SENTIMENT ANALYSIS AND INTERPLAY STUDY USING NEWS HEADLINES AND TWEETS

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the award of degree
Of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Balkar Lathwal
M.Tech (CSE)
Enrolment No. 14535009

Under the guidance of
Dr. A.K.Sarje
Emeritus Professor, Dept. of Computer Science & Engineering

COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
INDIAN INSTITUTE OF TECHNOLOGY
ROORKEE – 247667
May -2016

DECLARATION

I hereby declare that the work, which is being presented in the dissertation entitled **“Sentiment analysis and interplay study using news headlines and tweets”** towards the partial fulfillment of the requirement for the award of the degree of **Master of Technology** in **Computer Science and Engineering** submitted in the Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand (India) is an authentic record of my own work carried out during the period from July 2015 to May 2016, under the guidance of **Dr. Anil. K Sarje, Emeritus Professor,** Department of Computer Science and Engineering, IIT Roorkee. The matter presented in this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date:
Place: Roorkee **Balkar**

CERTIFICATE

This is to certify that the above statement made by the candidate is correct to the best of my knowledge and belief.

Date:
Place: Roorkee

(Dr. Anil. K Sarje)
Emeritus Professor
Department of Computer Science and Engineering
IIT Roorkee

ACKNOWLEDGEMENTS

I express my deep gratitude for enthusiasm and valuable suggestions that I got from my guide Emeritus Prof. Anil. K Sarje, Department of Computer Science & Engineering, Indian Institute of Technology Roorkee, for successful completion of this project. His guidance and commitment has inspired me a lot.

I am obliged towards professors and technical staff of Department of Computer Science and Engineering, IIT Roorkee, for helping me in exploring the possibility of carrying out the experiments in their laboratory. Finally, I would like to acknowledge with gratitude, the support and love of my family and friends.

Balkar

ABSTRACT

Every individuals and/or organizations need to know about people's feelings about their product and services provided by them during making any decision. If organizations or individuals have the information about sentiment of people toward the entities of their interest and how sentiment of one entity varies when sentiment of other related entities varies over a period of time, then decision making process can be very easy, efficient and effective.

In this report, after giving a brief introduction of sentiment analysis and related literature survey we discuss about our proposed work, "Sentiment interplay study of entities across various domains using news headlines and tweets". In literature survey section, we discuss about the evolution of the sentiment analysis research area and related work to our proposed work. In our proposed work, we present a model which after collecting entities in current news headlines and tweets related to these entities, performs sentiment analysis and interplay study for these entities. Interplay study means how consistently change in sentiments of one entity triggers a time delayed change in sentiment of other related entities over a period of time. By using this model we can predict the behavior of the related entities by looking at past patterns and can make right decision very effectively.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

1.1 Sentiment analysis

A sentiment can be considered as a quintuple [1], (e, a, oo, h, t), where ‘e’ denotes an entity, ‘a’ denotes an aspect of entity, ‘oo’ denotes the orientation (positive, negative, neutral) of the opinion/sentiment expressed by the opinion holder (h), at time t towards the aspect of an entity. The opinion orientation can be expressed with different strength / intensity levels. We can apply sentiment analysis and opinion mining at different levels to an opinionated text i.e. Document level, sentence/ phrase level.

“Sentiment analysis” is a process of extracting people‘s opinions, appraisal, attitudes, and emotions toward entities, individuals, issues, events, topics. This can be done by applying techniques of machine learning, natural language processing and data mining on social media data (e.g., Twitter, Facebook, and Blogs etc.). Information about sentiments of people toward anything is very useful because it helps in decision making process.

These day people are pouring their opinions about almost anything on social sites in form of facebook posts, tweets, blogs etc. News also contains subjective data about many entities. So, sentiment analysis using social media data is very interesting area of research due to its usefulness and efficiency because a lots of text data can be analyzed in very less time with very good accuracy which is not possible manually.

1.2 Sentiment Interplay between entities

Lots of research has been done for the problem of finding polarity in a piece of text i.e. positive, negative or neutral. Most studies [2] in this area are limited to the identification of sentiments and do not investigate the interplay between sentiments i.e. how change in the sentiment of one entity is affecting the sentiments of other related entities. A predictive model which can tell how entities are related to each other can help in decision taking. In this report a model is proposed which performs sentiment analysis over a time period on various news entities and performs interplay study between those entities. To get sentiments of people over period of time we used tweets and news headlines.

There is very natural motivation behind this proposed model. Things are related to each other and change in one affects others also. For example, if sentiment toward a country is negative due to any reason then foreign investment will drop. If sentiments towards a leader changes sharply then it will affect the sentiments of things related to it like party, persons, brands etc. "*Sentiment interplay study of entities across using news headlines and tweets"*, as name suggests in this we collected news entities appearing in news headlines and tweets related to these entities over a time period. After collecting data we performed sentiment analysis to get polarity score for each entity according to timeline then we tried to capture interplay between entities i.e. How change in the sentiment of one entity trigger a change in other entity/entities.

1.3 Applications

[3]Sentiment analysis is very useful when one want to know other's opinion towards anything about which decision is to be made. Some of them are following:

a) For safe decisions, "Which bank should I invest in?", "Which hotel should I go for??"

b) Opinions at present can help in predicting future events for example opinions of voters can predict election results and market trends can be guessed based on sentiments appearing in news, blogs, tweets etc.

c) Business's always needs public or customer's opinions about the products or services provided by them i.e. why they aren't purchasing their products?

d) For recommendation systems based on the sentiments.

There can be many more applications because everyone wants to listen others before taking decision and lots of data is available in form of text i.e. facebook, posts, tweets, blogs and news which is easily available.

If any individual or organization wants to monitor what are the entities which are related to their product or services and affects them directly or indirectly then they need to perform sentiment interplay between these entities to find correlation between them i.e. how change in the sentiment of one entity trigger a change in other entities. According to authors in [4] having correlation patterns between entities known in advance can be used in data analysis in many application like simulation, impact analysis, forecasting of results etc.

1.4 Challenges [1]

Sentiment analysis deals with the subjectivity i.e. Emotions which is very easy for humans to deal with but computer science finds it challenging things and sometimes fails in accuracy due to following challenges:

Languages complexity:

Any human language is very vast and complex. A sentence may be ambiguous and may contain implicit sentiment/opinion in it which is very difficult for a machine to detect precisely. For example consider this sentence “how can anyone sit through this movie”.

Domain Dependency

In human languages same sentence can have different sentiment/opinion of different domain. Many opinion words can have positive sentiment in one domain and negative in other. For example, consider these two sentences “*The story was unpredictable” and “The steering of the car is unpredictable.”*

Entity Identification:

In sentiment analysis, one or more entities must be targeted because a text may have many entities in it. So identifying entities of our interest is also very challenging task. For example consider following sentences: “*Samsung is better than Nokia”*. “*Ram defeated Hari in football”.* Here opinions about Samsung and Ram are positive but they are negative about Nokia and Hari.

Detecting Spam and Fake opinion:

While analyzing some data for opinion mining, one must take care of the fact that the data can be spam or fake. Because of spam and fake opinions, we can get biased summary. Some agents of a company may put false reviews on a product/service.

Negation:

A single misinterpreted negation can reverse the opinion overall. The reason behind it is, negation sometimes expressed in a subtle ways even without the explicitly using a single negative opinion word. For example, “Not only did I like the acting, but also the direction”, the sentiment is unchanged because of “not”. So this type of situations of “not”, “only” and “not only” needs special attention.

Subjectivity Detection:

To improve the performance of sentiment/opinion mining algorithm, It is required to filter out all the objective text i.e. which don't contains opinion in it. For examples: "*I hate love stories"* and *"I do not like the movie because i hate this story*".

The first sentence is an objective sentence having no sentiment but second example it has a sentiment on a movie.

Thwarted Expectations:

Sometimes the author writes reviews in such a way that he express positive sentiment in his review but in last he write something which make whole sentiment negative. Consider this example, *"This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up"*.

1.5 Organization

This section describes how this dissertation is organized in following chapters. In chapter 2, Literature survey and related work described. In this chapter, we present a brief summary of literature related to sentiment analysis followed by the research work which is related to our research work. In chapter 3, proposed work i.e. statement, definition and motivation behind the proposed work is explained. In this chapter, we clearly defined what actually this work is. In chapter 4, all steps followed to implement this project and results are described. In each stage corresponding algorithm and result is shown. Chapter 5 presents the conclusion of this dissertation. It also tells about the things that can be added to make this work more efficient and useful.

CHAPTER 2. LITERATURE SURVEY AND RELATED WORK

In this section, we present a brief summary of literature related to sentiment analysis followed by the research work which is related to our proposed work. In literature survey, we discuss how sentiment analysis related research evolved and what are the various techniques to perform sentiment analysis. In related work to our research work we mention papers which are intersecting with proposed work.

2.1 A brief Literature survey of sentiment analysis

With the huge growth of social media (blogs, social networks, reviews etc.) on the web, there is lots of data present in form of opinionated text which can be mined and used in decision making processes by organizations. Sentiment analysis is the process of classifying a given piece of text into 3 classes of polarity i.e. positive, negative and neutral. A score also can be given which indicates degree of polarity. Topic based text classification has been very old and well settled research area but text classification based on sentiment is relatively young but trending research area.

According to Bo pang et al in [4] most of the work has focused on ***"topic based classification"*** (e.g., sports, political etc.).But apart from the topic of an article, there is always an emotions or sentiments about entities are also present in it and sentiments can be very crucial factor.

According to authors in [4], in 1998 Biber et al Presented a ***"non-topic based text categorization system"***, which classified documents according to their source(e.g., Which author, which publisher etc) with statistically-detected stylistic variation serving as an important cue.(e.g., The New York Times vs The daily News). In 2001, Wiebe et al attempted to find features which can indicate if some subjective language has been used in an article. These features can be used to to categories a text on the *basis of genre.(subjective or objective)*.In this they used the idea of a classifier which can classify subjective and objective data by using subjective nouns which were learned by using bootstrapping algorithms and they achieved 77% precision. But they do not explicitly address the task of determining what the opinion actually is. They just separate text documents into to class objective and subjective.

Bo pang et al applied ***“supervised learning methods [4]”*** (e.g. naïve Bayesian classification, Support vector machines (SVM). They used these techniques to classify the polarities of movie reviews. They proved that standard supervised machine learning approaches (e.g. Naïve Bayes, support vector machine etc.) performs better than baseline produced by human. They used standard bag-of-feature approach. According to them, “m” features (e.g. word as feature, “still” or the bigram, “really sinks”) can be there in any document and a document can be denoted by a vector “d_v:= (n_1(d), n_2(d),…… n_x(d)) ,where n_j(d) denotes how many times ${}_j$th feature appeared in document”. According to their Experimental results, they applied all the combination of unigrams, bigrams, POS and adjectives with term presence and term frequencies. They found SVM by using unigram + bigram + term presence gives best result.

Most of the previous research on sentiment analysis has been at least partially knowledge based. Some of this work focuses on classifying the semantic orientation of individual words or phrases, using linguistic heuristics or a pre-selected set of seed words. In 2002, Turney and Littman in [5] presented an algorithm which determines semantic orientation i.e. ***“unsupervised learning of semantic orientation”***_from very big volume of corpora. In this method they are using a Web search engine and a point wise mutual information (PMI) function to find the semantic orientation. In the training corpus that they used for evaluating the result of algorithm contains nearly 100 billion words.

They performed classification of reviews using unsupervised learning technique based on the mutual information between document phrases and words “excellent” and “poor”.

$$PMI(term_1, term_2) = \log_2\left(\frac{Pr(term_1 \wedge term_2)}{Pr(term_1) \cdot Pr(term_2)}\right)$$

Pr($term_1$ ∧ $term_2$) denotes the probability of occurring of term1 with $term_2$ and Pr($term_1$)*Pr($term_2$) is the probability of co-occurrence when both the terms are not depending statistically. The ratio “PMI” tells about the statistical dependence between two terms. The semantic orientation (SO) tells about the opinion of phrase and computed by considering the fact that how much associated with the some positive word for e.g. “excellent” and with some negative word for e.g. “poor”:

SO (phrase) = PMI (phrase, “excellent) – PMI (phrase, “poor)

In 2004, [6] Hu, Minqing, and Bing Liu authors have explained how we can get sentiment analysis at aspect level and how to summarize customer reviews. They described a three- step method for extracting aspect expressions (1) mining product features (aspects) on which customers have commented; (2) filtering opinion sentences in a review and finding polarity of each opinion sentence i.e. "positive, negative"; (3) make a summary of all the results to give resultant opinion. According to this paper, all aspects are nouns and noun phrases and opinion words are adjectives. They have mainly used unsupervised learning approach for sentiment classification. To find frequent features or aspects they used POS-tagging to find frequent product features. To find opinion words are mainly adjectives and aspects are mainly noun and noun phrases.

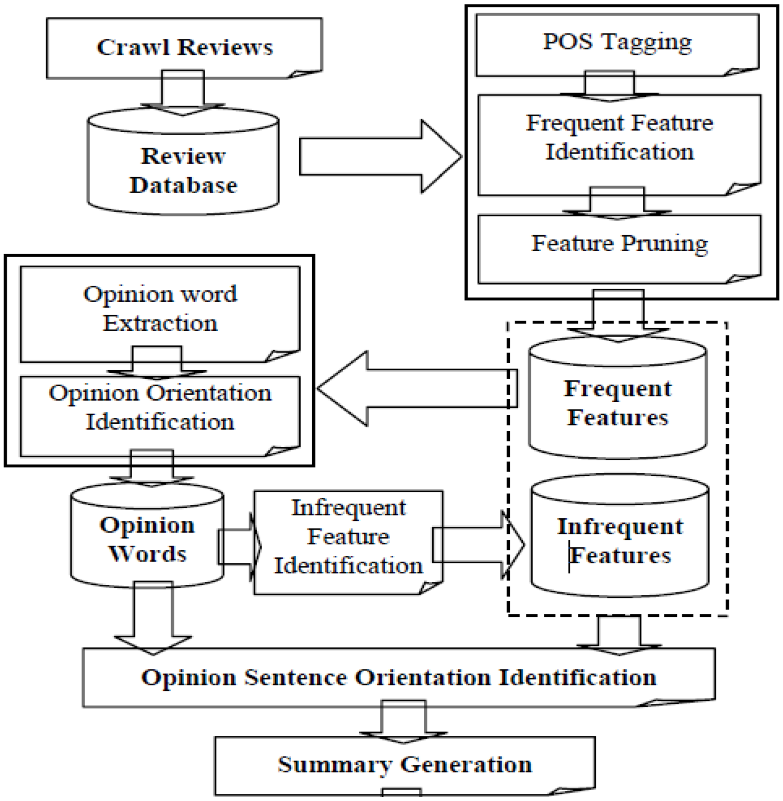

Fig: - 2.1 working model summarizing customer reviews

In [6] this paper, Bing liu et al used ***"lexicon/dictionary based approach"*** for mining and summarizing customer reviews. In their approach for expanding their opinion lexicon, they used bootstrapping which starts with small set of seed words with known opinion/sentiment and grow this set with the help of online dictionary ("e.g., Word Net or thesaurus"). In order to grow opinion lexicon, they used word Net or thesaurus (online dictionaries) by making use of synonyms and antonyms. Words which are newly discovered are added to the seed list and iterative process stops when there are no any new words/phrases are found.

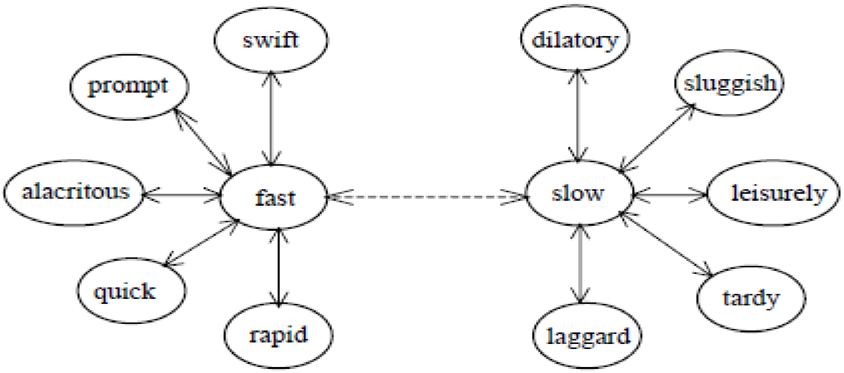

Fig: - 2.2 Growing Opinion Lexicon

Domain and context specific words are difficult to handle here. In this example, "for a speaker of phone", being quite sounds negative. But on the case of a car, "if it is quiet, it is positive".

More research work related to sentiment analysis is discussed in related work because they are helpful and somewhat related to our proposed model.

2.2 Related work to proposed work

In our research work, we are using twitter data and news headlines. For sentiment analysis and polarity score, we used lexicon based methods i.e. dictionaries. For getting sentiments of people we are using tweets and for getting current entities we are using news headlines. Following is the research work which is related to our work and have some intersection either in term of data i.e. tweets or in the way sentiment analysis is performed i.e. lexicon based.

In **[**7], authors extracted data related to political domain from web logs over a two year period. They used Naïve Bayes classifier and SVM classifier to predicate the sentiment/opinion of political blog posts. They targeted only political domain. But they have not studied the interplay between sentiments of entities related to their respective domain. In [9**]** paper they concentrated only on financial data extracted from bloggers towards companies and their stocks. In this, they prepared a corpus of financial blogs and by using text extraction techniques to generate topic specific sub-documents. Then they applied document based sentiment classification techniques on those topic based sub-documents.

[8] In this paper, they considered newspapers and blogs as source of data. In newspapers and blogs, there is opinionated data about of news entities (people, places, things). In their model, they assigned scores which indicates positive or negative sentiment/opinion to each distinct entity in their text corpus. In the first phase of their system i.e. "sentiment identification phase", they identified sentiments toward each entity, associate it with respective entity, and in second phase i.e. "sentiment aggregation and scoring phase", they summarized the overall sentiment of all the entities.

In [9] and [10], authors used Twitter, "the most popular microblogging platform" in order to perform sentiment analysis. They used micro-blogging features like emoticons, hash tags etc. in addition to traditional NLP features for sentiment analysis.

In [11], authors present a target-dependent sentiment classification approach using twitter i.e. It takes a target as a query from user. According to query, they collected tweets related to that entity and performed sentiment classification i.e. classification of the sentiments of the tweets as "positive, negative or neutral" .Here query defines the target for sentiment classification. It involve two steps 1) Defining features based on target provided by the query, 2) Collecting tweets about target and performing sentiment classification.

In [12], authors proposed a method for extracting sentiment out of text which is based on lexicon i.e. dictionary method. They used dictionaries which have words tagged with polarity of sentiment and strength of sentiment. They found this method performed very consistently across various domains and for unseen data because words mainly adjectives are tagged with polarity score.

2.3 Research Gap

After going through the many papers related to sentiment analysis, we found a research gap. As mentioned earlier most of the research work in area of sentiment analysis is focused on how to assign a polarity score to a piece of text. Papers mentioned in related work has focused on finding sentiment for individual entities in a single domain but not focused on interplay of sentiment between entities across multiple domains. Most studies in this area are limited to the identification of sentiments and do not investigate the interplay between sentiments i.e. how change in the sentiment of one entity is affecting the sentiments of other related entities in many domains.

However some papers are published on sentiment analysis on various domains individually but they have not focused on the relation between entities according to change in their sentiments. In our approach we are using tweets and news headlines, work which is related or intersects with our research work has been discussed in the section 2.2 i.e. related work.

CHAPTER 3. PROPOSED WORK

3.1 Problem Statement and Definition

“Sentiment analysis and interplay study of entities using news headlines and tweets”

In our proposed work, we want to find how closely two real world entity are sentimentally correlated with each other i.e. how change in sentiments of people towards them over a time is correlated. If change in sentiment of one entity triggers a change in other entity consistently over a period a time then we can say these entities are correlated and there is some sentiment interplay between them. We can collect some sentiment interplay patterns then it can of great use in decision making process.

This is very natural that when something happens to an entity or set of entities then all the related entities also affected. When some event happens with any person, product, or any real world object then sentiment/opinion of people towards it also changes, then how this change in sentiments affects the sentiments of people towards all the related entities. If we have frequent sentiment interplay pattern known in advance then decision making process can be very effective.

For e.g., “when price of patrol increases, prices of many things increase in market” Or “when crimes are increasing in some country and sentiment of people about that country is negative then tourism business and many other domains also suffers” Or “when some political party wins, this event can trigger changes in the sentiment of many related entities in various domain”, “If sentiments toward Aam Aadmi Party (AAP) and kejriwal are positive then it is very natural that people will have positive sentiment towards everything related to AAP or kejriwal.”

Interplay study of entities means when a considerable change in sentiments of an entity is found at time T in a time series then how it is going to affect other related entity/entities i.e. is there any considerable change in the sentiment of related entity at its time series in T+t where t can be any time depending on requirement of experimental setup. Two entities are said to be correlated if they are consistent in their interplay. Following graph shows the 90 days sentiments change of Modi and BJP in the months of feb, march and April day wise.

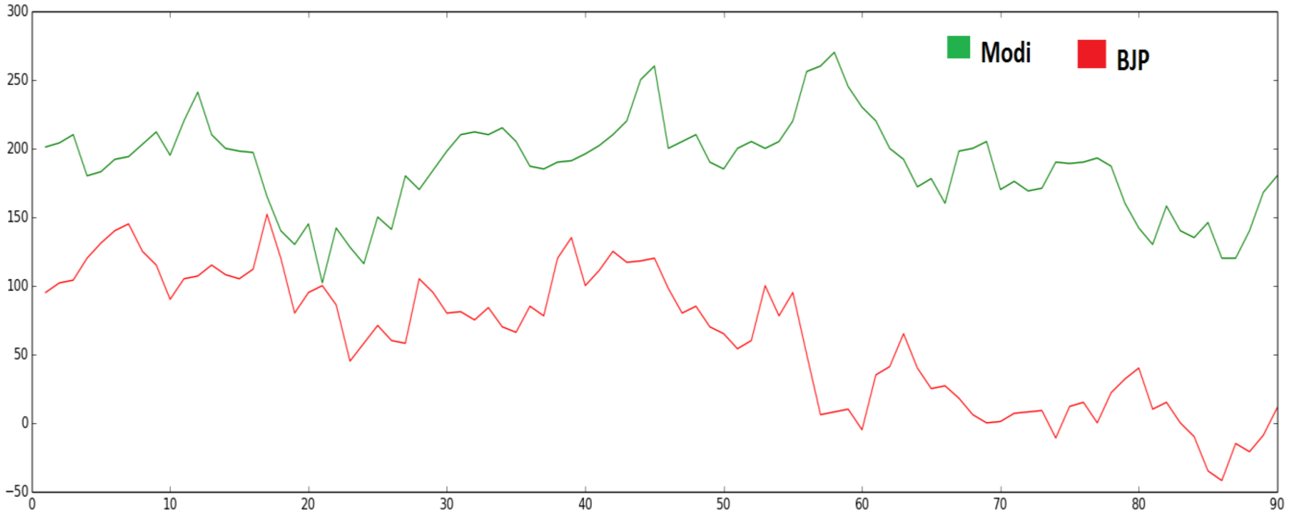

Fig: - 3.1 Sentiment score timeline of Modi and BJP for 90 days

As the above graph shows change in the sentiments of Modi and BJP are correlated in after doing sentiment analysis of tweets in two months.

We represents a model which can collect data from social networks site twitter, can divide and perform sentiment analysis for many entities and monitor the interplay of sentiments across many related entities. Overall design, implementation details and results are discussed in next chapter.

3.2 Motivation

The motivation for this idea is lying in following two facts:

1. There is huge amount of information existing and increasing every moment on online social networks in form of tweets, blogs, news, and forums. People are expressing their opinions and sentiments about almost everything on online social networks sites. By using this data, we can mine and analyze this data for various decision making and prediction purposes.

2. We found a research gap in literature of sentiment analysis area. Lots of research has been done for the problem of assigning score to a piece of text describing its polarity i.e. positive, negative or neutral. Most studies in this area are limited to the identification of sentiments and do not investigate the interplay between sentiments i.e. how change in the sentiment of one entity is affecting the sentiments of other related entities.

CHAPTER 4. IMPLEMENTATION AND RESULTS

In this project, "*Sentiment interplay study of entities across various domains using news headlines and tweets"***,** we are using twitter data and news headline to perform sentiment analysis due to following reasons:

a. Twitter is most popular microblogging social site where people are pouring their opinions about everything. Users express their opinion on products and services, politics and religions. Also [10] Variety of users i.e. "regular users to celebrities, company representatives, politicians, and even country presidents" of twitter makes it very good source of subjective data. So twitter provides texts data from various social groups and people of different interests.

b. News can be either good or bad but it is seldom neutral. News is something which is required by almost everyone. Everyone wants to know what is happening in the world because news headlines give most popular and current events. Organization and individuals always want to know the effect of current news entities on the entities of their interest.

Steps followed for implementing the proposed research idea

Overall process is divided in mainly two phases. In Phase-1 i.e. Data collection and preprocessing, news headlines are collected and tweets are collected for the entities appearing in these news headlines collected. In Phase-2 i.e. Sentiment analysis and Interplay study, for each tweet collected for a entity sentiment score is assigned depending on the kind of polarity it carries and after aggregating sentiment score to some level i.e. Hours, Days, Months etc sentiment interplay study is carried.

Following figure illustrates the overall process:

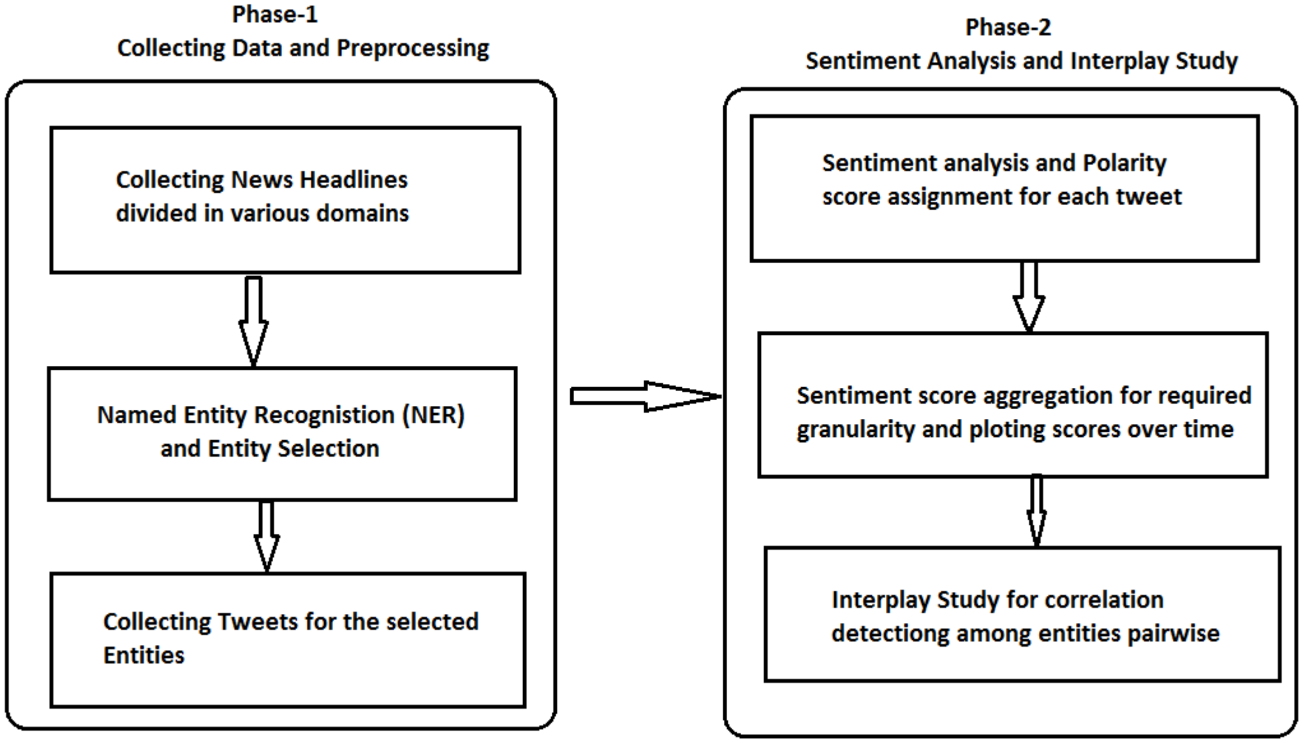
Phase-1
Collecting Data and Preprocessing
Collecting News Headlines divided in various domains
Named Entity Recognistion (NER) and Entity Selection
Collecting Tweets for the selected Entities
Phase-2
Sentiment Analysis and Interplay Study
Sentiment analysis and Polarity score assignment for each tweet
Sentiment score aggregation for required granularity and ploting scores over time
Interplay Study for correlation detectiong among entities pairwise

Fig: - 4.1 Overall system design for implementation

Two phase implementation method have followed steps which are listed below and described in next section:

1. Collecting news headlines classified into respective domain.
2. Finding entity/entities in each headline and selecting entities of interest among all.
3. Collecting tweets about each entity retrieved from the step 2.
4. Assigning sentiment polarity score according to the sentiment inside each tweet.
5. Aggregation of sentiment scores day wise or any granularity of user’s choice and Building a timeline graph i.e. time Vs. sentiment for each entity
6. Performing interplay study between entities and finding relationship between entities pairwise i.e. how much correlated two entities are.

4.1. Collecting news headlines and classifying them into respective domain

News headlines can be collected many ways. But we followed a smart approach to collect news headlines using twitter API Tweepy by using python programming language. Since we don’t need whole news article but only news headline in order to get current entities. After studying this API Tweepy we learnt that we can get the public timeline of a user and all the leading newspapers have their account on twitter on which they keep on updating news headlines. Also these news groups have their account for every domain (e.g., Sports, economy etc). By using the features of Tweepy we are able to get enough news headlines and get them divided in various domain with 100% accuracy. Table 1 shows news headline related to different domains.

General /Political	Economy/Market	Sport	Weather
Getting into live-in relationship is no crime:SC	Recovery! Rupee ends tad strong on Friday	avi Shastri to return as team director for tour of Sri Lanka	HEAVY widespread rain for Gujarat (many zones),U.P…..
“PM @modi launches country's first solar powered e-boats”	ensex crashes 400 pts, Nifty below 8400; metals bleed	India seal historic quarterfinals berth at @FIBA Asia	Moderate rain can push into city around 1pm.:: Chennai
Black flags shown to Kejriwal at Patna airport	“Stock markets pare early gains, Infosys up over 9 per cent”	@Venuseswilliams wins 700th match of career	Heavy rain will be along W,S-W,S Gujarat.
“VVIP chopper scam: Ex-deputy Air chief questioned for 8 hours”	Closing bell: #Sensex flat on weak global cues; #Nifty50 slips below 7,850; Airtel down 2%, ICICI 1%	@boxervijender beats Matiouze byknockout to register his fifth consecutive win in pro career:	Scattered moderate T showers will pop along W-ghats Kerala, Tamilnadu, S,central Kerala

Table 4.1: News headlines classified with respective domain

4.2. Finding named entity/entities in each headline and selecting entities of interest among all.

Named entity recognition (NER) is well known research area in Information Extraction (IR). NER is a process of labeling a word or phrases with names of persons, organizations, locations, expressions of times. In our project we are using “Stanford NER version 3.5.2”. Stanford NER is a Java implementation of a Named Entity Recognizer. But when we apply it on

Tweets, due to informal structure of tweets sometime it fails to detect named entity. So overall accuracy it provides is around 80 %. We pass each news headline tweeted by a newsgroup to the Stanford NER and it appends named entity found in that news headline at the end of the headline. Table 2 is a sample of output produced in this step.

News Headline	Named entities
Bharti Infratel reports 24% jump in profit to Rs 576 crore	Bharti Infratel
Delhi govt has sent the file pertaining to appointment of Swati Maliwal as DCW chief to LG: Deputy CM Manish Sisodia	Delhi ,Swati Maliwal ,Manish Sisodia
Give @narendramodi more time: Premji @narendramodi Premji	@narendramodi, Premji
Privilege notice against Robert Vadra over FB post against parliamentarians	Robert Vadra
It is a shame that even after 68 years of Independence, we're having debate over chanting of 'Bharat Mata ki Jai': BJP MP Gopal Shetty (ANI)	BJP, Gopal Shetty
PM has himself requested that this matter be investigated: @arunjaitley on #panamapapers (Pic: ANI)	#panamapapers, @arunjaitley
Congress undermined Antony’s concerns on Agusta tests: BJP	Congress, Agusta, BJP
@ArvindKejriwal appeals to @narendramodi, says 'don't be stubborn'	@ArvindKejriwal @narendramodi

Table 4.2: Named entities found in news headlines

Selecting entity set based on the connectivity

After passing all the tweets collected to “Named Entity Recognizer”, we get lots of entities and not all are useful and of interest. So to get only selected entities we followed a connectivity graph based approach. Two entities are said to be connected if they are appearing in the same news headline i.e. there exist a un-directional edge between these entities. By following this method we get a graph where vertex denotes an entity and an edge denotes a connection between two entities.

For example consider following sample entity sets i.e. a set contains one or more entities appearing in a single news headline i.e. (X, Y) means X←→ Y is an edge in graph.

(#OddEven, Delhi), (BJP, Gopal Shetty) (#Congress, #vadodara, @narendramodi), (Congress, Agusta BJP), (#OddEven, Arvind Kejriwal), (#panamapapers, @arunjaitley), (#panamapapers, @SrBachchan Aishwarya), (@ArvindKejriwal, @narendramodi, BJP), (Delhi, Swati Maliwal Manish Sisodia), (BJP, Himachal Pradesh, CMVirbhadra)

(Robert Vadra, Congress), (#panamapapers, @SrBachchan, Aishwarya, Amitabh)

By using the edges mentioned above we constructed connectivity graph:

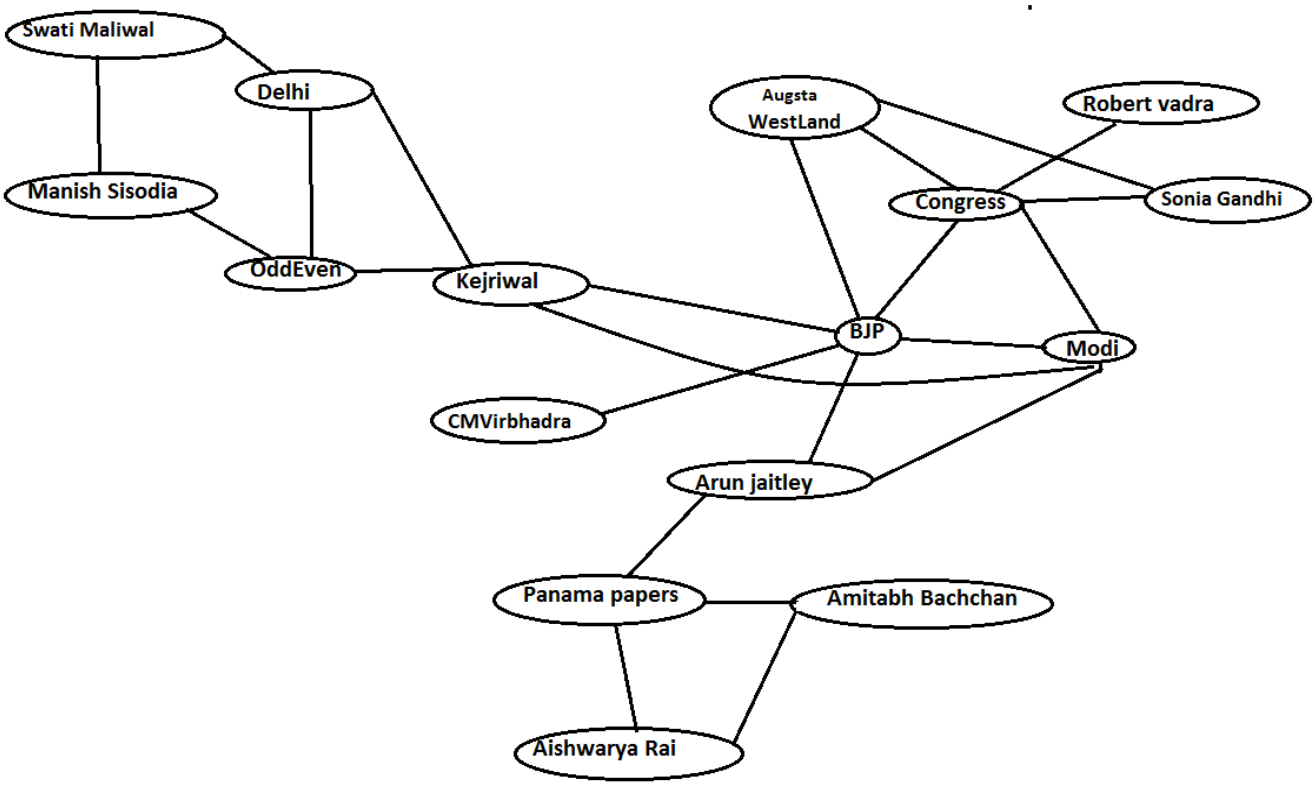

Fig:-4.2 Graph showing connectivity between entities which appears in news headlines

After making connectivity graph, we can get retrieve of entities for which we want to do interplay study. We performed sentiment analysis and interplay study on the sets which contains entities which are directly connected. If there are 'N' entities in connectivity graph then there are

‘N’ sets possible where in each set one entity is head and other are entities which are directly connected to head. We want to study if there is any sentiment interplay between head entity and rest of entities in each set. Following are some example sets From above graph:

{Modi → BJP, Congress, Arun jaitley, kejriwal }
{Congress → sonia Gandhi, Agustawestland, BJP, Robert Vadra, Modi }
{BJP→ Modi, Arun jaitley, kejriwal, Congress}
{Panama Papers →Arun jaitley, Aishwarya Rai, Amitabh Bachchan } etc.

There can be ‘N’ such sets and user can select any of set which is of interest.

4.3 Collecting tweets about each entity selected in previous step

By using API tweepy in python, we can gather any number of tweets according to keywords. “twitterStream.filter(track=[x],languages=['en'])” is used to collect tweets based on the keyword x (e.g.,Modi,AAP,kejriwal,Dhoni etc). Tweepy gives required tweets with many parameters like tweet-time,tweet text,user id etc. We can set limit on number of tweets and time period.

4.4 Assigning sentiment polarity score according to the sentiment inside each tweet

Earlier to assign polarity score to each tweet collected, we tried a python library “Sentiment_ Classifier 0.6”. In this they are using Word Sense Disambiguation using wordnet and word occurrence statistics from movie review corpus nltk. It classifies into positive and negative categories also assign a score according to polarity of tweet. But it is domain dependent and gives poor performance with negations.

To get a better result we write a new algorithm for assigning sentiment score to a piece of text .We implemented this algorithm in python which is a rule based approach in which we used some NLP techniques and 5 type dictionaries (positive, negative, incrementer, decrementer, inverter) to assign a score to each tweet. Following flow diagram illustrates our algorithm contains 3 parts explained below.

Positive dictionary which contains 2100 positive words, Negative contains 5000 negative words. We also used incrementer dictionary which contains those words which increases the strength of sentiment. Similarly decrementer which decreases and inverter which invert the strength of sentiment. For example “very good” is more positive than “good” because “very” here is sentiment incrementer , similarly “not bad” is positive but if we tag only by using negative words dictionary it would be negative sentiment but due to inverter word “not” is has positive sentiment.

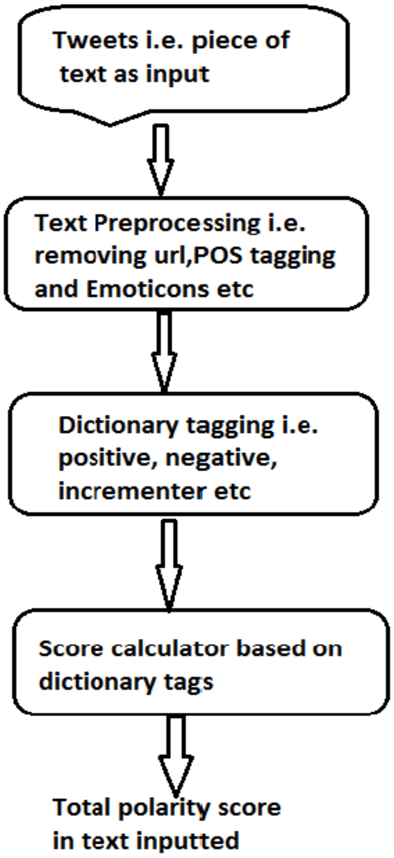
Tweets i.e. piece of text as input
Text Preprocessing i.e. removing url,POS tagging and Emoticons etc
Dictionary tagging i.e. positive, negative, incrementer etc
Score calculator based on dictionary tags
Total polarity score in text inputted

Fig:-4.3 Overview of Sentiment score assignment algorithm

Following steps explains the algorithm that we applied.

a) Text preprocessing
   - i) Split each tweet in sentences if more than one sentences are present.
   - ii) Remove the entire unnecessary thing from each tweet like usernames, URL etc.
   - iii) Split each sentence in tokens and apply POS tagger to tag each token.

b) Dictionary tagging
   - i) Find each taken for an expression i.e. positive, negative by using dictionaries. Since it is twitter data, It usually have emoticons :) :(etc. which denotes the sentiment of twitter.
   - ii) Tag each word which expresses some positive or negative sentiment according the dictionaries. Tweets which have positive emoticon in it then we avoid check it in negative words dictionary and same applies with negative emoticons.
   - iii) We used five type of dictionary discussed above to take care of proper sense of sentiment strength and orientation.

c) Sentiment score calculator
   - i) Measure the sentiment score according to the dictionary tag i.e. if positive then 1 if negative then -1 otherwise 0.
   - ii) Modify scores according to modifier dictionaries i.e. if Incrementer then multiply by 2, if decrementer divide by 2, if invertor or sentiment flipper then multiply score by -1.
   - iii) Return total sentiment score.

Some example tweets and with sentiment polarity score analysed by the above algorithm are shown in table below.

Tweet	Polarity Score
Modi is not delivering all promises he made	-1
Stock markets pare early gains, Infosys up over 9 per cent	2
Heavy rain will be along west Gujarat	0
Getting into live-in relationship is no crime:SC	1
If Modi was not having degree then USA would not have placed Red Carpet to welcome him.	1
RT @CNNnews18: Mallya rejects the idea that PM Modi was behind the decision to issue his arrest warrant and revoke his passport	-3
RT @oldschoolmonk: Plot twist : Modi ji has an engineering degree but he is ashamed to admit. #DegreeDikhaoModiJi	-3
Modi always sent his political opponents behind jail then does it mean that Sonia Gandhi is not his political opponent?: @Ar…	-1
@CNNnews18: Mallya rejects the idea that PM Modi was behind the decision to issue his arrest warrant and revoke his passport	-3
Arvind ji is like open book in front of people, Modi ji should also come out clean about his education	2
RT @HappyAppy83: Modi will lodge Sonia in the same cell where Shiela Dixit is... BTW which jail is Shiela Dixit in??	0
@ArvindKejriwal kitna modi modi karto ho bhai.... He has more important work to do for people of country.....	2

Table 4.3: Polarity score assigned to each tweets related to entities recognised above

4.5 Aggregation of sentiment scores day wise or any granularity of user's choice and Building a timeline graph i.e. time Vs sentiment for each entity

To make analysis purpose easy, fast and accurate, it is very important to convert the fact into some visualization. We choose a Time Vs. Sentiment score in order to analyze at what time what is the sentiment of peoples towards an entity and to monitor how sentiment is changing with time. When we have Time Vs Sentiment graph for all the entities of our interest, we can easy detect patterns among entities. Also tweets are available in huge volume, so it became more important to choose the granularity level in time i.e. hours, days, months or years. In over timeline graph on Y axis we take sentiment score aggregated over a day and on X axis we have days.

For example we collected tweets for the keyword "Modi" for three months February, March and April 800-1000 tweets for each day. After applying sentiment score assignment algorithm and aggregating for each day we get following time-series of sentiments for the Narendra modi, BJP and Congress for 90 days.

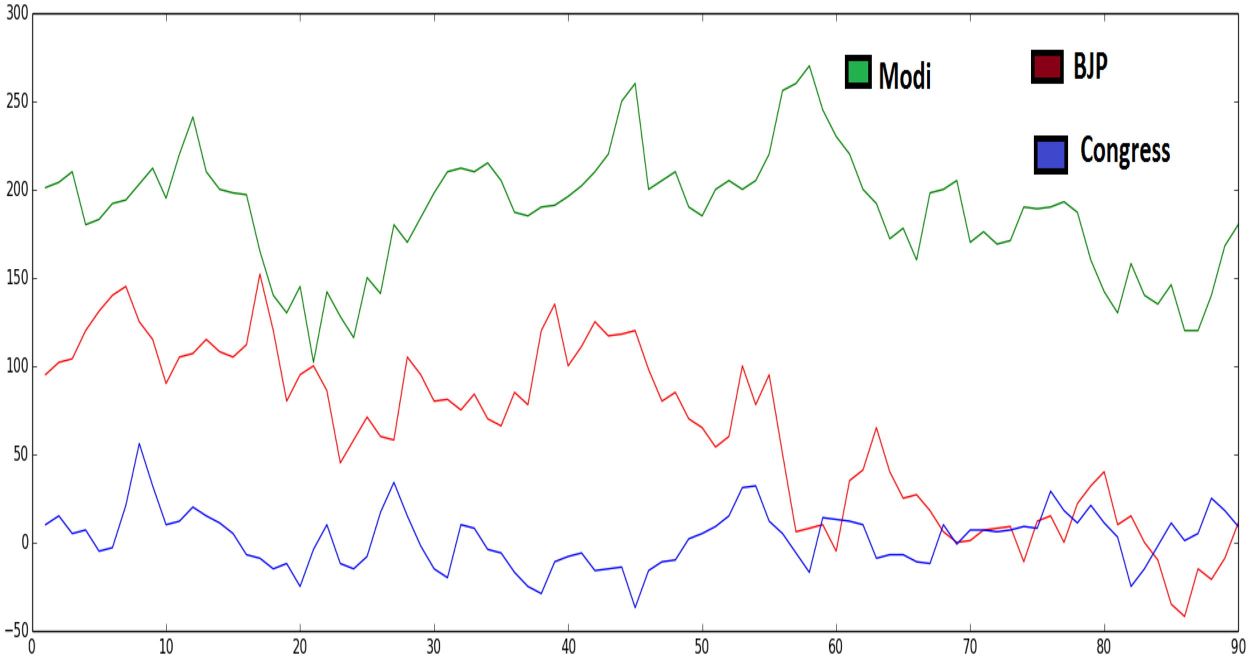

Fig:-4.4 Modi, BJP and congress's sentiments timeline for 90 days.

4.6 Performing interplay study between entities

In Interplay study we tried to find how change in sentiments of people towards entities over a time is correlated. If change in sentiment of one entity triggers a change in other entity consistently over a period of time to find relationship between entities i.e. correlation between two entities. To perform this we used and modify a statistical method [13] to detect sharp changes in time series of every entity of interest and then find how these sharp changes are correlated.

Interplays study requires two steps:

a) Finding change point in a time series of Sentiment score Vs time
b) Performing analysis among various time-series for the change points correlation i.e. if time series of two entities have some correlation regarding their change points.

4.6.1. Finding change points in a time series

As we are analyzing 2D Sentiment score vs. time graph where each data point <s, t > means s sentiment score at time t. [13]A data point is said to be change point if it denotes a sharp and considerable change in sentiment. Let d_1, d_2, d_3…... d_n are n data points in our time series. To find all such change point we are various methods. Initially we started with following cumulative summary based algorithms. This algorithm is explained below:

a) Calculate mean X of all the data points.
b) Initialise Cumulative sum CS_0=0. .
c) For finding the CS_i for all i=1 to n
    i) Do $CS_i = CS_{(i-1)} + d_i$-X
d) CS_i for each point is called as cumulative summary (CUMSUM) for each data point di which can be seen as $\sum_i (d_i\text{-}X)$ where i=1 to n.

After getting cumulative summary for each data point, these [14] CUSUM values need to be compare with two threshold values i.e. upper thresholds, lower thresholds. A data points is said to be a change point if its cumulative summary value (CUMSUM) is either more than upper threshold or less than lower threshold. For deciding these threshold values they are using SD i.e "standard deviation" but in some case it can be two constant values depending on the data and application.

But this CUMSUM algorithm is not very suitable for our model because we are interested in sudden and sharp changes but this is useful in detecting long term [14] general trends. Also in our data sentiment changes in inconsistent manner but this is useful in consistent and gradual changes.

So, in order to find change point in our approach we are using simple and efficient algorithm which is based on mean and standard deviation in sliding windows described following.

In our method, we are using a sliding window to calculate moving mean, SD on time-series data points so that we can have dynamic threshold values to determine if a data point is a change point or not. First we decided a window size, then we calculated mean ‘X_i ‘ and standard deviation “SD_i” for each i_{th} window sliding from left to right. After calculating previous two values for each window we set upper threshold (UT) and lower threshold (LT) where $UT_i = X_i + SD_i$ and $LT_i = X_i - SD_i$. Now we can declare those data point d_i as change point which are either below the LT_i or above the UT_i . Following are some example of sentiment score vs time (Sentiment scores on Y- Axis and time in days on X-Axis) over 90 days (months of February, March and April) time-series having their change point detected by using above algorithm.

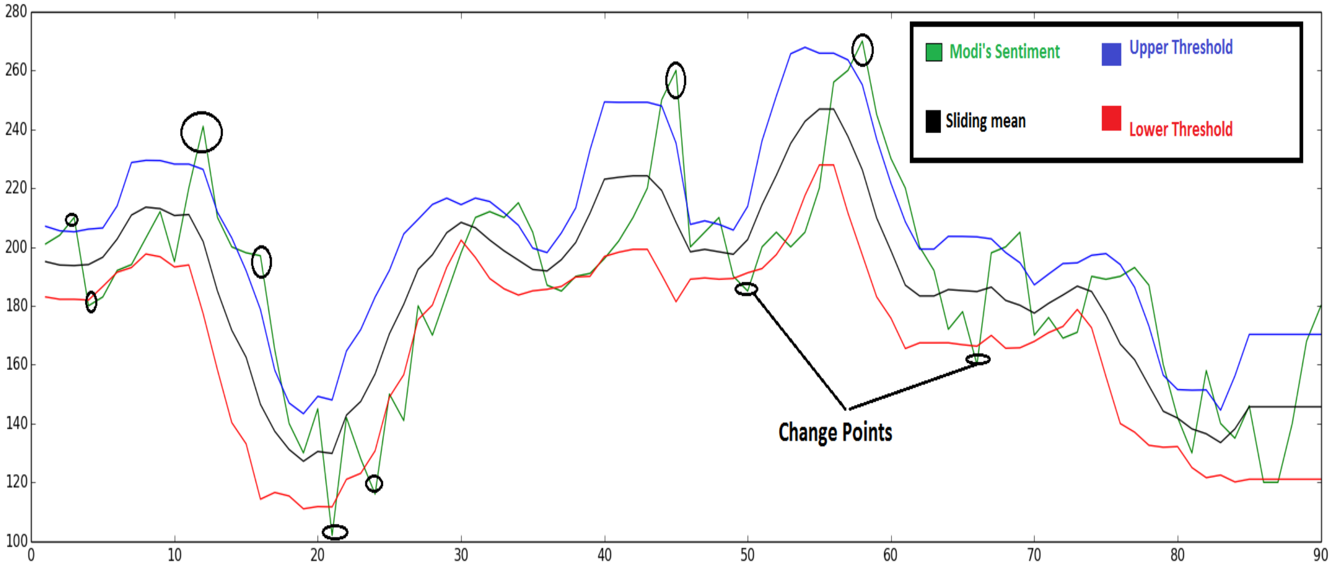

Fig. : 4.5 Modi’^{s} sentiment timeline with moving mean, upper threshold, lower threshold and change points

Here those points are considered as change points which deviate from mean of that window i.e. moving mean. If sentiment score of an entity change suddenly and are far from standard deviation then we considered it as change point.

Following is the sentiment score vs. days time-line for the entity ‘BJP’ with change points detected.

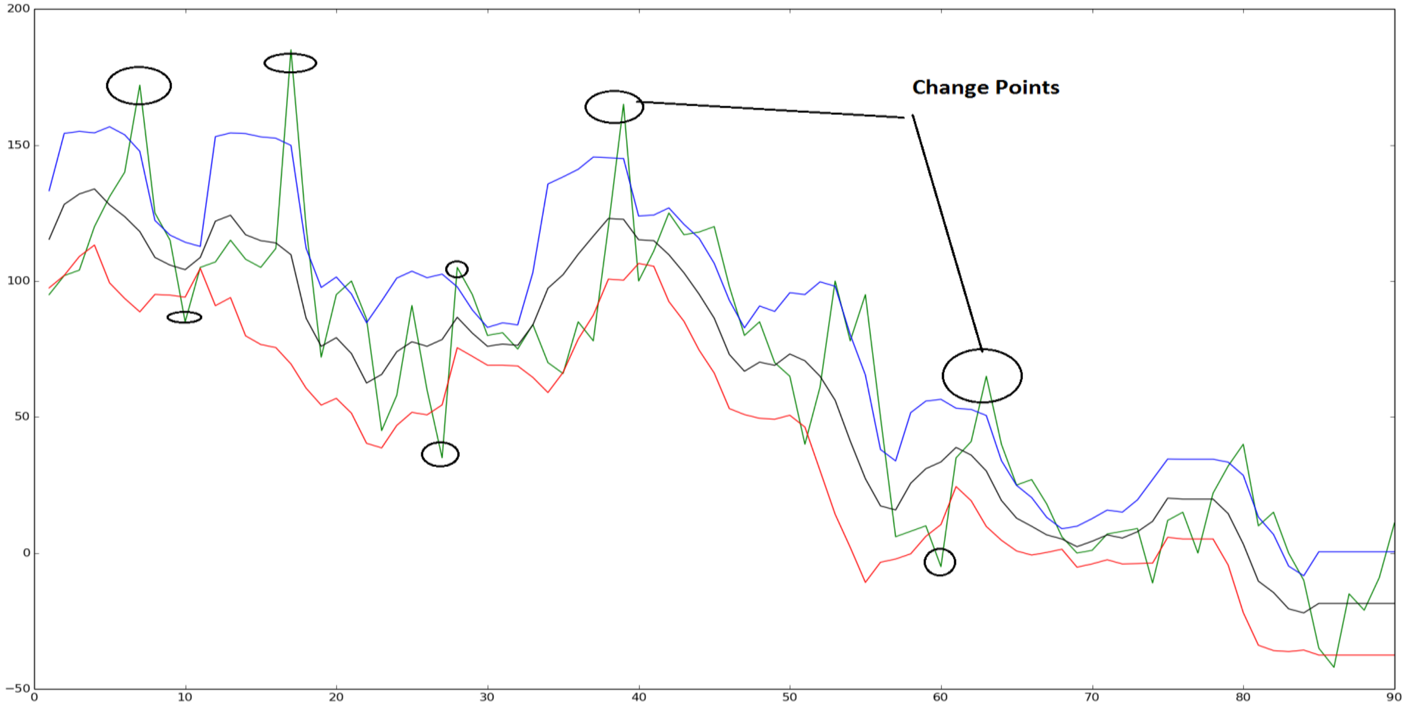

Fig:-4.6 Change points detected in ‘BJP’ timeline of sentiments score vs. time

There are two types of change points in timeline i.e. positive change point and negative change point. Points which are above upper threshold are positive change points and below lower threshold are negative change points.

4.6.2. Performing analysis among various time-series for detecting the change points correlation

After having all the change point detected, now we need to compare two time-series for finding any correlation in their change points. If there is a change point in a time series A affects or controls change points in other time series B consistently then these two entities are said to be correlated. Also degree of consistency determines how closely these entities are correlated to each other. To reflect how much two time series are correlated we define a score. This score if it positive then time series are positively related i.e. when sentiment value increase in one then it increase in other also or if negative then they disagree i.e. if sentiment increase in one then it decrease in other.

After performing all the steps mentioned above on these entities we get their time line of sentiment score vs days with all the change point detected. Now if we want to study the effect of entity X at entity Y i.e. X → Y. To get the final correlation score we followed following algorithm. We maintained two score denoting agreement score and disagreement score. Positive score tells that if a change point detected in X's timeline then same change point is detected in Y's timeline

a) For each change point at time T in X'^{s} timeline, look for a change point in Y'^{s} timeline in a window of size t where size depends on the granularity at which we want to analyze.

b) If a change point detected in Y'^{s} timeline up to the predefined window then we need to test if this change point is agreeing or disagreeing with change point of X'^{s} timeline. If it agrees then add one in agreement score else add one in disagreement score.

c) Now, here agreement score tells how positively or negatively correlated related two entities are.

To make analysis easy we are using a table which keeps records of all the change points like positive or negative, day at which change points detected. Following we are considering a set of related entities as an example and performed interplay study over these entities. The sets we are considering is{Modi, BJP, Congress}. So we perform interplay study pairwise for the the following.

After performing change point detection process we stored this information in tables so that we can analysis their interplay study.

Change points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Day	3	12	16	21	24	26	33	45	51	54	58	66	69	72	78	83
+ / -	+	+	+	-	-	-	+	+	-	-	+	-	+	-	+	+

Table:-4.4 Change point summary for the timeline of 'Modi'

Change points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Day	9	11	17	22	28	30	38	40	41	45	53	59	64	74	80	84
+/ -	+	-	+	+	-	+	-	+	-	+	+	-	+	-	+	+

Table:-4.5 Change point summary for the timeline of 'BJP'

Change points	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Day	7	9	12	20	24	28	31	32	37	45	53	58	63	68	76	80	82	86
+/ -	-	+	+	-	-	+	-	+	-	-	+	-	+	-	+	+	-	+

Table:-4.6 Change point summary for the timeline of 'Congress'

By using the tables above we calculated 2 score i.e. agreement score and disagreement score for pair of entities {Modi, BJP},{Modi, Congress},{BJP, Modi},{BJP, Congress}

In the set {Modi , BJP, Congress}.We can do this analysis for any number of desirable entities.

i) Agreement score which shows how positively correlated these entities are i.e. When there is change point occurred at time T there is change points of same nature detected in the T+t time. Where t is window size which is 5 here.

ii) Disagreement score which shows how negatively correlated these entities are i.e. When there is change point occurred at time T there is change points of opposite nature detected in the T+t time.

Modi→ BJP

There are 16 change points in Modi's timeline. Out of 16, 9 are positive and 7 are -ve change points. Timeline of BJP also detected with same number of change points, positive and negative changepoints. By using the above algorithm and respective tables of Modi and BJP we got following.

a.) Agreement score: In Modi's timeline 9 Change points found a change point in BJP timeline within the window time of same nature. So agreement score is 9/16=0.56.

b.) Disagreement score: In Modi's timeline 5 Change points found a change point in the window time of same nature. So disagreement score is 5/16=0.31.

Modi→ Congress

There are 18 change points in Congress's timeline. Out of 18, 9 are positive and 9 are -ve change points. By using the above algorithm and respective tables of Modi and Congress we got following.

a) Agreement score: In Modi's timeline 6 Change points found a change point in Congress timeline within the window time of same nature. So agreement score is 6/16=0.37

b) Disagreement score: In Modi's timeline 8 Change points found a change point in Congress's timeline within the window time of opposite nature. So disagreement score is 8/16=0.50

<u>BJP → Congress</u>

a) Agreement score: In BJP's timeline 5 Change points found a change point in Congress's timeline within the window time of same nature. So agreement score is 5/16=0.31

b) Disagreement score: In BJP's timeline 10 Change points found a change point in Congress's timeline within the window time of opposite nature. So disagreement score is 10/16=0.62

<u>BJP → Modi</u>

c) Agreement score: In BJP's timeline 5 Change points found a change point in Modi's timeline within the window time of same nature. So agreement score is 7/16=0.43

d) Disagreement score: In BJP's timeline 9 Change points found a change point in Modi's timeline within the window time of opposite nature. So disagreement score is 6/16=0.37

CHAPTER 5. CONCLUSION

In this dissertation, we present a model which performs sentiment analysis and interplay or correlation study among entities appearing in news headlines by using tweets. After giving brief introduction about what sentiment analysis and interplay study is, why it is needed, we discussed literature survey about sentiment analysis and correlation study. In literature survey part, sentiment analysis problem and various methods for Sentiment Analysis, mainly Machine Learning methods and lexicon based methods were discussed.

In Presented research idea, we used News Headlines and Tweets as our data source. To get news headlines we used twitter API tweepy and for people's opinions on entities appearing in news headline we used tweets. We proposed a dictionary based method for sentiment analysis and statistical method to perform sentiment interplay study for the entities appearing in news headlines. We found that it can be very useful if we have enough interplay patterns of entities of our interest then it can help in decision making process. We learned that most of the researches are focused on sentiment analysis only in but not about how sentiment of entity/entities depends on the change in sentiments of the related entities i.e. monitoring the change in the sentiments of one entity due to other entities. So there can many scope of advancement to design more methods and algorithm to perform sentiment interplay study.

In future, we will try to enhance this model of sentiment interplay study to achieve more accurate and efficient methods.

REFERENCES

[1] Liu, Bing, and Lei Zhang. " A survey of opinion mining and sentiment analysis." Book "Mining text data", Page 415-463, Springer US, 2012.

[2] Godbole, Namrata, Manja Srinivasaiah, and Steven Skiena. "Large-Scale Sentiment Analysis for News and Blogs." *Proceedings of the ICWSM* 7 (2007) *conference on International Conference on Weblogs and Social Media.*

[3]Pang, Bo, and Lillian Lee. "Opinion mining and sentiment analysis."*Foundations and trends in information retrieval* Volume 2,1-2 (2008):Page no. 1-135.

[4] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002

[5] Peter, D. "Turney: Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews." *Proceedings of 40th Annual Meeting of the association for Computational Linguistics*. ACL 2002

[6] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews. "*Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

[7] Durant, Kathleen T., and Michael D. Smith. "Mining sentiment classification from political web logs." *Proceedings of Workshop on Web Mining and Web Usage Analysis of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (WebKDD-2006), Philadelphia, PA*. 2006.

[8] O'Hare, Neil, et al. "Topic-dependent sentiment analysis of financial blogs. "*Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*. ACM, 2009.

[9] Pak, Alexander, and Patrick Paroubek. "Twitter as a Corpus for Sentiment Analysis and Opinion Mining." *Proceedings of the* LREC *Conference* on Language Resources and Evaluation. Vol. 10. 2010.

[10] Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. "Twitter sentiment analysis: The good the bad and the omg!." Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (*ICWSM* 11) page 538-541,2011.

[11]Jiang, Long, et al. "Target-dependent twitter sentiment classification. *“Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

[12]Taboada, Maite, et al. "Lexicon-based methods for sentiment analysis." Proceedings of the 37th Volume, Issue 2, of the Association for Computational Linguistics: Pages 267-307 37.2 (2011) June 2011

[13]Sayal, Mehmet. "Detecting time correlations in time-series data streams."*Hewlett-Packard Company* .Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto HPL-2004-103 June 9, 2004.

[14]Taylor, Wayne A. "Change-point analysis: a powerful new tool for detecting changes." *preprint, available as http://www. variation. com/cpa/tech/changepoint. html* (2000).