CHAPTER 2
LITERATURE SURVEY

This chapter presents an in-depth literature review on techniques used in social media data analytics in various domains such as political election, healthcare, and sports. Several attempts had been carried out on Internet-related data for making predictions have been done in different areas. There are various techniques which are used by authors for making social media data fit for decision making and considered as primary data for analysis. The chapter also reviews techniques used for predictive mapping along with decision-making in healthcare.

2.1 OVERVIEW

Social networking sites have become fast and low cost communication that enables quick and easy access to political information among potential users. Social media users express their sentiments and opinions towards different parties and their leaders during political elections. Initial studies often presented optimistic results regarding the predictive capacity of Twitter data relative to election results. Some researchers had found that the volume of candidate or party mentions alone reflected election results [Gordon, (2013)].

There exist many mediums, where people can express themselves on the web. Blogs, wikis, forums and social networks are examples of such mediums, where users can post information, give opinions and get feedback from other users. In their own right, they collectively represent a rich source of information on different aspects of life, but more importantly on different topics, ranging from politics and health to product reviews and travelling. The increasing popularity of
personal publishing services of different kinds suggests that opinionative information will become an important aspect of the textual data on the web.

Due to the ever-growing size of the information on the web, we are now barely able to access the information without the help of search engines. This problem gets harder, when we want to aggregate the information from different sources. Multiple solutions have been proposed to solve this problem, and they are mainly specialized in factual information retrieval.

The field of sentiment analysis has recently witnessed a large amount of interest from the scientific community. Sentiment analysis has traditionally been applied to a single domain at a time, such as movie reviews or product reviews [Li & Wu, (2010)]. More recently, much effort has been invested into development of sentiment analysis methods that can be used across multiple domains like movie and product reviews, election result prediction; disease outbreak etc [Wanga et al., (2014)].

### 2.2 SOCIAL MEDIA ANALYTICS

Social media data have been a primary focus in the field of information retrieval (IR) and text mining due to an excessive amount of unstructured data in real time. Every tweet, comment and blog post might reflect their sentiments [Jain and Kumar, (2015a)]. These unstructured data provides valuable knowledge which constitutes a big opportunity for creating new services for governments, businesses or individuals. Exploiting these unstructured data created a new field called opinion mining and sentiment analysis.

Recently, authors developed sentiment analysis methods that can be used across multiple domains like movie and product reviews, election result prediction; disease outbreak, stock market etc. Indian social media users have rapidly evolved over the past few years to form a
complete ecosystem which deals in several areas such as news, politics, health, government policies and finance [Jain and Kumar, (2015b)].

2.2.1 Social Media Analytics in Political Election

Today every political campaign have presence in social media because of it researchers from various disciplines interested in mining knowledge from theses data. Social media data have two main types of textual information on the web: facts and opinions. On one hand, facts are assumed to be true and on the other, opinions express subjective information about a certain entity or topic. Social media websites like Twitter is one of the fast and convenient medium to diffuse news and political messages. Election experts argue that usage of social media was one of the reasons for Obama to won the US presidential election in 2008[Matthew and Dutta, (2008)]. Unique communicational characteristics play important role in winning the election through micro blogs such as Twitter [Kaplan and Haenlein,(2010)]. Platform such as Xbox gaming is also used for forecasting election[Wanga et al.,(2014)]. The question raised by McComb whether the mass media sources actually reproduce the political world perfectly [McComb and Shaw,(1972)]. Whinston and Huaxia (2010) argue that “the unique innovation of social media is recognizing and connecting people’s need for information and attention” and as such, its design should facilitate such a connection. Acquiring political information on the internet is associated with political discussion and online civic messaging, which are associated in turn with participation is explained by Shah et.al(2005). Xenos and Moy(2007) demonstrated direct effects of online information on political knowledge and differential effects on participation moderated by political interest. Bollen et al. (2011) have done another similar study, where they analyze the tweets and check how the social, political, cultural and economic sphere produces an effect on public mood expressed on Twitter. Hughes and Palen(2009) argue
in favour of using micro blogs as a public information channel used by authorities, for instance in emergency situations. Jungherr et al. (2012) argue that methods of prediction using social media analytics are frequently contingent on somewhat arbitrary experimental variable. Tumasjan et al. (2010) investigate how Twitter is used in political discourse and check if political sentiment on Twitter reflects real-life sentiments about parties and politicians. O’Connor et al. (2010) relate the opinions over Twitter to outcomes of a large-scale poll, i.e., they show the correlation between events occurring in the same time period during which the Twitter analysis is carried out. Conover et al. (2011) applied label propagation to a re-tweets graph for user classification, and found the approach to outperform tweet content based machine learning methods. Chung and Mustafaraj (2011) found that merely counting tweets is not enough to obtain good predictions and measure the effect of sentiment analysis and spam filtering. Sang (2012) reduce this bias by using the number of users rather than the number of the tweets. Gayo-Avello (2011) pointed other issues such as, large data does not make such collections statistically representative samples of the overall population. Second, not all tweets are trustworthy; there are many spam tweets and campaign tweets that do not represent the sentiments or opinions of the users. Nooralahzadeh (2013) used more complete keywords by adding the campaign and election hashtags. Many other authors have used different techniques in different types of election.

### 2.2.2 Social Media Analytics in Healthcare

The crisis occurred due to epidemic activity caused a massive damage to human life, it is very important for governments and public health agencies to communicate with accurate, timely, direct, and relevant messages to the public by using social media or other broadcasting medium. Diffusion of accurate information during health emergencies is necessary to overcome the impact of epidemic diseases. Improving public health information system using social media data can
overcome the massive damage of human life by detecting outbreaks. Some authors carried used social media data for detecting disease outbreak such as Chew(2010) proposed a method during 2009 H1N1 pandemic, using Twitter. The method is based on specific keywords. Various authors also used some other techniques for finding keywords like Google web search queries related to influenza epidemic [Signorini,(2009)]. Technique based on some specified search terms(flu, vaccine, tamiflu, “h1n1”) give high accuracy when applied for in the tweeter related data set. Content analysis and regression models are used to measure and monitor public concern and levels of disease during the H1N1 pandemic in the United States by Hu et al.(2011) and Lampos and Cristianini(2010). Some other diseases like cholera is also investigated by Chunara et al.(2012) to find the cholera outbreak. Authors also developed a framework which provide to quantify users affected by influenza (swine flu) within a community or group and based on this introduce a user rank algorithm[Tang and Yang, (2010)]. Some authors used machine learning techniques like SVM to predict influenza rates using twitter dataset in Japan[Aramaki et al.,(2011)]. Lampos and Cristianini(2010) proposed a method for tracking the various epidemic activities. Some authors also developed tool for real-time analysis of disease using Twitter data, showing daily activity of the disease and symptoms by analysis the text [Lee et al.,(2010)]. Stewart and Diaz(2012) discussed early warning system, as well as Outbreak Control and Analysis Systems. Hansen et al. (2011) analyzed which tweets attract the biggest attention. Bodnar and Salathe(2013) applied various classification techniques for detecting influenza. Parket et.al (2013) developed a low cost frame work for tracking public health condition trends via Twitter. Ghosh et al.(2013) forecast rare disease outbreaks and used spatiotemporal topic forecast outbreaks.
2.2.3 Social Media Analytics in Sports

There are number of methods used to forecast success (winners and losers), both for single games and team games in the sport. These events have a great uncertainty towards the results if teams and players are strong contenders. Limited research is carried out to forecast results using social media data. According to Wang (2013) supporter of the team and players use social media to express their emotion and opinions. Yu and Wang (2015) use twitter data for FIFA world cup 2014 to analysis the emotions users and also describe event based tweets response. Some authors also studies to predict outcomes of EPL games played during the 2013-2014 season based on sentiment analysis [Godin et al.,(2014); Radosavljevic et al.,(2014)]. Sinha et al.,(2013) used n-grams from Twitter data sets to predict outcomes of the National Football League (NFL) and compared it with other simple statistics methods. Lock & Nettleton(2014) applied machine learning technique to classify tweets and also used situational variables. UzZaman et al.(2012) used a framework (TwitterPaul) to extract tweets and find the outcome of FIFA World cup 2015.

2.3 SENTIMENT ANALYSIS

Sentiment analysis is the computational study of people's opinions, sentiments, emotions, and attitudes. It focuses on developing automatic systems that can analyze natural language texts to determine the sentiment expressed in them. The word “sentiment” is often used in a wide sense to refer to expressions of subjectivity, opinion, affect, attitude, orientation, feelings, emotions, and tone in the text [Jain & Kumar, (2015b)]. Much of the current work in sentiment analysis has focused on the task of determining the presence of sentiment in the given text, and on determining its valence, that is, the classification of sentiment according to positive or negative orientation. This fascinating problem is increasingly important in business and society.
2.3.1 Sentiment Analysis Methods

The existing work on sentiment analysis can be classified as the problem of text classification in different forms such as Document-level sentiment analysis, Sentence-level sentiment analysis, Aspect-based sentiment analysis, Comparative sentiment analysis and Sentiment lexicon acquisition [Jain and Kumar,(2016c)].

1. **Document-level sentiment analysis:** This technique assumed that every document contains an opinion or opinion related words. Classification of these documents is carried out using two approaches: supervised learning and unsupervised learning.

2. **Sentence-level sentiment analysis:** This technique is based on specialization of documents which contains various sentences and every sentence contains an opinion behind it.

3. **Aspect-based sentiment analysis:** Attribute(aspect) based classification is performed such as in customer reviews of a mobile phone which have different attributes like battery, camera, screen size or resolutions, processor, memory etc. Every attribute opinion of the product is considered for better understanding sentiments.

4. **Comparative sentiment analysis:** In this technique identification of sentences which contain comparative opinions are filter out and opinions are extracted by considering sequential patterns as features.

5. **Sentiment lexicon acquisition:** This is most popular and crucial resource for most of the sentiment analysis algorithms. In this technique, acquisition of the sentiment lexicon is performed by manually, dictionary-based approaches and corpus-based approaches.
2.3.2 Emotion Extraction using Social Media

Emotional states have multiple cognitive bases which are formed by number of factors. The scope of this chapter is limited to provide important features relevant to recognize emotion and the process of determining the emotional orientation of customer reviews. Emotion present in multiple languages or simply in a language can be expressed in several ways and understanding of this expression can be interpreted differently by different readers.

During the last few decades, Computational techniques such as Natural language processing, Artificial Intelligence (AI), Machine learning has been used to develop intelligent systems that interpreted human emotions [Jain et al., (2016)]. Real-time processes have been modeled to develop effective intelligent systems by providing learning, perception and reasoning. Development of affective interfaces is an important research area in emotions. Affective interfaces are the system which provides emotional inputs; emotional responses using this facilitate online communication with the help of animated affective agents. These affective interfaces provide better user experience in following areas such as Computer-Mediated Communication (CMC) and Human-Computer Interaction (HCI). Text-to-Speech (TTS) synthesis systems are also used these interface for developing real-time systems.

Recognizing the polarity of sentiment present in written text such as in social media posts, blogs and forums is the most popular in sentiment and opinion mining domain [Pang and Lee, (2008)]. Among the less paying attention area related to sentiment is the identification of emotions present in customer reviews and social media data. Computational techniques related to emotion analysis present in text mainly deals with their emotion modalities and emotion-annotated data [Jain et al., (2016b)]. However, only limited work has been done in developing
automatic emotion recognition system which benefits customer and e-commerce companies in decision-making. The focus of this chapter is to give advantages of recognizing emotions present in customer reviews which provide an insight related to customer’s intentions.

2.3.3 Theories of Emotion

Theories of emotion mainly categorized in terms of the context within which the explanation is developed. There are different emotion theories are used in physiology and other relevant fields [Cowie et al.,(2011)]. Recently, researchers have investigated a number of aspects related to human emotions in multiple domains. Authors categorized emotions terms in multiple emotion classes based on emotion-bearing words which are accepted across the world. Multiple milestone work has been carried out by researchers to provide right judgment related to emotion categories [Tomkins ,(1962); Plutchik ,(1980); Izard,(1977); Ortony et. al., (1988); Raghavan, (2007) and Ekman, (1992)]. Emotion categories recognized by the different researchers has been presented in Table 2.1.

According to the popular emotion theory (Rasa theory) given by Bharta Muni, emotions are the primary gastric juices which represent every part of the world into the sentiments [Raghavan,(2007)]. Rasa emotional theory is based on the Natyashastra (the Textbook on Drama) and given by Bharata Muni. According to Natyashastra[Raghavan,(2007)], there are eight rasas corresponding to Bhava (mood). Ekman[Ekman,(1992)] has defined six basic emotions on the basis of distinctive facial expressions which are universally accepted has been presented in Table 2.1.

Watson and Tellegen in 1985 has presented a compressive theory i.e The Circumplex Theory of Affect which identifies two main classes as positive and negative affects which range from high to low[Watson and Tellegen, (1985)]. Osgood’s in 1957 has proposed theory of Semantic
Differentiation [Osgood et al., (1957)]. In this theory words are assigned by emotive meanings using three factors such as activity factor, evaluative factor and potency factor.

Clore et al. (1987) proposed number of words which convey emotion explicitly, while other are depending on the context. The classification of affective words into two classes 'indirect affective words' (implicit) and 'direct affective words' (explicit) has been performed by Strapparava and Valitutti [Strapparava and Valitutti, (2006)]. The chapter has utilized both these types of words given by Strapparava and Valitutti (2006).

Table 2.1 Emotion Categories or classes Identified by Authors

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Hasya (Mirth)</td>
<td>Joy</td>
<td>Disgust</td>
<td>Joy</td>
<td>Joy</td>
<td>Fear</td>
</tr>
<tr>
<td>Krodha (Anger)</td>
<td>Anguish</td>
<td>Enjoyment</td>
<td>Sorrow</td>
<td>Sadness</td>
<td>Anger</td>
</tr>
<tr>
<td>Rati (Love)</td>
<td>Shame</td>
<td>Sadness</td>
<td>Anger</td>
<td>Fear</td>
<td>Fear</td>
</tr>
<tr>
<td>Bhaya (Terror)</td>
<td>Fear</td>
<td>Fear</td>
<td>Anger</td>
<td>Disgust</td>
<td>Fear</td>
</tr>
<tr>
<td>Jugupsa (Disgust)</td>
<td>Anger</td>
<td>Surprise</td>
<td>Disgust</td>
<td>Disgust</td>
<td>Anger</td>
</tr>
<tr>
<td>Soka (Sorrow)</td>
<td>Disgust</td>
<td>Surprise</td>
<td>Surprise</td>
<td>Acceptance</td>
<td>Disgust</td>
</tr>
<tr>
<td>Vismaya (Astonishment)</td>
<td>Surprise</td>
<td>Shame</td>
<td>Acceptance</td>
<td>Anticipation</td>
<td>Surprise</td>
</tr>
<tr>
<td>Utsaha (Energy)</td>
<td>Interest</td>
<td>Shyness</td>
<td>Guilt</td>
<td>Interest</td>
<td>Disgust</td>
</tr>
</tbody>
</table>

2.3.4 Emotion Recognition in Text

Emotion recognition in text is just one the numerous dimensions of the task of making the computers make sense of and respond to emotions. Emotion represents a complex, subjective experience which involves thinking, excitement, feeling and activation. This can be seen using native languages, physiological changes, and behavioral changes. The word “affect” is often used alternatively with “emotion” in the literature.
Emotion detection has been focused on the sentence-level analysis or document level analysis for learning emotions in customer reviews. Emotional words are used as keywords to identifying emotion in input sentences. Osgood used multidimensional scaling to visualize the affective words for calculating similarity ratings between them. Three dimensions were used by Osgood which includes “evaluation”, “potency” and “activity”, where evaluation can be used to measure how much a word can refer to a pleasant or an unpleasant event [Osgood et al.,(1957)].

Read et al. have carried out appraisal annotation of a corpus of book reviews; a genre that provides ample instances of the various kinds of appraisal classes [Read et al.,(2007)]. They found that out of the three subsystems of the Appraisal Framework, the attitude subsystem’s instances (which include emotion expressions) were the easiest to identify. Mihalcea and Strapparava (2004) present results in favor of automatic recognition of humor in texts [Mihalcea and Strapparava, (2005)]. They perform experiments to identify humorous one-liners, which are one-line sentences generally characterized by simple syntax and use of rhetoric which gives them a humorous connotation. Neviarouskaya et al.(2007) propose a system for augmenting online conversations with a graphical representation of the user, which displays emotions and social behavior in accordance with the text. This system performs automatic estimation of affect in text on the basis of symbolic cues such as emoticons, popularly used IM (Instant Messaging) abbreviations, as well as word, phrase, and sentence-level analysis of text [Singh et al.,(2011)].

Ghazi et al.(2010) proposed a multiple levels of hierarchy classification to classify emotions words. Strapparava et al.(2006) has developed a linguistic resource to lexically represent affective knowledge named WordNet – Affect. The WorldNet-Affect contains affective words and available freely as open-source lexical recourses. Wang et al.(2012) proposed a novel
approach to detect emotion from Chinese language. The algorithm proposed was segment based fine grained emotion detection model that is a supervised learning approach.

Dey et al. (2014) proposed a system for extracting emotions from the data collected from chat messenger. The author built a lexicon of emotion conveying words to extract emotions from the data sets. Shaheen et al. (2014) had proposed a framework for classification of emotions based on generalized concepts extracted from the English sentences. Perikos and Hatzilygeroudis (2013) developed a system that can automatically recognize emotions in natural languages. They used Tree tagger and Stanford Parser. They also used WordNet Affect lexical resource in order to spot the emotional words. After that analyzed emotional words dependencies in order to specify the strength of emotional word’s strength and also determined overall sentence emotional statues that are based on dependency graph of sentence.

2.4 SCOPE OF WORK AND PROBLEMS IDENTIFICATION

This comprehensive literature review on the social media data analysis provides following benefits to the readers. First, the literature outlines the important factors, which help in understand the unstructured data and how to uncover the hidden knowledge presented in it.

Second, the data collection and quality of data has been discussed. Third, the applicability of social media based prediction has been highlighted. Fourth, few machine learning approaches have been discussed which are widely used for analysis the data. Finally, identification of factors used suite for prediction events from social media data. While extensive, the review is also not complete in some aspects.

Based on the literature review, following problems are identified:

1. Several attempts had been carried out on Internet-related data for making predictions have been done in different areas. There are various techniques which are used by authors
for making social media data fit for decision making and considered as primary data for analysis. The data collection and pre-processing techniques play a key role in prediction.

2. The mining of emotions from multilingual text posted on social media by different categories of users is one of the crucial tasks in the field of opining mining and sentiment analysis. Every major event in the world have an online presence and social media users use social media platforms to express their sentiments and opinions towards it. An advanced framework has been needed for automatic detection of emotions of users in Multilanguage text data using emotion theories which deals with linguistics and psychology,

3. The medical diagnosis system varies in the level to which they attempt to deal with different complicating aspects of diagnosis such as the relative importance of symptoms, varied symptom pattern and the relation between diseases themselves. Delay in identifying the beginning of infectious epidemic results a big damage towards a society.


5. Deep learning has become popular in all aspect related to human judgments. Most machine learning techniques work well which includes text classification, text sequence learning, sentiment analysis, question-answer engine, etc.