CHAPTER 2

LITERATURE SURVEY

Social media has become this significant part of modern civilization. Social media is changing existing information behavior by giving users access to real-time online information channels without the constraints of time and space. This gives scientist an enormous scope for data analysis challenge. Most previous studies have adopted broad-brush approaches that typically result in limited analysis possibilities.

Author Min Song, MeenChul Kim and Yoo Kyung Jeong [1][2] has studied a twitter dataset for 2012 Korean election by collecting real time tweets. Topics extracted from tweets and related real time events relation was identified and they were traced chronologically using term co-occurrence retrieval technique.

For India 2014 general election; User's orientation towards parties and candidates was studies by **Author Abhishekbhola [3]** using Tweeter. A dataset consisting of 17.60 million of tweets was analyzed for identifying user, candidate, and party popularity based on peak of time, topic or location. A sentiment analysis was performed using classification algorithms

Voting advice applications (VAAs) [4] are online tools designed for election in countries like Greece. It is becoming increasingly popular and they are helping users in deciding which party/candidate to vote. It is designed based on the concept called community-based recommendations system. It provides the comparison of users' political opinions, and becomes a channel for user communication. This system proposed various approaches for community-based vote recommendation. The approaches were evaluated on five real VAA datasets in terms of prediction accuracy.

Using Data from Facebook and Twitter**; Lars Kaczmirek and his team from GESIS [9]** gathered various aspects of the communication structures. They compared data gathered from social media with local survey and added new insights to social media by providing how social media can be used during elections. Based on this studies; German Longitudinal Election Study (GLES), a long term research project is designed; that examines the German federal elections for the years 2009, 2013, and 2017. The main aim of this project is to track the German electoral process over an extended period of time; by collecting Twitter and Facebook data about the German Bundestag elections.

There is multiple works related to sentiment analysis. **Unnamalia K copied Ohbyung Kwon and Namyeon Leea [6]**. Showed that sentiment analysis of tweets is a challenging task due to multilingual and informal messages. In this study, a research model is proposed to explain the acquisition intention of big data analytics mainly from the theoretical perspective of data quality management and data usage experience. Our empirical investigation reveals that a firm's intention for big data analytics can be positively affected by its competence in maintaining the quality of corporate data.

Chetashri Bhadanea also with William D. Abilhoa and Leandro N. de Castro [7]. This paper proposes a keyword extraction method for tweet collections that represents texts as graphs and applies centrality measures for finding the relevant vertices (keywords). To assess the performance of the proposed approach, three different sets of experiments are performed. The first experiment applies TKG to a text from the time magazine and compares its performance with that of the literature. The second set of experiments takes tweets from three different TV shows, applies TKG and compares it with TFIDF and KEA, having human classifications as benchmarks. Finally, these three algorithms are applied to tweets sets of increasing size and their computational running time is measured and compared. Altogether, these experiments provide a general overview of how TKG can be used in practice, its performance when compared with other standard approaches, and how it scales to

larger data instances. The results show that TKG is a novel and robust proposal to extract keywords from texts, particularly from short messages, such as tweets [3].
Bo Pang and Lillian focus on applications of sentiment analysis that go beyond extracting a sentiment value from a single text. Their application range from sentiment computation towards identifying topics of a text, the visualization of sentiment as well as automatically defining the usefulness of a customer review [5].

Sentiment analysis is very popular technology in today's world. Most of work has been done in this field. Following are most popular approaches in today's world in this technology. There is most of research in the area of this analysis. Bo Pang and Lee was the pioneer's king in this field. Now the Current works in this area includes using a mathematical approach which uses a formula for the sentiment value depending on the proximity of the words with adjectives like ' excellent', ' worse', 'bad' etc. Our project uses the **Nai"ve-Bayes** approach and a HADOOP cluster for distributed processing of the all type of data.

A comparative study between different methodologies has been reviewed and analyzed including subjectivity detection, feature selection for opinion mining, and different machine learning approaches[1]. Various mechanism has been implemented until now, which includes bags of words, training corpus, document level, sentence level and feature-level opinion mining[3]. Different polarity measures exist according to the external system wherein sentimental analysis is utilized. The linguistics feature and domain relevant features are essential for providing the better classification of text[2]. Hence, in this system, consideration the gamut of keywords associated with the feature is essential for successful classification. The algorithm explained revolves around the expansion and better understanding of the model proposed by **Dave, Lawrence and Pennock[8]**.

Turney et.al. [9] Used bag-of-words method in which the relationships between words was not considered at all for sentiment analysis and a sentence is simply considered as a collection of words. To determine the sentiment for the whole sentence, sentiment of every individual word was determined separately and those values are aggregated using some aggregation functions. **Pak and Paroubek [10]** proposed a model to classify the tweets as positive and negative. By using Twitter API they created a twitter corpus by collecting tweets and automatically annotating those tweets using emoticons. Using that corpus, the multinomial Naive Bayes sentiment classifier method was developed which uses features like POS-tags and N-gram. The training set used in the experiment was less efficient because they considered only tweets which have emoticons. **Po-Wei Liang et.al. [11][12]** Used Twitter API to collect data from twitter. Tweets which contain opinions were filtered out.

Unigram Naive Bayes model was developed for polarity identification. They also worked for elimination of unwanted features by using the Mutual Information and Chi square feature extraction method. Finally, the approach for predicting the tweets as positive or negative did not give better accuracy by this method. That [4], proposes a linguistic approach system for aspect based opinion mining, which is a clause/Sentence level sentiment analysis for opinionated texts. For every message post sentence it generates a syntactic dependency tree, and splits the sentence into clauses. It then determines the contextual based sentiment score for each clause using grammar dependency of words and uses SentiWordNet which has prior sentiment scores for the words and also from domain specific lexicons**. Hussein [13]**, this paper explains the previous works, the goal is to identify the most significant. Challenges in sentiment and explore how to improve the accuracy results that are relevant to the used techniques. All the above mentioned work uses the corpus data in this paper the real streaming data based on the filters used and it does not require any memory to store the tweets.