) )
user system elapsed
1.84 0.05 1.91
> system.time( SVM_pred <- predict(SVM_classifier, na.omit(test_SVM)) )
user system elapsed
0.85 0.08 1.14
> SVM_conf.mat <- confusionMatrix(SVM_pred, test_SVM$class,positive = "Pos")
> SVM_conf.mat
Confusion Matrix and Statistics

          Reference
Prediction Neg Pos
Neg      25  13
Pos       5  17

              Accuracy : 0.7
              95% CI   : (0.5679, 0.8115)
              No Information Rate : 0.5
              P-Value [Acc > NIR] : 0.001335

              Kappa : 0.4
              Mcnemar's Test P-Value : 0.098960

              Sensitivity : 0.5667
              Specificity : 0.8333
              Pos Pred Value : 0.7727
              Neg Pred Value : 0.6579
              Prevalence : 0.5000
              Detection Rate : 0.2833
              Detection Prevalence : 0.3667
              Balanced Accuracy : 0.7000

              'Positive' Class : Pos

> SVM_conf.mat$overall['Accuracy']
Accuracy
0.7

```

Chapter 6: Conclusion & Future Work

In this thesis project we analyze the sentiment on speeches of leaders and on tweets twitted by leader, parties, and people by using natural language process. We make model to predict the sentiment using machine learning algorithms like naïve bayes and support vector machine.

6.1 Conclusion: In this project we have taken two top leader speeches of BJP and INC respectively. Ana the result are showing that there is no big difference in emotions words used in their speeches. Yes there are difference in their speeches and one of the major reasons for difference is one leader is from the ruling party and on from opposition. So that's obvious ruling party leader will appreciate his party in his speech and that's will increase the score of positive, joy, trust etc sentiment and for apposition the score of anger, sadness, negative etc increase. This project will help to find out the overall sentiment of all speeches during the election campaigning by using sentiment score and world cloud. And we can find out what was the main agenda of the leader during the election campaigning.

Huge number of user like to share his or her feeling on social networks site such as facebook twitter etc, And this is emerging as a effective place for exploring, tracking public sentiment. Social networking or social media is one of the largest platforms where huge amount of instant messages are publish each day which is making it an best ideal source of capturing the opinion of public on various topics. Through this project work, I have represented an model for doing the text mining and sentiment analysis of tweets which are politically motivated. We have collected the tweets by using different resources. And as per the analysis result we found that Narendra Modi is far away to Rahul Gandhi in terms of no of tweets, positive sentiment tweets, favorite count on tweets, and retweets count. And Narendra mod has only 6% negative sentiment out of 3200 and Rahul Gandhi has 23%. And this huge difference is just because the Narendra Modi currently ruling the country and Rahul Gandhi currently in opposition so that one of the major reason of negative sentiment because he usually criticize the government policies.

We also try to find the people opinion on Narendra Modi and Rahul Gandhi by analyzing the tweets which have the string @narendramodi and @RahulGandhi (Twitter user name). the result show that around 71% opinion on Narendra Modi is positive and for Rahul Gandhi 58% positive

tweets. But in the negative sentiment there is huge difference between these two leaders for Narendra Modi negative sentiments are 7% and for Rahul Gandhi it is 23%. So we can clearly say that on twitter Narendra Modi has better opinion in public mind. And one more result suggest that at time of election the retweets and favorite count is increasing rapidly for both the leaders so we can assume that people actively sharing the tweets of leaders during the election campaigning. That can be see as they are promoting their leader on social media. This sentiment analysis result are suggesting that Narendra Modi and his party have better positive opinion in public as compare to Rahul Gandhi and his party. And the result of election show the same result as well

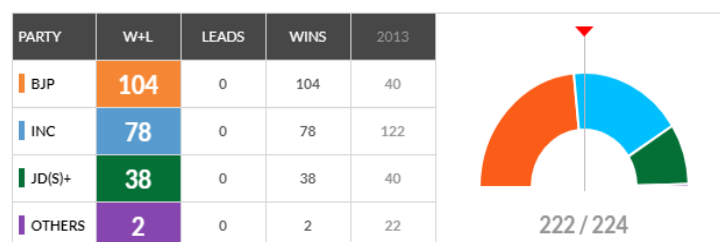


Fig: Karnataka assembly election result 2018

6.2 Future work: opinion mining is one of the most effective ways to understand public opinion. In this project we had done sentiment analysis on leader speeches. And to understand their speeches impact on public we had used tweets. To make this process of analysis we can improve the accuracy by doing following steps.

- **Spam Detection:** As in the second part we are using tweets for analysis. If we include and process the tweets through the spam detection method we can have more accurate sentiment analysis result.
- **More Data of Speeches:** In this analysis we have only used 11 speeches of each leader. And if we have more accurate data then we can have better results.
- **Include other social media data:** We can also have data from other social media. Because in this analysis we only use data from twitter.
- **Language Independent Technique:** Improving proposed technique to work on multi-languages in the scientific domain.
- **Proposing an additional evaluation:** Proposing additional evaluation criteria to categorize the leaders on the basis of their speeches. We can implement this by using the sentiment score present in their speeches

References

https://en.wikipedia.org/wiki/Sentiment_analysis

<https://cran.r-project.org/>

Nath Banamali, "Goods and services tax: A milestone in Indian economy", *IJAR* 3.3, pp. 699-702, 2017.

Sourav Das, Anup Kumar Kolya. "Sense GST: Text mining & sentiment analysis of GST tweets by Naive Bayes algorithm", 2017 Third International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN), 2017

Farid, Dewan Md., Li Zhang, Chowdhury Mofizur Rahman, M.A. Hossain, and Rebecca Strachan. "Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks", *Expert Systems with Applications*, 2013.

Shahnawaz, Parmanand Astya. "Sentiment analysis: Approaches and open issues", 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017

Harshali P. Patil, Mohammad Atique. "Sentiment Analysis for Social Media: A Survey", 2015 2nd International Conference on Information Science and Security (ICISS), 2015

Fuji Ren, Ye Wu, "Predicting User-Topic Opinions in Twitter with Social and Topical Context", *IEEE Trans. on affective computing*, vol. 4, no. 4, October-December 2013.

Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao Li, "Dual Sentiment Analysis: Considering Two Sides of One Review", *IEEE Trans. on Knowledge and Data Engineering*, 2015

Xin Chen, Mihaela Vorvoreanu, and Krishna Madhavan, "Mining Social Media Data for Understanding Students' Learning Experiences" *IEEE trans. on learning technologies*, vol. 7, no. 3, July- September 2014.

Danushka Bollegala, David Weir, and John Carroll, "Cross-Domain Sentiment Classification Using a Sentiment Sensitive Thesaurus", *IEEE trans. on knowledge and data engineering*, vol. 25, no. 8, August 2013.

Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, "A Practical Guide to Support Vector Classification", <http://www.csie.ntu.edu.tw>, web, July 22 2014 .

Shulong Tan, Yang Li, Huan Sun, Ziyu Guan, Xifeng Yan, Jiajun Bu, Chun Chen, and Xiaofei He,” Interpreting the Public Sentiment Variations on Twitter”, *IEEE trans. on knowledge and data engineering*, vol. 26, no. 5, May 2014.

Chenghua Lin, Yulan He, Richard Everson, Member, IEEE, and Stefan Ru”ger,” Weakly Supervised Joint Sentiment-Topic Detection from Text”, *IEEE trans. on knowledge and data engineering*, vol. 24, no. 6, June 2012.

Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, Ming Zhang,” Topic Sentiment Analysis in Twitter: A Graph-based Hashtag Sentiment Classification Approach”, *ACM, CIKM’11*, October 24–28, 2011, Glasgow, Scotland, UK, 2011. Andrius Mudinas, Dell Zhang, Mark Levene,” Combining Lexicon and Learning based Approaches for Concept-Level Sentiment Analysis”,

WISDOM’ 12, August 12, 2012, Beijing, China Copyright 2012, *ACM*. Xiaohui Yu, Yang Liu, Jimmy Xiangji Huang, and Aijun An,,” Mining Online Reviews for Predicting Sales Performance: A Case Study in the Movie Domain”, *IEEE Trans. On knowledge and data engineering*, vol. 24, no. 4, April 2012.

LI Bing, Keith C.C. Chan, Carol OU,” Public Sentiment Analysis in Twitter Data for Prediction of A Company’s Stock Price Movements”, 2014 *IEEE*, 11th International Conference on e-Business Engineering.

Walaa Medhat , Ahmed Hassan , Hoda Korashy,” Sentiment Analysis Algorithms and Applications: A survey”, *Ain Shams Engineering Journal* (2014).

N Cristianini, J Shawe-Taylor, “An Introduction to Support Vector Machines and Other Kernel-based Learning Methods”, Cambridge University Press, 2000 .

Hiroshi Shimodaira, “Text classifying using Naïve Bayes ”, Document models, <http://www.inf.ed.ac.uk/teaching/courses/inf2b/learnnotes/inf2b-learn-note07-2up>, 11 Feb 2014, web, 15 August 2014 .