CHAPTER-2: LITERATURE REVIEW

LITERATURE REVIEW

In this Literature Review Chapter 2 shows important significant literature most related to this sentiment analysis research work with real time twitter application. On the one hand, it discusses the Studies which were concerned with sentiment analysis for prediction purpose and in election campaigns, and on the other hand, it considers studies on the active learning approaches applied to data mining. Some sort of algorithm and techniques found related to this research work. Here discusses the motivations for performing knowledge discovery on opinion data sources, and illustrates how the field of opinion mining is closely related to that of data mining. The relevance of opinion mining techniques to the field of knowledge management is also explored. This research is concerned with prediction of popular content from social media tasks in identifying the interesting cases from the twitter datasets with sentiment polarity of information. Traditional approaches for deriving knowledge from data rely strongly on manual analysis and interpretation.

For any domain Election prediction, Scientific data analysis, Marketing, Finance, Health, Business, etc. The success of a traditional analysis depends on the capabilities of one more specialists to read into the data scientists go through manual data analysis to mark interest objects or opinion, such as impact of election analysts go through the people’s opinion to determine which are prone to end in defaults. Such an approach is slow, expensive and with limited results, relying strongly on experience, state of mind and specialist know how.

The huge data generation is on a dramatic raise, which makes traditional approaches impractical in most domains. Hidden strategic pieces of information for fields such as marketing, education, science and health or business of such knowledge. Besides the possibility to collect and store large volumes of data, the information era has also provided us with an increased computational power. The need to discover the interesting knowledge in the raw data and classifying them suitably is undertaken in the research.
The big boom of Social Media reveals the greater opportunities Available to the Machine Learning system.

Today traditional learning approach alone cannot provide effective learning methodology. The conventional learning systems think about the modern pedagogical strategies which are mainly focused on the user centric approach. The growing popularity of social media and mobile devices and user centric approach lead to new innovative pedagogical strategy such as the Machine Learning system. Opinion Mining is a relatively new and challenging field dedicated to detecting subjective content in opinionated documents, Sentence and aspect of user with a variety of uses in real world applications.

The main concern of this research is to perform opinion mining tasks and enhancing the classification accuracy of the classifiers through evolutionary techniques. It is an emerging area of research to extract the subjective information in source materials by applying Machine learning and lexicon approach for classification the polarity of the opinion stated. Sentiment analysis in Machine Learning system helps to enhance the classify the popular content from social media. Here for this research following are the related literature for this research and find the appropriate techniques for sentiment analysis.

Year 2002 Rolf Mitsdorffer, Joachim Diederich works on Rule Extraction from Technology Ipos in the Us Stock Market. Machine learning methods are used Bayesian classifications, Support vector machines, decision tree techniques, rule learners and artificial neural networks. C4.5 achieves-the highest level of accuracy with 94% of positive cases predicted correctly and RBP/RULEX results with 89.9% of cases. Ripper achieves 82.7% accuracy with only 3 rules and maximum of 3 antecedents, while SVM Light predicts 80% accuracy [1].

Year 2003 Hong Yu, Vasileios Hatzivassiloglou works on Towards Answering Opinion Questions Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. Necessary component for an opinion question answering system: separating opinions from fact, at both the document and sentence level. they also present a first model for classifying opinion sentences as positive or negative in terms of the main perspective being expressed in the opinion
Results from a large collection of news stories and a human evaluation of 400 sentences are reported, indicating that achieve very high performance in document classification (upwards of 97% precision and recall), and respectable performance in detecting opinions and classifying them at the sentence level as positive, negative, or neutral up to 91% accuracy[2].

Year 2004 Minqing Hu and Bing Liu works on Mining and Summarizing Customer Reviews. The study shows to mine and to summarize all the customer reviews of a product. This summarization task is different from traditional text summarization because only mine the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. In this study, author proposed a set of techniques for mining and summarizing product reviews based on data mining and Sentiment Analysis methods. The objective is to provide a feature-based summary of a large number of customer reviews of a product sold online [3].

Year 2005 Jian Liu, JianXin Yao and Gengfeng Wu works on Super Parsing: Sentiment Classification with Review Extraction. The study shows describes the sentiment classification of review extraction. Whole process can be illustrated logically as extract the review expressions on specific subjects and attach sentiment tag and weight to each expression, calculate the sentiment indicator of each tag by accumulating the weights of all the expression with the corresponding tag and given the indicators on different tags, apply a classifier to predict the sentiment label of the text. A system Approximate Text Analysis (ATA) is used for review extraction. It follows the idea of Super Parsing, which enables non-adjacent constituents to be merged to deduce a new one. To traverse the valid constituent combinations in Super Parsing, an algorithm named Candidate List Algorithm (CLA) is proposed. Then the performance of three kinds of classifiers simple linear classifier (SLC), Support Vector Machine (SVM) and decision tree is studied. The experiments on on-line documents show that the SVM algorithm achieves the best performance and SLC achieves lowest accuracy [4]. 2006 Qiang Ye, Wen Shi, Yijun Li work on Sentiment Classification for Movie Reviews in Chinese by Improved Semantic Oriented Approach. Sentiment classification aims at mining reviews of customers for a certain product by automatic classifying the reviews into positive or negative opinions. Semantic Orientation approach for sentiment classification to Chinese Reviews. In this
study obtains 170 pieces qualified movie reviews, and labeled them into two parts, 40 pieces of them positive and 130 pieces of them negative. Opinion Classification by Semantic Orientation approach. Author selected 72 reviews, 36 positive and 36 negative to conduct opinion classification. For both approaches author are calculated Accuracy, Recall, and Precision.

The accuracies were between 60% to 80%, the proposed SO approach for Chinese reviews performs not bad. This study also introduced an optimal Reference-Word-Pair (RWP) selecting process and improved the selection of optimal SO value threshold. The research process implied that when SO approach is applied to customers’ sentiment analysis in different fields to different products, different RWP and different threshold are required [5].

2007 ZHANG Zi-Qiong, YE Qiang, LI Yi-Jun works on Using Naive Bayes Classifier to Distinguish Reviews from Non-review Documents in Chinese. In this study proposes an approach to perform subjectivity classification on Chinese text based on a supervised machine learning algorithm, Naive Bayes. Experiment studies have been conducted on two kinds of documents: movie reviews and movie plots written in Chinese. Author developed a tool for Chinese words segmentation. Another candidate for this task is the segmentation component of ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) from CAS (Chinese Academy of Sciences). Author applied both tools in the experiments. To evaluate the performance of the machine learning method for subjectivity tagging, Author adopted three indexes which are generally used in text categorization: accuracy, recall and precision. In this study presented a Bayesian classifier to predict subjectivity for Chinese documents. The results of experiments are very promising with the highest accuracy of 87.6% [6].

2008 Bin Shi, Kuiyu Chang works on Generating a Concept Hierarchy for Sentiment Analysis. In this study, author study the problem of automatically creating a product concept hierarchy based on textual information contained in on-line customer reviews. Author proposed and evaluated two methods to label each intermediate clustering node with the most representative member context feature vector As demonstrated in both the quantitative benchmark results and qualitative sub-hierarchy comparison with Sanderson’s method, approach can automatically group similar and related nouns into an intuitive hierarchy that is both user friendly and up-to-date [7].
Features to Predict Popular Content in Social Media Mining

2009 Kerstin Denecke works on Are SentiWordNet Scores Suited for Multi-Domain Sentiment Classification SentiWordNet scores are exploited together with additional features to assign a polarity to a text using machine learning. We present a rule-based approach that exploit SentiWordNet scores and machine-learning based approach. The rule-based approach achieved only results of very limited accuracy, while good results can be reported for the machine-learning based approach. The accuracy values are very low and reach from 40% for the classification of news articles to 65% for the classification of reviews of the electronic domain. Machine-learning based approach the accuracy varies between 66% and 82% [8].

2009 Si Li, Hui He, Wei-Ran Xu, Jun Guo Works on Automatic Chinese Sentiment Word Extraction Based On Maximum Entropy. In this study presents an approach for Chinese sentiment analysis at phrase-level. Maximum Entropy (ME) model is employed to extract sentiment words. Author evaluates method with parts of the first Chinese Opinion Analysis Evaluation (COAE2008) corpus, about 10000 sentences. According to the context, Author labels the sentiment words. In experimental procedure, author select word, POS label and word position as features in the LMR template and Chinese sentiment word has been extracted automatically with a good performance [9].

2010 William Simm et.al works on Classification of Short Text comments by Sentiment and Actionability for Voice Your View. In this study collected over 2000 individual short comments on the topic of library Refurbishment. This paper looks towards applying methods from the literature to dataset with the aim of evaluating methods of automatic sentiment and actionability analysis for VoiceyourView application data and has relevance to data from other applications. Two machine-learning classifiers, a Naïve Bayes machine-learning classifier and the ReadMe classifier are used to estimate sentiment and actionability. 2000 comments in the dataset, Author randomly sampled 300 comments for analysis. The Naïve Bayes classifier proved to be the most accurate approach for actionability analysis. The Naïve Bayes (NB) approach proved more accurate for actionability estimation than sentiment estimation [10].

2010 David Alfred Ostrowski works on Sentiment Mining within Social Media for Topic Identification. Unique source of data provides rapid means of customer feedback that is
used to support a number of business areas. Towards this purpose, describe a methodology for the identification of topics associated with peoples Sentiment. This process first employs a Fisher Classification based approach towards sentiment analysis. Data is then retrieved on a regular (daily) interval according to specified keywords. Author has chosen four distinct categories towards research goals. Bad, Neutral, Good, Great training set consisted of 1000 manually evaluated messages from the Twitter. Input set consists of 5748 data samples. Author chose the Fisher algorithm to utilize a normalized approach across four identified categories. Fisher classification method outperformed the standard Bayesian method by providing greater support towards smaller data sets within each category [11].

2011 Alexander Pak, Patrick Paroubek Works on Twitter for Sentiment Analysis: When Language Resources Are Not Available. In this study use Twitter, the most popular microblogging platform nowadays, to collect a dataset of emotional texts in French. Using the collected dataset, estimated affective norms of words to construct an affective lexicon, which have use for polarity classification of video game reviews. Experimental results show that method performs comparably to classic supervised learning methods. They propose to use Twitter to collect a dataset of emotional texts using emoticons as polarity indicators that use to estimate the polarity scores of the most frequent words. The constructed lexicon can be used in emotion studies, sentiment analysis and opinion mining systems [12].

2012 Israa Bukhari et.al author works Social Media Use for Large Event Management Using social media analytics tools, Radian6 and Visible Technologies, Purdue Homeland Security Institute (PHSI) researchers were able to monitor, capture and analyze publicly posted online information pertaining to the Super Bowl XLVI. The study collected and analyzed data regarding the public’s perceptions of the Super Bowl XLVI marketing campaign, as well as Indianapolis’ hospitality, accommodations, and safety. Data was collected from three different platforms: Facebook, Twitter, and blogs. The results of the study provided insights into public sentiment, public dialogue regarding specific citywide events, and trending social media topics associated with each topic’s keyword analysis. In this research they further explored the potential usage and
application of social media analytics tools within local government and found that social media analytics can be of great value for the government in both special events and routine activities. The results are based on the entire body of data collected, since it is not feasible to present results for each of the 36 days monitored. Daily reports were provided to local government that enabled quick action to improve the public’s experience of this event and the hosting city [13].

Ernesto Diaz-Aviles Claudia Orellana-Rodriguez Wolfgang Nejdl works on taking the Pulse of Political Emotions. In Latin America Based on Social Web Streams Social media services have become increasingly popular and their penetration is worldwide. Micro-blogging services, such as Twitter, allow users to express themselves, share their emotions and discuss their daily life affairs in real-time, covering a variety of different points of view and opinions, including political and event-related topics such as immigration, economic issues, tax policy or election campaigns. Taking The Pulse of Political Emotions Authors approach consists of the Data collection process, Emotion and Polarity Analysis, Pattern recognition from the sentiment analysis and Author concludes that the extracted polarity and isolated emotions alone are not good predictors for the opinion captured in an independent opinion poll. But the linear combination of basic emotions opposing pairs and polarity, achieved a prediction performance of 81% in terms of AUC, whereas using polarity alone the AUC dropped to 61%. Author also noticed that the pair of emotions joy–sadness dominated the model [14].

Sang-Hyun Cho, Hang-Bong Kang works on Statistical Text Analysis and Sentiment Classification in Social Media. In this study, they propose a new method of classifying tendencies and opinions in texts of multiple sentence length extracted from social media and covering both formal and informal vocabularies. To extract contextual information from the texts, author carry out computations based on keywords, the position of the sentence and the flow of sentiments in the multiple texts. A feature vector for the given text is constructed from the contextual information, and is then classified with a Support Vector Machine (SVM) classifier as Positive, Negative or Neutral. Method performs well in classifying the gradient of sentiments expressed in social media.

The feature vector is classified into one of three sentiments (Positive, Negative, Neutral) by sentence sentiment classifier. They propose a new approach to sentiment classification
Features to Predict Popular Content in Social Media Mining

at paragraph length using contextual information and sentiment-based domain dictionaries covering formal and informal vocabularies. Contextual information such as key words, the position of the sentence, and the flow of sentiment are computed in texts of multiple sentence length. A feature vector for a given text is constructed from the contextual information and is then classified by the Support Vector Machine (SVM) classifier as positive, negative or neutral. Methodology performs well in classifying the sentiments expressed in the multiple texts of social media [15].

Yulan He, Chenghua Lin and Amparo Elizabeth Cano authors works on Online Sentiment and Topic Dynamics Tracking over the Streaming Data. They propose a dynamic joint sentiment-topic model (dJST) which is able to effectively track sentiment and topic dynamics over the streaming data. They also studied three different ways of accounting for such dependency information, (1) Sliding window where the current sentiment-topic-word distributions are dependent on the previous sentiment-topic specific word distributions in the last S epochs (2) Skip model where history sentiment topic-word distributions are considered by skipping some epochs in between and (3) Multiscale model where previous long- and short timescale distributions are taken into consideration. The experimental results show that while these three models give similar values, both the skip model and multiscale model generates better sentiment classification results than sliding window [16].

Mark Rahmes et.al works on a Qualitative and Quantitative Method. For Predicting Sentiment toward Deployed U.S. Forces the ability to automatically predict likelihood of reaction to specific events and situational awareness is important to many military and commercial applications. The propose a method for predicting sentiment towards deployed U.S. Forces in near real time to efficiently propitiate manpower resources, allocate equipment assets, and reduce cost of analyses. Sentiment prediction is becoming an increasingly important and feasible task based on social media, open source data, physical imagery and abundance of video data feeds. Predicting reaction to events can be time consuming.

The novel idea was to combine modeling technologies and visualization viewer as the front end of a usable tool which can supply data mining information for integrated sentiment analysis. Author use Feature search with a prediction artificial intelligence (AI)
algorithm to calculate sentiment probability. An automatic decision threshold is calculated for sentiment analysis. Global Site eliminates a number of human oriented tasking into one system by combining multiple data feeds, reduction of information, and assignment of weight classifiers to assist in understanding information for planning type activities. Global Site provides an efficient alternative to a full manual review of large amounts of disparate data. Author has defined a novel application of sentiment prediction that is relevant in this environment. The key to success is finding and extracting the right information from the environment quickly, performing accurate and relevant data fusion, and finally effective and efficient presentation to the user [17].

2013 Clay Fink, Nathan Bos et.al author works on Twitter, Public Opinion, and the 2011 Nigerian Presidential Election. In this study analyzes a corpus of Nigerian Tweets collected during the run-up to the 2011 Nigerian Presidential election and compares it with official election returns and polling data. The language model was trained using the Ling Pipe Java API.6 only a small number of Hausa tweets were found for each candidate, supporting author assumption that this domain was mostly restricted to English and Pidgin English. The Hausa tweets found were excluded from analysis. This research demonstrates that it is possible to collect social media data from the developing world and identify enough political content to support country-level analysis. Using tweets mentioning candidates in the 2011 Nigerian presidential election [18].

Neethu M S, Rajasree R works on Sentiment Analysis in Twitter using Machine Learning Techniques Sentiment analysis deals with identifying and classifying opinions or sentiments expressed in source text. In this study, they are analyzing the twitter posts about electronic products like mobiles, laptops etc using Machine Learning approach. By doing sentiment analysis in a specific domain, it is possible to identify the effect of domain information in sentiment classification. They present a new feature vector for classifying the tweets as positive, negative and extract peoples’ opinion about products. Machine Learning techniques use a training set and a test set for classification. Training set contains input feature vectors and their corresponding class labels. Using this training set, A classification model is developed which tries to classify the input feature vectors into corresponding class labels. Then a test set is used to validate the model by predicting the class labels of unseen feature vectors. A number of machine learning techniques like
Naive Bayes (NB), Maximum Entropy (ME), and Support Vector Machines (SVM) are used to classify tweets.

In this study dataset is created using twitter posts of electronic products. Tweets are short messages with full of slang words and misspellings. So author performs a sentence level sentiment analysis. This is done in three phases. In first phase preprocessing is done. Then a feature vector is created using relevant features. Finally using different classifiers, tweets are classified into positive and negative classes. Based on the number of tweets in each class, the final sentiment is derived.

Creation of a Dataset. They are used Naïve Bayes, SVM, Maximum Entropy and ensemble approach for this work [19].

Soha Ahmed, Michel Pasquier, Ghassan Qadah works on Key Issues in Conducting Sentiment Analysis on Arabic Social Media Text. In this study highlights key issues researchers are facing and innovative approaches that have been developed when performing subjectivity and sentiment analysis (SSA) on Arabic text in general and Arabic social media text in particular. A preprocessing phase to sentiment analysis is proposed and shown to noticeably improve the results of sentiment extraction from Arabic social media data. Language is much more than grammar and semantics it conveys sentiment as well.

There are mostly two levels Document and Phrase level. At the document level, researchers classify long text as subjective or objective then the subjective as positive or negative. Text categorized as documents include forum posts, blog articles, product reviews, and Tweets. At the sentence level, short text or even single sentences are classified. Sentence level categorization is also very popular among Arabic sentiment analysis research. In general, the techniques used for sentiment classification can be grouped into three categories machine learning, link analysis, and score-based approaches. Machine learning techniques are the most popular with Support Vector Machine (SVM) and Naïve Bayes (NB) being the most frequently used [20].

Xinzhi Wang, Xiangfeng Luo works on Sentimental Space based Analysis of User Personalized Sentiments With the development of Social Media, online documents such as the comments of news articles, blogs and microblogs have received great attention and the sentiment analysis via online documents has become one popular research area. In this study focuses on establishing user sentimental space obtained from online documents.
to analyze user’s personalized sentiments, which aims to identify users sentimental feature. Affection, sentiment and attributes of user are firstly employed to build user’s personalized sentimental space. Then, the general constrains of user sentiments space are proposed to calculate user’s personality. In this way, author can seek out the personalities of every web users, such as controllability and transmissibility. At last, case studies based on sentimental distribution are illustrated. Though computing personalized sentimental model, author can find sentimental leaders. This paper focuses on user’s personalized sentimental analysis and leaves the classification of users based on their sentimental model alone. It will be interesting to explore the explicit modeling of user classification too. Furthermore, data is limited, it will be beneficial to measure social emotions based on big data of microblogs [21].

Ahmad Kamal, Muhammad Abulaish works on Statistical Features Identification for Sentiment Analysis using Machine Learning Techniques. In this study, they have proposed a sentiment analysis system which combines rule-based and machine learning approaches to identify feature-opinion pairs and their polarity. The efficiency of the proposed system is established through experimentation over customer reviews on different electronic products. Web opinion sources such as merchant sites, Web forums, discussion groups, and blogs are rapidly emerging containing precious information useful for both customers and manufacturers.

A study in reveals that a complete opinion along with its relevant feature is always expressed in one sentence. If a sentence contains product feature, its nearest opinion word can be associated using Parts-Of-Speech (POS) information and dependency relationship. The rule-based approach is exploited to identify feasible feature-opinion pairs from opinion sources, whereas the machine learning approach is used to determine the polarity of opinion bearing words. The rule-based approach is augmented with resolution mechanism which uses backtracking to identify features expressed using pronouns. They presented a feature-based opinion mining and sentiment analysis system. The novelty of the proposed method lies in it’s enrich set of statistical features, and their formulation in a way to produce an effective word-level sentiment classification system [22].

2014 Jasmine Bhaskar, Sruthi K, Prema Nedungadi works on Enhanced Sentiment Analysis of Informal Textual Communication in Social Media. Sentiments may be
Features to Predict Popular Content in Social Media Mining

positive, negative or objective and the method of assigning sentiment weights to terms and sentences are important factors in determining the accuracy of the sentiment classification. For experiments author have used the data set taken from Amazon.com and ebay.com for the products reviews of digital camera from which author have taken 24000 sentences for training and 20000 sentences for testing.

The performance of the classifier can be measured in terms of the four possible outcomes: True positive (TP), true negative (TN), false positive (FP), and false negative (FN). True positive\negative means that a sentence is classified to a positive and negative class when this sentence really belongs to the positive and negative class respectively. Both true positive and true negative are correct classifications. False positive and negative means that a sentence is incorrectly classified to a negative and positive class when this sentence belongs to a positive and negative class. In this study they proposed a new method to improve the sentiment classification of product reviews by considering the objective words and intensifiers.

Here they have used SentiWordNet, a publically available lexical resource to extract the sentiment word from the online reviews. After performing pre processing, negation handling, intensifiers handling and modification of objective words, polarity of the sentences is calculated. In this work they used SVM for sentiment classification. Proposed model can handle classification in an efficient manner than traditional methods. The experiment results show that proposed approach achieved better accuracy than existing methods [23].

Fernando Leandro Dos Santos, Marcelo Ladeira Works On The Role Of Text Pre-Processing In Opinion Mining On A Social Media Language Dataset. In this study describes the development and results of an opinion mining application based on a particular dataset. Opinion mining, or sentiment analysis, is the study field that tries to identify and extract subjective information from nonstructural data the web has many comments, reviews and texts that reflect people’s opinions. Thus, one can think of a way to label these opinions in categories such as”positive”, ”negative” or ”neutral”, and then, automatically identify each review as belonging to one of these categories.

In this study, large amounts of reviews about Android apps were analyzed. In total, 759,176 Brazilian Portuguese reviews were extracted from the Google Play app store.
These reviews have some particularities. Machine learning algorithms are used to train and classify reviews in this work. In this approach, documents are converted to term-document matrices. The study case conducted in this research involves two phases. First, a small amount of the total dataset was analyzed using Weka4[24].

Aliza Sarlan, Chayanit Nadam, Shuib Basri works on Twitter Sentiment Analysis. The objectives of the study are first, to study the sentiment analysis in microblogging which in view to analyze feedback from a customer of an organization’s product; and second, is to develop a program for customers’ review on a product which allows an organization or individual to sentiment and analyzes a vast amount of tweets into a useful format. To associate with Twitter API, developer need to agree in terms and conditions of development Twitter platform which has been provided to get an authorization to access a data.

The output from this process will be saved in JSON file. The reason is, JSON (JavaScript Object Notation) is a lightweight data-interchange format which is easy for humans to write and read. Moreover, stated that, JSON is simple for machines to generate and parse. JSON is a text format that is totally language independent, but uses a convention that is known to programmers of the C-family of languages, including Python and many others. However, output size depends on the time for retrieving tweets from Twitter [25].

Jonnathan Carvalho, Adriana Prado, Alexandre Plastino works on A Statistical and Evolutionary Approach to Sentiment Analysis. In this study Statistical methods have been inspired by the observation that if two words frequently appear together within the same context. In current statistical approaches, such paradigm words have been selected following different criteria, without any prior evaluation. Motivated by this observation, they propose to classify tweets via a statistical method where the paradigm words are selected by means of a genetic algorithm. This algorithm explores a set of paradigm words to find a subset of such words that leads to a significant improvement of the classification accuracy and appropriate set of paradigm words may vary according to the data domain. One of the objectives is to identify the polarity of individual sentences in the context of newswire articles using, as paradigm words, subsets of the 1,336 adjectives that were manually classified as positive or negative in.
In experiments, they showed that the higher the number of paradigm words, the better the results. They study the behavior of different word similarity measures, for the classification of documents in different domains. In this proposed methods to classify tweets via a statistical method, where the paradigm words are selected by means of a genetic algorithm. When classifying the polarity of tweets, the experiments showed that a more flexible way to select the paradigm words to be used in the classification process outperforms the idea of working with a predefined set, as is frequently done in the literature. The demonstrated experiment that it may be advantageous to vary the set of paradigm words according to the data domain. Tweets of different subjects were more accurately classified with different paradigm words in experiments [26].

Vandana Singh, Sanjay Kumar Dubey works on Opinion Mining and Analysis: A Literature Review Machine learning techniques are the techniques which makes machine to learn from the past experiences and make the machine able to do some pre-defined task. In sentiment analysis, machine learning techniques play an important role Machine learning techniques for sentiment analysis are generally supervised learning approaches. Supervised learning is the approach in which have pre-defined classes as per data. In machine learning, usually need two different sets of documents one for the training set and other for the test set. Training set has the functionality to differentiate the features of dataset. Whereas test set is used for validation of the performance, done by the classifier.

For classification there are number of machine learning techniques, some of them are SVM, Naive Bayes and Maximum Entropy.

There are number of machine learning techniques available like Naive Bayes, Maximum Entropy and Support Vector machine (SVM). Naive Bayes is a simple technique based on the probabilistic model. It is generally used for the document classification. SVM, a deterministic classification technique it is considered the best text classification method. Bayesian method is a simple technique which has the function of the probabilistic model. It is generally used for the document classification. SVM, a deterministic classification technique, it is considered the best text classification method. Support vector machine provides a discriminative platform to distinguish the points of data in to the pre defined classes and then make the decision that relies on the support vectors.
In the starting phase, centroid vector has been calculated and then on the basis of the similarities found document is classified to the class from which the calculated value match SVM gives the maximum marginal difference between the two classes, so it is considered as the most appropriate technique for the classification. In this literature review, the sentiment analysis is done on the social issues and events taking data from the social sites. Sentiment analysis has a wide variety of application in summarizing reviews, classifying reviews, information system, market analysis and decision making. Sentiment analysis is a broad range of fields of natural language processing and text mining. It is found that different types of features and classification techniques are combined in an efficient way to enhance the sentiment classification. The proposed system is able to collect useful information from the social sites, blogs and microblogging sites, and effectively perform sentiment analysis on the data [27].

Zhaoxia Wang, Victor Joo Chuan Tong David Chan Works on Issues of Social Data Analytics with a New Method for sentiment Analysis Social Media Data. The proposed method overcomes the limitations of the existing methods by not only improving the accuracy of the algorithm but also having the capability to perform analysis on non-English languages. Several case studies are included In this study to illustrate the application and utility of proposed method. Author developed a new sentiment analysis scheme to address the above limitations. By using an adaptive fuzzy inference method (FIM) with linguistics processors to minimize semantic ambiguity as well as multi-source lexicon integration and development, new scheme enables us to derive dominant valence as well as prominent positive and negative emotions. The English-based sentiment analysis has shown that this is a viable approach which achieves significant improvements in predictive performance as compared to existing machine learning techniques without the need of a training dataset. For this study, real-world datasets will be used instead of limited historical datasets. In this study combined adaptive fuzzy inference method (FIM) with linguistics processors to minimize semantic ambiguity then merged this combined method with multi-source lexicon integration and development. As a result, were able to derive dominant valence as well as their prominent emotions without the need of a training dataset.
The system can classify sentiments related to public transportation with an accuracy of 88.9% on a pre-classified dataset which was identified by the two domain experts and classified by eight annotators into positive and negative items. Machine-learning methods such as Support Vector Machines (SVM), Naïve Bayes (NB) and Maximum Entropy (MaxEnt) have also been explored with accuracy of 75.0%, 69.4%, and 75.0% respectively for analyzing the same datasets. In this paper, they have outlined several key issues faced by researchers when performing sentiment analysis of social media data [28].

2015 Shokoufeh Salem Minab, Mehrdad Jalali, Mohammad Hossein Moattar works on A new sentiment classification method based on hybrid classification in Twitter. In this study, Hoeffding stochastic gradient descent, has not changed the time while increasing the accuracy. Studies on the tweets that have been randomly selected shows that the proposed algorithm outperforms when assessing the sentiment stream data. Hoeffding’s tree algorithm is one of the best classification methods of flow data. Standard Hoeffding’s Tree can predict majority of the classes in leaves. Add Bayesian algorithm results in more favorable accuracy in facing noise and big data. In this study, they show that replacing the stochastic gradient descent algorithm in Hoeffding’s tree improves accuracy and speed in textual data. The proposed framework to sentiment analysis in twitter has two phases including: preprocessing and online process. First, obtain appropriate amount of tweets which have been detached the English language then in order to analyze them, they use first phase namely pre-processing.

Pre-processing has three main steps that wrapper approach of transformation, filter and wrapper approaches and each one has several parts. In this study, they offer hoeffding's tree of stochastic gradient descent to obtain the best accuracy in the text stream data by prediction gradient descent in each training samples and comparison of prediction with the majority of the class. In this study 2,475 tweets that are randomly selected among 50,000 tweets. Study shows, a new method is given for classifying sentiments on Twitter based on combination classification called stochastic gradient descent Hoeffding’s tree which is a Hoeffding’s tree with stochastic descent gradient in leaves and prediction of gradient descent has good result when there is more accuracy in average compared with majority of class. Evaluation has conducted on tweets and shows the stability of the
Features to Predict Popular Content in Social Media Mining

model compared to other methods that can be used to analyze sentiment in text analysis and management practices should be applied on various topics [29]. Aparup Khatua et al. works on Can #Twitter Trends Predict Election Results Evidence from 2014 Indian General Election. In this study uses a new dataset from Indian context to explore the explanatory power of twitter for predicting 2014 Indian general election. Experiment performs both twitter volume as well as sentiment analysis. Broadly analysis confirms previous findings but it also highlights that a nuanced understanding of research context is extremely important for collecting data. Thus predicting election results is challenging in a large and politically heterogeneous country like India where both national as well as regional parties participate in general election and election is a month long process.

Data Collection: Collected tweets posted by common people, political candidates, and electronic media as well as political parties to overcome the above mentioned bias. The collected tweet data by using ‘twitteR’ and Application Program Interface (API) offered by twitter. The considered data three weeks prior to phase 1 election to the end of phase 9 election i.e. March 15, 2014 to May 12, 2014.

Data Cleaning: they used an exhaustive set of keywords for crawling purpose. As a consequence some tweets, especially if it contains more than one keyword, got extracted for multiple times. Authors apply remove duplicate function on tweet text filed. This way the individual score of each tweet can be calculated and a summation of the same would give a cumulative sentiment score. Conceptually favorable sentiments regarding a political party would be a better indicator of their electoral performance. The contributions are in number of fronts. First, they studied has made an attempt to develop a kind of template for data collection and cleaning which can be used by researchers for similar kind of work.

They have domain driven data mining model has significantly improved data collection and relevant tweet identification process. In this study understand essential for rejecting junk tweets and selecting relevant tweet feeds. Future research should explore how to improve the trade-off between rejecting junk and selecting relevant tweets. Second, in countries like India where regional parties have strong presence in their respective constituency, it is important to consider and control regional dynamics for efficient
prediction. Author has made an attempt in this direction but this can be a potential research area for future studies. Third, findings regarding tweet volume and sentiment score confirm prior studies. It leads to ambiguous outcome when author are using both of them simultaneously in the same model. This requires further exploration in other contexts. Probably the study is fraught with the pitfall of small sample size. Fourth, authors find that significant portion of voters are loyal to a political ideology, and sentiment score can be an effective predictor of vote swing. The empirical evidences support this proposition [30].

Nima Dokooohaki et.al works on Predicting Swedish Elections with Twitter: A Case for Stochastic Link Structure Analysis. In this study make a case for studying evolution of link structures surrounding political tweets to capture the popularity of parties and their respective members. The implementation entices an exploratory analysis comprised of basic and advanced social network mining studies mainly at understanding the dynamics of party’s interaction within and without their neighborhood. They presents an in-depth study of dynamics of party and individual member’s link-structure based popularities along the time-line of two European and general elections during 2014 on Sweden. They validate the resulting popularity estimates using the official statistics of voting outcomes. They consider a combined selection of mentions and retweets as the source of interactions. They studied a link mining approach that leverages the structural features of the interaction network underlying the conversation with politicians during the timeline of two elections.

The presented evidence of how reveals an authoritative structural link formation within which the popularity of the political accounts along with their neighborhoods shows strong correlation with the vote outcome. The public time-lines of two electoral events from 2014 elections of Sweden on Twitter were studied. By distinguishing between individual and official party accounts, The study that estimated popularities reveal strong statistical similarities with vote outcomes also revealed strong ranked dependence between standings of selected politicians and general election outcome, along with official party accounts and European election outcome[31].

Sanchita Kadambari ,Kalpana Jaswal , Praveen Kumar , Seema Rawat works on Using Twitter for tapping public minds, predict trends and generate value. The data sets in the
world are growing at an exploding rate, research and analysis to derive value from this data has gained ground. Social media is a prime contributor to this data most of which is unstructured. The tasks in sentiment analysis cover multiple disciplines like machine learning, NLP and web mining. Tasks are several and include subjectivity and sentiment classification, and object holder and object feature extraction. These task can be performed at different levels of Word Level, Phrase Level, Document Level and Feature Level. The twitter messages are typically short, the granularity of the word level is appropriate for Sentiment Analysis in twitter. Dictionary based and Corpus based approaches are used at this level.

Twitter for sentiment analysis offers numerous opportunities in the field of scientific research, strategic planning, enterprise decisions, market analysis, politics, recommender systems, summarization, reviews, government intelligence, detecting epidemics, innovation, advertising. The twitter data is wide ranged and enormous it offers opportunities to accurately detect, analyze and predict sentiment for marketing strategies, poll prediction, product comparison and planning. The applications are numerous. Substantial work is being done in this field and offers rich opportunities for real-time automated sentiment analysis which would help us to predict trends and generate value [32].

Anurag P. Jain, Mr. Vijay D. Katkar works on Sentiments Analysis of Twitter Data Using Data Mining. This Paper presents approach for analyzing the sentiments of users using data mining classifiers. It also compares the performance of single classifiers for sentiments analysis over ensemble of classifier. Experimental results obtained demonstrate that k-nearest Neighbor classifier gives very high predictive accuracy. Result also demonstrates that single classifiers outperforms ensemble of classifier approach. Sentiments analysis is a method of computing and stratifying a view of a person given in a piece of a text, especially in order to identify persons thinking towards a specific topic product etc. is positive or negative or neutral.

In this study presents a mechanism to predict the overall sentiments inclination of Indian people towards political situation and training tweets are collected by using Twitter API v 1.1. After collecting a raw tweets various preprocessing method get applied to clean the data. Same methods are applied for collecting and cleaning raw tweets for preparing
testing dataset. After preparation of training and testing dataset various classifiers get applied to analyze the performance of classifiers.

Data Collection

Training and testing tweets collected from twitter by using twitter searched API v 1.1 for various political leaders and parties in India. Tweets are of only an English language tweets. Preprocessing Tweets are sometime not in the usable format. To get a tweets in usable format various preprocessing method for cleaning a tweets get applied. All userID, twitterId, user info from the tweets is removed. All special character and hyperlinks is removed from the tweets. Duplicate tweets also get removed from training dataset.

Training data set does not contain retweets. After applying all cleaning methods text that is only with the tweeted text is remained. Training Dataset and testing dataset for classifying a dataset, three classes, Positive, Negative, Neutral are used. To classify tweets in to these category SentiWordNet 3.0.0. Dictionary is used. SentiWordNet 3.0.0.dictionary contains the 117659 word. Each word is assign with its positive negative polarity.

If positive and negative polarity of word is 0 then word considered as neutral. Basically single tweet is spilted into words, after splitting polarity of words is calculated from SentiWordNet 3.0.0. After deciding a polarity of each word all positive and negative words polarity get added separately. And then comparison between positive words polarity with negative words polarity get done. If sentence having more positive polarity then classify sentence (tweet) as positive polarity. Same for negative and neutral polarity.

Duplicates tweets were not considered in a trainingdataset. 2,102,52 tweets were collected about various political leaders and parties. Data cleaning process leaves us with 35% of original tweets.2 lack tweets collected for various political leaders. Experiment performed with following classifier 1) k-nearest Neighbor 2) Random Forest. 3) Naive Bayesian 4) Baysnet. The prediction accuracy of all classifiers when stop words are not removed from training and testing tweets Algorithm Accuracy k-nearest Neighbour 99%, Random Forest 99%, BaysNet 75%, Naïve Bays 60. K-nearest Neighbour and random forest gives much better accuracy compared to naive baysin and bays Net. K-nearest Neighbour gives a highest accuracy. Prediction accuracy of k-nearest Neighbour 99%.
prediction accuracy of all classifiers when stop words are removed from training and testing tweets’-nearest Neighbor gives better accuracy compared to all three classifier. Prediction accuracy of k-nearest Neighbour when stop words are removed is 96.6398%. prediction accuracy when ensemble of classifier is used to classify tweets without stop words removal. Ensemble of random forest, BaysNet and k-nearest Neighbor gives a higher accuracy that is 99%. It can be observed from the experimental results that data mining classifiers is a good choice for sentiments prediction using tweeter data. In this study, k-nearest Neighbor outperforms over all three classifier namely Random Forest, BaysNet, and Naïve Bayesian. Random Forest also gives good prediction accuracy. There is a no need to use of ensemble of classifier for sentiments predictions of tweets as single classifier gives a better accuracy over all combinations of ensemble of classifier [33].

2016 Christos Troussas, AkriviKrouska, Maria Virvou works on Evaluation of Ensemble-based Sentiment Classifiers for Twitter Data. They run the ensemble algorithms, namely Bagging, AdaboostM1, Stacking and Vote, using the four base classifiers. For the validation phase, percentage split method was implemented, in which the 70% of data was used as training set, while the rest as test one. The reason using this method is that ensemble approaches need a satisfactory number of training data, as they are split further for constructing the model. In this study choose the stand-alone base classifiers used as baselines to be compared to the ensemble ones. The performance results have been tabulated and a descriptive analysis has been conducted. All phases of classification were implemented with Weka data mining package.

Comparing stacking and voting, two methods that combine multiple base classifiers, to each individual base classifier, observe that these ensembles outperform mainly the weak classifiers, and C4.5. These ensembles may not necessarily improve the performance of the best classifier in the combination it certainly reduces the overall risk of making a mis-classification, as it is unlikely that all classifiers will make the same mistake. Twitter sentiment analysis constitutes a prominent and an active research area with significant applications in industry and academia. For the classification of tweets, the researchers mainly use stand-alone classifiers, like Naïve Bayes, SVM, decision trees etc.

The combination of classifiers may boost the accuracy of classification. In this study, they try to increase the efficiency of four well known machine learning algorithms using
different ensemble techniques: Bagging, Boosting, Stacking and Voting. A variety of experiments were conducted on three datasets, one with no specific domain and the others with certain topics. The outcomes show that the use of ensembles of multiple base classifiers can improve the accuracy of Twitter sentiment Analysis [34].

Peiman Barnaghi, and John G. Breslin, Parsa Ghaffari works on Opinion Mining and Sentiment Polarity on Twitter and Correlation Between Events and Sentiment. The trained model is based on the Bayesian Logistic Regression (BLR) classification method. They used external lexicons to detect subjective or objective tweets, added Unigram and Bigram features and used TF-IDF to filter out the features. Using the FIFA World Cup 2014 as used case study, In this study used Twitter Streaming API and some of the official world cup hashtags to mine filter and process tweets, in order to analyze the reflection of public sentiment towards unexpected events. In this study examine the effectiveness of a commonly used text categorization method called Bayesian Logistic Regression (BLR) Classification for providing positive or negative sentiment on tweets. They also use extracted Twitter sentiment to look for correlations between these sentiment and major FIFA World Cup 2014 events as used case study.

Analyzing Twitter posts allows the extraction of detailed insights into opinions and trends around sporting events such as the FIFA World Cup, players, teams, etc. and how they change over time during a critical event or after unethical behavior. In this study, a sentiment classification model was trained based on Twitter data using text features. Author extracted sentiment polarity for some major events that occurred during the World Cup using trained model. The experimental results show the positive and negative reaction of people towards such events and how it can change based on incidents during those events. This kind of sentiment analysis helps us to use Twitter data for extracting patterns based on opinionated texts. In addition, teams, players, etc. can receive an overall sentiment in relation to their performance and behavior that could be used to help to improve the quality of matches by highlighting controversial ethical issues as well [35].

Parul Sharma, Teng Sheng Moh works on Prediction of Indian Election Using Sentiment Analysis on Hindi Twitter. They performed data (Text) mining on 42,235 tweets collected over a period of a month that referenced five national political parties in India,
Features to Predict Popular Content in Social Media Mining

during the campaigning period for general state elections in 2016. They made use of both supervised and unsupervised approaches. In this study used Dictionary Based, Naive Bayes and SVM algorithm to build the classifier and classified the test data as positive, negative and neutral. They identified the sentiment of Twitter users towards each of the considered Indian political parties. The results of the analysis for Naïve Bayes were the BJP (Bharta Janta Party), for SVM it was the BJP and for the Dictionary Approach it was the Indian National Congress. SVM predicted a 78.4% chance that the BJP would win more elections in the general election due to the positive sentiment they received in tweets. As it turned out, BJP won 60 out of 126 constituencies in the 2016 general election, far more than any other political party as the next party (the Indian National Congress) only won 26 out of 126 constituencies. It is probable to predict the popularity of any political party and therefore extrapolate their chances of winning the election by utilizing sentiment analysis of Twitter data.

To analyze the problem statement, they obtained tweets, filtering for Hindi language and then applied sentiment mining and prediction operations. They accumulated models including Maximum Entropy, Support Vector Machines and Naive Bayes and concluded that SVM performed the best amongst various others, attaining more precision which lead SVM to be the best performer of all the classifiers. They recorded that all distinctive models were beaten by the unigram model. To gather subjective information, they compile the tweets ending with emoticons. As it is very difficult to predict the results of elections using other methods, including public opinion polls, and with the growing prevalence of social media, such as Facebook and Twitter, In this study to utilize sentiment analysis of Twitter tweets to predict the results of the Indian general election the accuracy of the Naïve Baye's algorithm was 62.1% and the accuracy of Support Vector Machine was 78.4%. Author made final prediction utilizing SVM, since the accuracy of the algorithm is higher. In this study predicted that the party that had a better chance of winning the 2016 general election is BJP [36].

Andy Januar Wicaksono, Suyoto, Pranowo works on A Proposed Method for Predicting US Presidential Election by Analyzing Sentiment in Social Media. US Presidential election is an event anticipated by US citizens and people around the world. In this study two stages in research methodology which is data collection and implementation. Data
Features to Predict Popular Content in Social Media Mining

used in this research are collected from Twitter. The implementation stage consists of preprocessing, sentiment analysis, aggregation, and implementation of Electoral College system to predict the winning party or candidate. The proposed method still unable to use real time data due to random user location value gathered from Twitter REST API, and researchers will be working on it for future works. Every four years, United States (US) held political event to elect a new president. This event known as US presidential election.

The process to run the US presidential election is called Electoral College. This event eagerly anticipated not only by US citizens, but also people around the world. The popularity of predicting the US presidential election has been growing, especially in the academic realms. This paper proposes a method to predict the winning party or candidate in US presidential election in November 8th, 2016. The data gathered from social media will be processing four phases: Pre-processing, Sentiment Analysis to classify the sentiment of tweets by using Binary Multinomial Naïve Bayes Classifier, sentiments aggregation to collect the votes, and implementation of Electoral College to predict the chosen party or candidate. Data are collected from Twitter REST API by applying queries about parties and candidates.

It must be written in English. In this study, data are grabbed from Twitter REST API and must be written in English. The collected data must contain tweet status/text, post time, username, and user location. Tweet status should mention any keyword about parties or candidates that participating in the US presidential election 2016. Corpus is needed to do learning-based sentiment analysis method such as Binary Multinomial Naïve Bayes.

This study using Sentiment140 tweet corpus that contains a total of 1,600,000 data train and 497 data test. To minimize the processing time yet still generating good results, this research limits the data training used to 10,000 positive data and 10,000 negative. As for the data test, 181 positive data and 177 negative data were analyzed to test the performance of sentiment analysis method. Abbreviations dictionary is needed to expand the abbreviations found in tweet status. The purpose of this stage is to classify the sentiments of tweets by calculating the sentiment score using Binary Multinomial Naïve Bayes Classifier. This classifier will generate positive and negative score for each tweet. Sentiment score is the subtraction of negative score from positive score. If the sentiment
score is positive, then the tweet is classified as positive, and vice versa. There is a possibility for classifier to give zero sentiment score, if that happen, and then the tweet is classified as neutral. Researchers have conducted a test using 3 type of sentiment classifiers, i.e. Binarized Multinomial Naïve Bayes Classifier, SentiWordNet, and AFINN-111.

The purpose of the test is to measure precision, recall, accuracy, and F1 score for each sentiment between mentioned sentiment classifiers. Data test used to test those methods are 181 positive data and 177 negative data provided by Sentiment140. The purpose of this stage is to aggregate the sentiments of tweets to decide the winning electors for each state.

The tweets sentiment will be used to represent the vote casted by the user who wrote the tweet status. If the sentiment of a tweet is positive, then the vote is given to the mentioned party or candidate. But, if the sentiment of a tweet is negative, the vote is given to the opposite of mentioned party or candidate. The electors chosen in a state are decided by popularity vote. The winning electors are the one who has the most votes. The electors, who win the popularity vote in its state, must cast their electoral votes to the party that appointed them. The number of electoral votes is distributed according to the allocation of electors for each state. The party or candidate who gets the most electoral votes is predicted to be the winner of US presidential election 2016 [37].

Jyoti Ramtekeet.al. Works on Election Result Prediction Using Twitter Sentiment Analysis They compare the performance of two popular sentiment analysis algorithms, Naive Bayes and SVM. Both the algorithms belong to the category of Supervised Learning. Supervised Learning is a branch of Machine Learning which needs a training data set to perform classification. A training data set comprises of training examples which is basically a pair consisting of an input object and a desired label. The quality of classification is dependent upon the quantity as well as the quality of the training set. While performing classification, machine learning models should be provided with training set which not only has sufficient training using two stage frameworks. This study shows Data Collection from Twitter data for two candidates namely Donald Trump and Hillary Clinton were collected for the dates March 16th 2016 and March 17th, 2016. In this study used the Twitter Streaming API to fetch data relevant to the presidential
candidates and Data Preprocessing and Data Labeling. They used two algorithms, Multinomial Naive Bayes and Support Vector machines to determine the polarity of tweets. SVM algorithm from the Scikit learns package provides the best accuracy for classification [38].

2017 Lei Wang and John Q Gan works on Prediction of the 2017 French Election Based on Twitter Data Analysis. In this study proposes a method to predict election results based on Twitter data analysis. The method extracts and analyses sentimental information from micro blogs to predict the popularity of candidates. The proposed method was used for predicting the result of the 2017 French presidential election. It has been shown that the proposed method significantly outperformed the Tumasjans method, a well-recognized method for election prediction based on Twitter data. In this paper, a new method for election prediction based on Twitter data analysis is proposed and applied to predict the 2017 French Election [39].

Munaza Ramzan, E. Annapoorna, Shikha Mehta works Are Tweets the Real Estimators of Election Results? A medium with a huge amount of information like Twitter can be analyzed for surveying the opinions or sentiments of the public towards future leaders. In this study focuses on predicting the winning party in UP elections with the help of public opinion on Twitter. The paper intends to understand whether tweets can be used as a useful method in predicting the election results or is it just a social hype. Twitter API is used to extract tweets about the UP elections. Public opinion is then ascertained with the help of sentiment analysis on the tweets. After sufficient amount of tweets are collected for analysis, author are simply expanding on positive tweets and getting 45.13% for BJP, the highest percentage as compared to other parties. On comparing this result with the exit polls and the actual election results, the prediction of winning party by the twitter data analysis is correct.

In this study, the aim is to analyze the Indian Election data using Twitter API to extract Tweets regarding elections. As Twitter is a popular micro blogging service where users create status messages. These tweets sometimes express opinions about different topics. In this work, a prediction model is built for elections that would be based on sentiment analysis done on the tweets. This paper focuses on four main things: Extracting relevant
tweets, analyzing the sentiment of each tweet. Storing the tweet and the observations in a NoSQL Database.

In this study, tweets are used to find out which political party is most supported by the public during UP elections. The study has implemented and shown the results of the prediction for UP Elections 2017 which are similar to the actual results. It can be seen that doing sentiment analysis on Twitter data for opinion mining can be used for prediction of election results. Accuracy can be increased if the data is collected just before the commencement of the elections [40].

Hernandez Suarez, G. Sanchez-Perez, V. Martinez-Hernandez, H. Perez Meana works on Predicting Political Mood Tendencies based on Twitter Data. Online social media has changed the way of interacting among users, nowadays, is used as a tool for expressing polarized opinions related to a global or specific context. Valuable information can be gathered in real-time basis and can help to determine if such data has a social impact on users represented as comfort or discomfort on a political domain. Analyzing data related to political domains like government, elections, security and defense and health insurance are important for measuring social mood and predicting whether there is a positive or negative tendency on selected populations.

In this study presents a mood analysis methodology on Twitter data to predict social sentiment on political events. The proposed methodology is done by gathering streams of Twitter’s information, then converted into trained data for processing and classification such that study can statistically predict if there is a positive or negative tendency on political events. Comparing results with historic polls can be achieved by using well known websites like Real Clear Politics, a platform that score public opinion related to political events by comparing online media poll results measured by percentage of favoritism for candidates.

When classification is done author scored daily negative and positive tweets as inputs for a prediction task. The objective of predicting is to find a correlation between users’ mood polarity and a trend which clearly do not establish a threshold for predicting which candidate is going to win elections, but can be useful to observe online behavior towards political issues. For a case study showed that using regularized regression can reduce redundant features and can establish a suitable correlation coefficient between the
Features to Predict Popular Content in Social Media Mining

computed scores and real-time polls. September 2016 was a high competitive month in which there was a minimum difference of positive between candidate preferences and can predict a positive mood [41].

Rodrigo Castro, Leonardo Kuff’o and Carmen Vaca works on Predicting Venezuelan States Political Election Results through Twitter. The large adoption of Twitter during electioneering has created an unprecedented opportunity to capture the citizen’s behavior nationwide. The real-time access to information published by citizens has motivated researchers to design methods in order to enrich traditional political polling with insights from this rich source of data. Less work has been done to capture the political scenario in Latin American countries, given that some methods rely on the use of English words, the reproducibility of such studies in Spanish speaking countries is a challenging task.

In this study a framework in which to apply social network analysis techniques and unsupervised machine learning to infer the political alignment at state level during Venezuelan Parliamentary election, which was performed on December 6, 2015. This electoral process took place in the middle of an acute political polarization in the country the masses were organized around two political coalitions with opposite ideology: Government and opposition. In order to discover automatically the corresponding state political preferences, author analyzes 60K tweets posted within the Venezuelan geographic boundaries during one week before the Election Day. Applying our framework, Author is able to infer a given state of political alignment starting from the quantified differences in communication patterns and linguistic profiles of the state aggregated tweets.

In this study demonstrate that the online political atmosphere reflects the offline tendency at state scale given that author are able to predict the election tendency in Venezuela states with an accuracy of 87.5% with respect to official election results publicly available. During the election period, political preferences reflected in the content published in Twitter inside Venezuela, exhibit clear spatial pattern. In order to predict this behavior automatically, Author propose a methodology to quantify the political alignment embedded in messages authored by residents in each state of the country.

The Proposed methodology produce a text corpus, grouping at state level, tweets posted during the analyzed period. Next, represent the text corpus as a document-term matrix M,
where each row represents an aggregated document, i.e. a set of tweets coming from the same state, each column represents feature and encodes the Term Frequency-Inverse Document Frequency (TF-IDF) score for the with respect to the document. in order to produce a more accurate analysis, Author apply a social network analysis-based strategy to discriminate political features from other non-political ones, starting from communities detected in a word co-occurrence network. The spherical k-means clustering algorithm using the matrix, to find groups of states with the same political alignment. These results are compared side by side with the ones given by the National Electoral Council of Venezuela. Using the Linguistic Inquiry and Word Count (LIWC) software, Attempt to characterize the clusters with additional insights that are not captured by neither the TF-IDF score not by the traditional polls [42]. Tajinder Singh, MadhuKumari, Triveni Lal Pal and Ahsan Chauhan works on Current Trends in Text Mining for Social Media. Online social media has created new paradigms of information sharing which not only provides appropriate platform for the contributors but also for active information seekers. Numerous forms of social media have gained the widespread attention internet users’ almost on explosion level. Availability of data on such behemoth scale mandates regular and critical analysis of this information for various perspectives’ plausibility. As text mining plays a significant and crucial role in discovery of these insights therefore its challenges and contribution in social media analysis must be explored extensively.

In this study focuses on description of assorted social media text mining methods with their viabilities and summarizes various commercial and open source tools for such data analysis. Text mining is a Combination of Machine learning (ML) and the Statistical techniques which are helpful to model and structure the data contents of textual resources for research, e-commerce data analysis and for further investigations or processing. The process of text mining can be visualized as two step progression such as Text refining and knowledge distillation. Information Extraction (IE) helps to extract entities and keywords which are required for further processing and it helps to mine some valuable, knowledgeable pattern from collected keywords to take decision Text mining uses several techniques for the information extraction such as bag-of-words (BOW) model for
document ranking, matching and clustering etc Social Media mining, introduce necessary concepts and principal algorithms suitable for investigating enormous social media data. In this study methodologies from different disciplines such as computer science, data mining, machine learning, social network analysis, network science, sociology, ethnography, statistics, optimization, and Social media mining encompasses the tools to properly represent, measure, model, and mine meaningful patterns from large-scale social media data. It cultivates an inventive kind of data science which is well versed in social and computational theories, specialized to analyze unruly social media data and skilled to help bridge the gap from what author want to know about the infinite social media world with computational tools. Social media helps to advertise the business, news, articles, events and information according to trend all over the world for marketing business perspective Sentiment analysis (SA), Sentiment Polarity Disambiguation (SPD), Event Detection and Tracking, Trend Analysis and prediction, Spammed Profile Detection, Information Diffusion etc[43].

Jaspreet Singh, Gurvinder Singh and Rajinder Singh works on Optimization of sentiment analysis using machine learning classifiers. Words and phrases bespeak the perspectives of people about Products, Services, Governments and Events on Social Media. Extricating positive, negative and neutral polarities from social media text denominates task of sentiment analysis in the field of natural language processing. The exponential growth of demands for business organizations and governments, impel researchers to accomplish their research in sentiment analysis. In this study four state-of-the-art machine learning classifiers viz. Naïve Bayes, J48, BFTree and OneR for optimization of sentiment analysis. The experiments are performed using three manually compiled datasets; two of them are captured from Amazon and one dataset is assembled from IMDB movie reviews.

The efficacies of these four classification techniques are examined and compared. The Naïve Bayes found to be quite fast in learning whereas OneR seems more promising in generating the accuracy of 91.3% in precision, 97% in F-measure and 92.34% in correctly classified instances. The online service providers are hooked on assessing social media data on blogs, online forums, comments, tweets and product reviews.
This assessment is exploited for their decision making of their services or quality of products. The applications of sentiment analysis encompass the areas like social event planning, election campaigning, healthcare monitoring, consumer products and awareness services. The immoderate use of internet by business organizations all around the globe has noticed that opinionated web text has molded to the business plays and socio-economic systems. The computational power is fueled by burgeon of machine learning techniques. This work focused on four text classifiers utilized for sentiment analysis viz. Naïve Bayes, J48, BFTree and OneR algorithm. The “Machine learning techniques for sentiment analysis”. The preprocessing of proposed methodology is limited to extract foreign words, emoticons and elongated words with their appropriate sentiments [44].

2.1 Summary
In this chapter 2 shows the literature related to Sentiment Analysis, Social Media, Data Acquisition techniques, data preprocessing techniques, data classification approaches, data representation techniques and also discuss the results of different classifiers used with twitter datasets. Here total 44 research papers are reviewed and make a literature survey. For this research found the proposed methods from literature survey like Naive Bayes, Support Vector Machine and Maximum Entropy that machine learning approach and Stanford Lexicons and Bing Liu lexicons from Sentiment Lexicons based approach.

2.2 References


Features to Predict Popular Content in Social Media Mining


Features to Predict Popular Content in Social Media Mining


Features to Predict Popular Content in Social Media Mining


[38] Jyoti Ramteke, Darshan Godhia, Samarth Shah, Aadil Shaikh “Election Result Prediction Using Twitter sentiment Analysis” International Conference on Inventive
Features to Predict Popular Content in Social Media Mining


