CHAPTER 1
INTRODUCTION

In the past 40 years, data was mainly used to record and report business activities and scientific dealings, and in the next 40 years data will be used also to influence business decisions and to speed up scientific discovery. During the recent years the Web turning into an open platform, where individuals can express their emotions and can be heard. There are numerous online social media platforms that enable individuals to post opinions, for example, social networking sites, wikis, blogs, Twitter, forums and others. These posts provide rich and valuable information on various points or events, which can be used in different applications for fruitful outcomes. Exploiting this online data by extracting key information from it is one of the great challenges in data mining and knowledge discovery. These unstructured data provides valuable knowledge which constitutes a big opportunity for creating new services for governments, businesses or individuals.

1.1 OVERVIEW
In the past 40 years, data was mainly used to record and report business activities and scientific dealings, and in the next 40 years data will be used also to influence business decisions and to speed up scientific discovery. Every day, we create 2.5 quintillion bytes of data so much that 90% of the data in the world today has been created in the last two years alone[Bollen et al.,(2011)]. The quantity of available data has exploded in the past years because of new social behaviors, societal transformations as well as the vast increase of software systems.
A recent study indicates that unstructured data account for at least 80% of the world's data. This means that many companies today are making critical decisions with only 20% of the data they have the 20% of data that is structured and stored in relational databases.

Huge data sets and various data types lead to new types of problems and require the development of new types of techniques for modern intelligent data analysis. An important objective of intelligent data analysis is to reveal and indicate diverse non-trivial features or views of a large amount of data. Many techniques and models in data mining, machine learning, pattern recognition, statistics, and other fields have been proposed. Each technique or model focuses on one particular view of the data and discovers specific type knowledge embedded in data[Cameron,(2013)].

The rapid development of Web 2.0 applications such as micro-blogging, social networks, news portals and web-forums etc provides a major source of user generated data in the form reviews, comments, recommendations, ratings and feedbacks which are useful for businesses, governments and individuals [Alghamdi,( 2013);Rana et al.(2015)] .The major attraction where digital business lies is in the area of sentiment analysis. Sentiment analysis has been widely used in areas like healthcare, stock markets, sports, customer review analysis etc. The concept behind this task is to classifying the polarity of a given document or text[Jain & Kumar,(2016b)].

Exploiting social media data by extracting key information from it is one of the great challenges in data mining and knowledge discovery Khatwanin & Srivastava,(2015)]. 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. These unstructured data provides valuable knowledge which constitutes a big opportunity for creating new services for governments, businesses or individuals [Pang &
Exploiting these unstructured data created a new field called opinion mining and sentiment analysis.

1.2 SOCIAL MEDIA

Social networking such as Facebook, Twitter, Instagram, Google +, LinkedIn and Blogs are directing their businesses with the same vision that tries to attract more users to develop social relations and virtual environment human interaction in societies around the world[Castillo et al., (2011)]. They conceived the trait that human beings have the tendency to share information regarding their opinion, state of mind, lifestyle, emotions, and sentimental feelings via different avenues. Hence, social media provides a platform for the sharing and witnessing of interactive information[Conover et al., (2011); Culotta, (2010)]. They come with attractive features that allow users to manage their own posts and provide the convenience to save link for sharing or to read later. With these facilities, the shared information is at the reach of fingertips of everyone who has access to that particular social media, at any time, and in any place[Fink, (2013)].

Many scholars apply machine learning approach and implemented text mining tool in social networking websites. Opinion mining/sentiment analysis has been an enduring and fascinating idea. Many researchers have been done in the past few years, such as natural language processing (NLP), web mining, data mining, text mining, and so on. All were done with the same purpose - to gather and analyze disparate words of thoughts and opinions about different topics [Rana et al., (2014)].

By applying sentiment analysis, information can be accurately classified and human thoughts can be analyzed as well. This is significantly important to refine a NLP tool in order to perform mining task efficiently. There were numerous attempts aimed to enhance the work of NLP, yet it is still confined by stark challenges[Grossman and Frieder, (2004)].
The growth of social media data has interested researchers from various disciplines to uncover the hidden knowledge by applying intelligent data analysis techniques. There are 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.

1.3 SENTIMENT ANALYSIS AND OPINION MINING

Sentiment analysis or opinion mining, as it is sometimes called, is one of many areas of computational studies that deal with opinion-oriented natural language processing. Such opinion-oriented studies include, among others, genre distinctions, emotion and mood recognition, ranking, relevance computation, perspectives in text, text source identification, and opinion-oriented summarization [Pang and Lee, (2008); Guan et al.,(2016); Hasan et al.,(2016)].

Sentiment analysis (or more specifically, sentiment polarity analysis) can be defined as the mapping of text to one of the labels (elements) taken out of a pre-defined set or placing it on the continuum from one end to the other [Pang and Lee, (2008)]. The elements of the pre-defined set are usually 'negative' and 'positive', but they can also be any other elements such as 'relevant' and 'irrelevant', 'in favour of' or 'against', or other more than two elements such as 'negative', 'positive', 'neutral' or a range of numbers such as from 1 to 10.

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.

All problems in opinion-oriented text analysis can be formulated as problems of text classification [Pang and Lee, (2008);Hasan et al.,(2016)]. Thus sentiment analysis can be redefined as classification of a textual entity into one of the elements or another of the predefined set described above. Sentiment analysis can be done at different levels - document, section,
paragraph, sentence, or phrase levels. Most of sentiment analysis work is done at a document level, but there are works on sentiment analysis at phrase and at sentence levels [Wilson et al., (2000), Hatzivassiloglou & McKeown, (1997)]. Studies have shown that sentiment analysis at phrase level proved to be more difficult than at higher levels [Wilson et al., (2005); Hasan et al., (2016)].

Quite a bit of early Sentiment Analysis examine centred around item product reviews, for example, items on Amazon.com, characterizing opinion as either positive, negative, or unbiased. These were advantageous marked information, as star evaluations were utilized as quantitative pointers of the creator's assessment. Afterward, annotated datasets were made for platforms, for example, online journals, website pages and news articles.

Late development of Twitter has created a plenty of research following subjects and estimation for a wide range of new applications: for example, try to predict box-office revenues of movies, track H1N1 epidemics, and monitor effects of an earthquake, all using Twitter [Jain and Kumar, (2016)].

Although exciting in their diverse applications, most sentiment analysis studies have focused on one social media source, tailoring their approaches to a subset of a wide variety of texts. Furthermore, analysis of political discussions proves to be quite challenging, putting in question whether conceptualization of sentiment and lexicon-based approaches developed for mining product reviews are suitable for analysis of such rich discourse.

1.4 SOCIAL MEDIA BASED DECISION-MAKING

Over the years, surveys have been the main method for answering the question what do people think. 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 [Hong and Fang, (2015)].

The automatic recognition of sentiment, particularly in large volumes of documents, can have a variety of applications, notably in summarizing popular sentiment about any product or issue. Such information may particularly be of interest to policy-makers, economists and market researchers, political analysts, and social scientists [Jain & Kumar, (2015b)].

This thesis contributes to a field of Sentiment Analysis (SA), which aims to extract emotions and opinions from text, and most notably from social media.

With the growing availability of large volumes of textual information on the Web, text mining techniques have been gaining a growing interest. One specific text mining problem that is increasingly relevant relates to the detection of textual expressions that refer to opinions on certain topics and services. A second text mining problem, which has also been gaining a growing interest, is the identification of the geographic location that best relates to the contents of particular documents.

1.5 Thesis Context and Problems

Individuals give personal experiences and opinions to encounters and suppositions towards various points in the Web: In general, they express their feelings and opinions with a particular platform or services present in the web. In blog postings, information, arguments, and opinions can be found. Gathering postings can be viewed as online discourses and exchange platform of arguments. Microblog messages bear a response of individuals towards different occasions and personal experiences. All these communicated conclusions are intriguing for various applications. Consider the accompanying situations:
1. Given a particular specific treatment, a doctor may be occupied with positive and negative experiences or opinions as for his treatment. Be that as it may, not just a quantitative impression is helpful to him, additionally how the supposition changed after some time and which contentions were utilized as a part of support or against a particular treatment. Another case of differentiating suppositions originates from political subjects, which are broadly talked about in the Web. Monitoring the assessments communicated towards these themes may be pivotal to think of the right forecast of the effect and support for specific arrangements or choices.

2. In some cases, opinion of a group of people deviates from the general population or from that of another group with different demographical attributes. Another sort of inclinations that we are keen on considers behavioral contrasts of demographical gatherings, e.g. at the point when individuals living in India think differently their nearby event contrasted with individuals from Europe as a rule.

3. Changes in public opinion are usually driven by new information or evidence from new sources. However, a major problem is that the text information is not mentioned explicitly, and it is not easy to recognize the cause of sentiment changes. Correlation of news trend and analyze the volume of news can provide valuable insights. This dynamics nature of social media helps in predicting these changes.

4. Evidently, such problems require analyzing significant amounts of data to produce a desired output, and special methods that can exploit this volume to improve the resolution and representativeness of sentiment analysis. Conventional sentiment aggregation methods may be inefficient for large scale analytics, especially when subsequent time intervals contain different amount of sentiments and when a simple average of diverse
sentiment values is taken. The information about real sentiment values can even be lost, when aggregating opposite values.

5. The aforementioned problems also require efficient sequential time series access methods with a possibility of hierarchical navigation over time. However, the databases commonly used for sentiment storage and access, are not optimized to track the evolution of sentiments on a large scale or to support fast update rates.

1.6 ORGANIZATION OF THE THESIS

The objective of our research is to extend the current work on social media analytics. The work reported in thesis is organized into eight chapters, as given below:

CHAPTER-1: INTRODUCTION

This chapter provides the introduction to the different approaches for social media analytics. A brief discussion of a sentiment analysis, opining mining, and decision making through social media has been covered. This chapter describes the important aspect related to social media mining with some important factors.

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. [Jain and Kumar, (2015b)].

CHAPTER-3: MULTI DIMENSIONAL EFFECTIVE APPROACH FOR EVENT PREDICTION USING SOCIAL MEDIA

In this chapter, an effective Multi Dimensional prediction framework using Twitter data has been developed. The data sources encompasses an edge of advantage over other medium due to its direct access to information about events, easily accessible at nearly zero cost, real time analysis and automated language processing capabilities. For the purpose of analysis three different events such as Political election, Healthcare and Sports have been considered. [Jain and Kumar, (2015a); Jain et al. (2016); Jain and Kumar (2017c)].

CHAPTER-4: A NOVEL FRAMEWORK FOR EXTRACTION OF EMOTIONS FROM MULTILINGUAL TEXT USING INTELLIGENT TEXT PROCESSING AND COMPUTATIONAL LINGUISTICS

In this chapter, an advanced framework for automatic detection of emotions of users has been developed. The proposed approach used Multilanguage text data using emotion theories which deals with linguistics and psychology. The emotion extraction system is developed based on multiple features groups for better understanding of emotion lexicons. Empirical studies of three real-time events in domains like a Political election, healthcare, and sports are performed using proposed framework. [Jain et al., (2017a)]
CHAPTER-5: ROUGH SETS BASED INTELLIGENT APPROACH FOR IDENTIFICATION OF SUSPECT USING SOCIAL MEDIA

In this chapter, an intelligent technique based on rough set theory for finding suspect patient for swine flu using social media data. The symptoms of swine flu are taken from reputed medical and research agencies as keywords and are used for the collection and examination of tweets. Identification of suspects is performed by identification of most important conditional attributes and discovery of decision rules characterizing the dependency between the values of conditional attributes and decision attributes. [Jain and Kumar, (2018)].

CHAPTER-6: A NOVEL APPROACH FOR EFFECTIVE SURVEILLANCE AND PREDICTIVE MAPPING OF MOSQUITO-BORNE DISEASES USING SOCIAL MEDIA

In this chapter, social media based mosquito-borne disease surveillance and outbreak management system has been presented. The system uses spatial and temporal information which help in identification, characterization, and modelling of user behavioural patterns on the web have been presented through this chapter. The proposed predictive mapping based on geo-tagging data has a significant impact on preventing and tracking mosquito-borne disease in the specific area with limited resources.[Jain and Kumar, (2017b)].

CHAPTER-7: SENTIMENT RECOGNITION IN CUSTOMER REVIEWS USING DEEP LEARNING

In this chapter, two objectives have been completed using Deep learning models. Firstly is to study the applicability of deep neural networks strategies for extracting sentiment present in social media data and customer reviews with effective training solutions. The second objective is
to design deep networks that can be trained with these weakly supervised strategies in order to predict meaningful inferences. This chapter presents the concept and steps of using deep learning for extraction sentiments from customer reviews.

CHAPTER-8: CONCLUSIONS AND SCOPE OF FUTURE WORK

This last section of thesis concludes the study to address the identified problems and discuss the scope for enhancement in future.

i. Event prediction framework using dynamic keywords has been proposed. This framework is applied in three different multidimensional events. The results show social media sources can help in forecasting election results. Analysis of disease outbreak detection results shows that, the developing countries, where public health monitoring system are not quick and reliable, can also track disease outbreak in feasible time and take measure for prevention. In another section; short term sports related social media has been collected for predicting the outcomes.

ii. Public emotions present in social media data offers unique challenges and opportunities for in decision-making in different domains. The major contribution of this research is to present that it is feasible to apply intelligent computational techniques for identification and classification of various types of emotions in texts.

iii. A rule based system for finding suspect and symptoms of swine flu using social media data has been developed. Tracking the symptoms of epidemics diseases from social media data and use rule-based system help in decision making. Given approach could give rapid information in various situations during epidemic and improve public health monitoring system in developing countries.
iv. Traditional methods of surveillance rely on bureaucratic, hierarchical, health-care system which adds a lot of time delays in detection that cannot usually gear up with the speed at which a disease is spread. The proposed framework focused on alternative methods of analysis and visualization of user’s opinions that do not depend upon the assumption of normality and historical data. An intelligent surveillance process model for decision making has been proposed through this chapter.

v. From the e-commerce perspective, receiving consumer reviews/feedbacks can significantly improve its business action plan in order to increase profits. These reviews help in gaining business intelligence and improving customer satisfaction. Deep learning has been used for predicting sentiments present in customer reviews using two popular deep learning models such as Convolution neural networks (CNN) and Recurrent Neural Network (RNN) using the Long Short-Term Memory (LSTM) architecture.