REVIEW ON SENTIMENT ANALYSIS OF TWEETS BY MACHINE LEARNING

Meena Jindal [1] Khushwant Kaur[2], ANU KAUL[3]
[1, 2 ,3] *Assistant Professor in Computer Science Department*
[1,2,3] *SRI Guru Gobind Singh College, Sec26 Chandigarh*

Abstract- **Sentiment analysis has turned out one of the most significant tools in natural language processing because it opens up numerous possibilities to understand people's sentiments on different topics. The purpose of an aspect-based sentiment analysis is to understand this further and find out what someone is talking about, and whether he likes it or does not like it. A real-world example of the perfect realm of this topic is the millions of available Indian welfare plans. These welfare schemes were launched by the government at all levels in schools, states and centre level. The schemes work with the collaboration of centre and state government. These welfare schemes are introduced for the different levels according to the peoples and their lifestyle. The welfare scheme mostly introduced to develop the weaker and minority section of the society.**

I. INTRODUCTION

The word sentiment represents feeling that can be joyful, confusing, irritating, distracting. The main principle of sentiment analysis is the accuracy of the analysis and to check how much it agrees with human judgment. Precision and recall over the two target categories of negative and positive text are the variant measure on which analysis depends [1] [2]. In general, the utility for practical commercial tasks of sentiment analysis as it is defined in academic research has been called into question, mostly since the simple one-dimensional model of sentiment from negative to positive yields rather little actionable information for a client worrying about the effect of public discourse on e.g. brand or corporate reputation. The concept of sentiment analysis is understood by combining the terms "Senitiment" and "Analysis".

1. Sentiment Analysis

Sentiment analysis is the task of finding the opinions and affinity of people towards specific topics of interest. The sentiments are the feelings based on certain attitudes and opinions rather than facts due to which sentiments are of subjective nature [7]. The sentiment implies an emotion usually motivated by opinion or perception of a person. The psychologists attempts to present multitude of emotions classified into six distinct classes: joy, love, fear, sadness, surprise and anger. The emotions based on sadness and joy are experienced on daily basis at different levels. We are mainly concerned about sentiment analysis detecting a positive or a negative response or opinion.

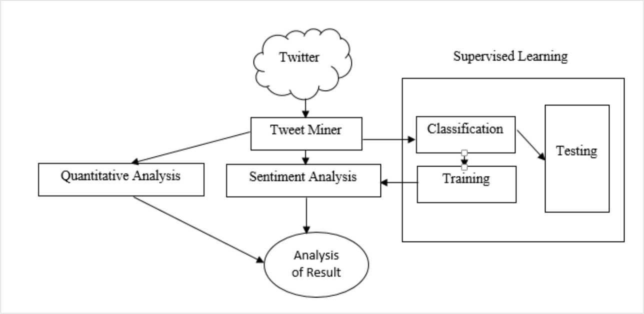

Figure.1 Twitter-based Sentiment

In twitter-based sentiment analysis, tweets generally contains spelling mistakes and problem of character limitation resulting in abbreviation mistakes. Many unconventional methods on linguistic basis like words elongation or capitalization are used. In addition, tweets consist of unique features like hashtags and emoticons having an analytical value. It is seen through research that 70% accuracy in classifying sentiments. Computer system will make very different errors than human assessor. Computer system will has the trouble with negation, exaggeration, jokes or sarcasm which is easy to handle for a human reader.

1.1 Types of Sentiment Analysis

(1) Comparative Sentiment Analysis: The main goal is to identify the opinion from the comparative sentence. For e.g.: "I drove the Verna, it does not handle better that Honda city superb."

(2) Document-Level Sentiment Analysis: In this type of sentiment analysis is performed on the entity on which and identify the positive negative view on the single entity by using the documents [9] [10] .

*(3) Aspect-Based Sentiment Analysis***:** Document level and sentence level analysis gives good results when they are used on single entity but when we want to analyze the multiple entity then we need aspect based sentiment analysis. For eg: "I am a Samsung phone lover. I like the look of the phone. In the aspect based sentiment analysis it allows us to analyze the positive and negative aspect of an item. This type of analysis is mainly domain specific. In this analysis firstly aspects are identified and then location of the aspect in the review.

1.2 Online Social Media

In the 1990s, clients can make and put their contents to show up by very well-known broadband Internet. In 1997, first social network website (SixDegrees.com) appeared. From 2002 onward, a many informal organizational locales were propelled. For e.g., some are like Friendster others are like MySpace. Later in the 2000s, online networking increased broadly and also quantities of clients. Various elements included into online networking interest. These incorporate accessibility of broadband connection, new programming devices, and development of more Personal Computers and cell phones; other developed elements are social and financial, by fast adoption of online networking by young generation and the expansion of Personal Computers, programming, commercialization in social media sites respectively. In the 21st century with the advancement of Web 2.0 Internet, social media platforms, for e.g. websites begin to allow clients to connect and work together in virtual groups. Rich client experience, openness, aggregate insight, adaptability etc are newly added attributes in online networking. Clients get additional facilities of ―like‖, make and post pictures, upload audio, video content on the online environment. The data posted can be shared with some selected clients or freely over the web. By the social media [6] has become quickly around the globe because of its mixing of innovation and social association for the co-making of worth.

1.2.1 Types of Social Media: Different types of social media are as follows [5] [7]:

(1) Wikis: Wikis offer an effective yet adaptable communitarian specialized apparatus for creating content-particular Web sites. Since wikis develop and advance as an immediate consequence of individuals adding content to this, further leads to an assortment of pedagogical necessities—understudy association, bunch exercises or group activities, and so on.

(2) Blogs: The significance of the blog is that it is dynamic. It can be redesigned and it permits the guest to impart through the remark segment joined to every individual post.

(3) Social network sites: Social network sites are updated continuously and adding new devices, for example, photographs /video /sharing and blogging.

- *Status-update services:* Also called micro blogging administrations. Micro blogging is a broadcast medium that exists in the form of blogging. A microblog varies from a conventional site that its content is little in size [3]. Microblogs allow users to use little texts of the substance, for e.g., "small tenses, singular images, or video joins".
- *Media-sharing sites:* Media sharing has made it workable for people and organizations alike to grow their impact and reach. These regions license customers to post recordings or photographs. It permits clients to post recordings or photographs on YouTube, Pinterest etc. Users can then share that media with the world or just a select group of friends.
- *Virtual world content:* A computer based simulated environment MMOW is massively multiplayer online world used to explore the virtual world. This is like a site offering a game environment. Imaginary universe is one of the examples of virtual world in which users create avatars interacting with each other.

1.3 Role of Twitter

Twitter is likewise a substantial long range interpersonal communication micro blogging webpage. The monstrous data gave by twitter, for example, tweet messages, client profile data and the number of devotees/followings in the system assume a huge part in information examination, which consequently makes most examinations explore and look at different investigation procedures to get a handle on the ongoing utilized innovations [8] [13]. Retweet in twitter is the assertion activity to a particular tweet, as now and again the client passes data to his/her gatherings of people to express their feeling on a specific tweet. The retweet rate of the first tweets and the number of notices identified with those tweets to research whether the quantity of retweet and number of notices are identified with a similar system.

1.4 Methodology of Sentiment Analysis

The methodology of sentiment analysis is discussed as follows:

(1) Labeled Data**:** SemEval 2014 was also responsible for the labeled data-set development used for experiments internally and deep analysis. Such kind of data set contained a total number of 1,655 tweets with a distribution of class analogous to the data training set [10]. In order to identify better algorithm performance and the features often very useful, the developed data was used which forms a standard practice in the field of machine learning analysis. More often, this data was used by many researchers and the adjustment of the system to perform well over the data-set was done. The dataset collected was annotated on Mechanical Turk i.e. an internet market place by hand. The organizers provided guidelines to annotators considering positive, negative or neutral behaviour in order to reduce the subjective nature at sentiment annotation. The publishers stated that the class distribution was reflective of the practical world tweets and highlighted that the set of data was cleansed from any kind of exception [6]. One such example of tweet labelled sentiment was published over the website as represented below:

Table.1 Training data-set Class Distribution

Labelled Sentiment	Share
Positive	38%
Neutral	47%
Negative	15%

2. ***Pre-processing****:* The process is initialized using the following steps. Such steps convert the plain tweet text into the elements of processing nature with an additional information utilized by the feature extractor [13]. This process was done before the tweet-based usage of feature extractor in order to design or build the feature vector. The tools of third party were used for all the steps specifically handling tweeted text unique nature.

Step1: Tokenization: It is a process of text conversion as a string into elements process-able known as tokens. In terms of tweets methodology, such elements can be emoticons, words, links, punctuations or hashtags. As shown in fig.3, "an insanely awsum time..." text was busted into "an", "insanely", "awsum", "time"…..

The elements here get separated by some space whereas the sentence based punctuation ending such as full stop or exclamation mark get separated more often by a space. The hastags along with symbol "#" that precedes the tag is required to get retained as the symbol "#" may suggest distinct sentiments than the word to be used regularly in text. Therefore, the Twitter-based particular form of tokenizer helps to extract tokens.

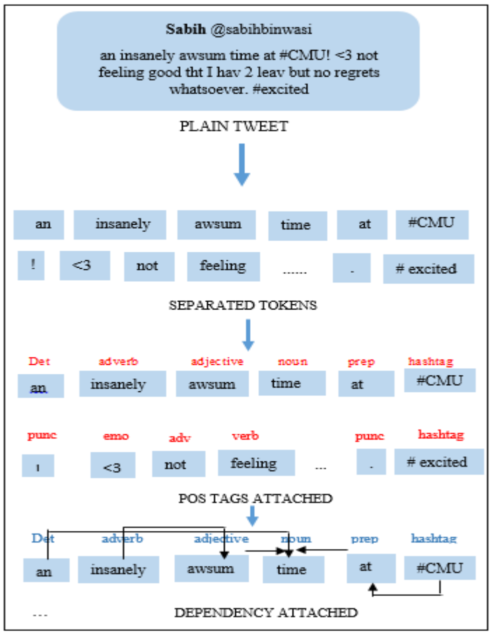

Figure.2. Pre-processing stages

Step 2: Parts of Speech Tags (POS): The POS tags are generally characterized by sentence based words dependent over the different categories of grammar in context to a word. Such an information or data is necessary for the process of sentiment analysis as words contain distinct values of sentiment that are basically POS dependent [2]. Considering an example, the word "good" acting as a noun has no sentiment whereas "good" in its adjective form reflects a sentiment in a positive way. Fig.2 above shows that each token gets extracted in its last step and gets assigned a POS tag. The accuracy maintained by a POS tagger is about 93%.

Step 3: Dependency Parsing: It represents the relationship extraction among sentence-based words. Such a method is very useful for identification of relationship between "good" and "bad" in form of phrases like "not really good" where there is only adjacent word relationship. It explains the parent-child relation between tweet tokens as shown in figure above. The accuracy maintained by dependency parsing is about 80%.

***(3)* Feature Extraction:** This is a process of designing a feature-vector from a known tweet. The entry (each) in case of feature vector is an integer that contributes on the attributing a class of sentiment to a tweet class [1] [8]. Such a contribution strong to negligible form. The algorithm here identifies the strength dependency between classes.

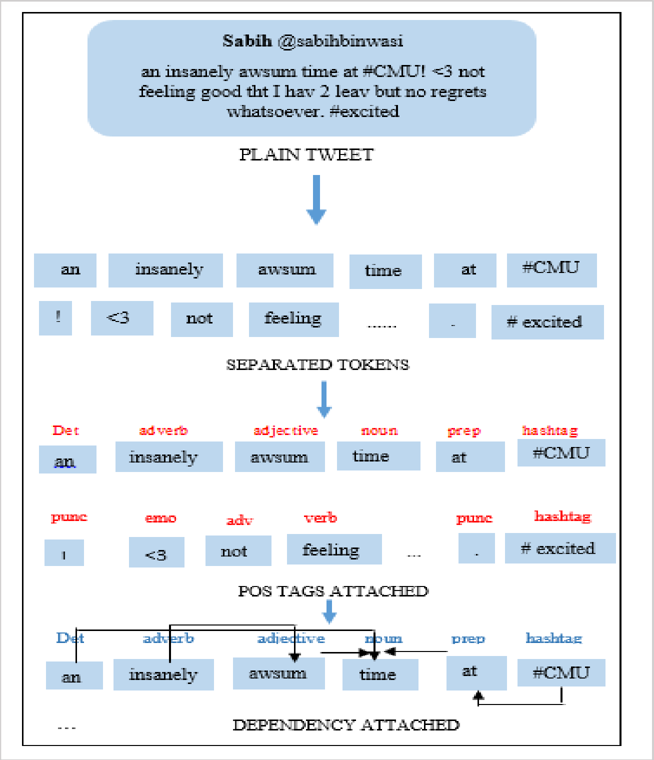

Figure.3 Feature Extraction

(a) Feature Set-Bag of Words: As shown in figure above, the order of token-based sequence or the structure of grammar is not at all preserved [12]. For instance, the token "awsum" forms a part of tweet hence it is marked or assigned as "1" whereas the token "hater" does not occur in tweet, hence it is labelled as "0". The token "hater" has a column for it in a feature vector. Hence, it would take place in some tweet in the trained data-set form. It was further analyzed that the indication of only word presence yields good performance than the word-based frequency. In such case, an entry is also assigned for specific ordered token pairs termed as bi-grams. The token pair "insanely awsum" where it is assigned or labelled as "1" if it forms a part of tweet, otherwise it is considered to be "0". It indicates that the system is equipped not only to indicate the token presence but also indicate its context.

(b) Feature Selection: The process notices that some of the features are not relevant for the operation of sentiment analysis, hence these un-necessary features are required to be removed. A feature attribution evaluation was conducted to analyze the impact of features. A very popular method named Chi-Squared Feature Selection was used which carried a classification algorithm evaluating the feature-based dependency value and the dependency of each of the class.

(c) Social Media Feature Set: Emoticons are used as symbols basically to express the gesture or feelings using language characters and punctuation. These emoticons strongly analyses the positive or negative sentiment.

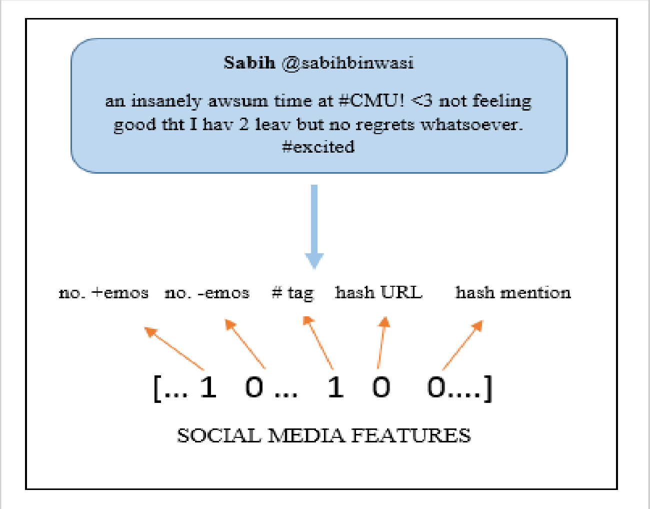

Fig.4 Social Media Feature Set

The figure above shows a tweet sample illustrating the social media feature set representing a number of positive emoticons labelled as "1" as per the dictionary while no such emoticons are found. In feature-based vector, the presence of hashtag, URL and mention ("@username") is also included. Such types of features are mostly used in the showing the differentiating impact between the polar i.e. positive and negative and non-polar i.e. neutral class. These type of indicators helps to identify the relation between sentiment class and the indicators itself [11] [14].

(d) Lexical Feature Set: These are basically driven by the use if lexicons. The task of sentiment lexicon analysis maps the n-grams or tokens to score-based polarity. These have been reported successfully with an ability to locate issues based on classification methods of sentiment analysis. As shown in figure below, Twitter having a range from -5 to +5, where, -5 is labelled as a negative type of token and +5 is considered as the most positive type of token. The tokens which does not contain any kind of sentiment are labelled as "0". Each of the tweet token is labelled independently in a polarity score. For instance, "time" is not labelled any type of score.

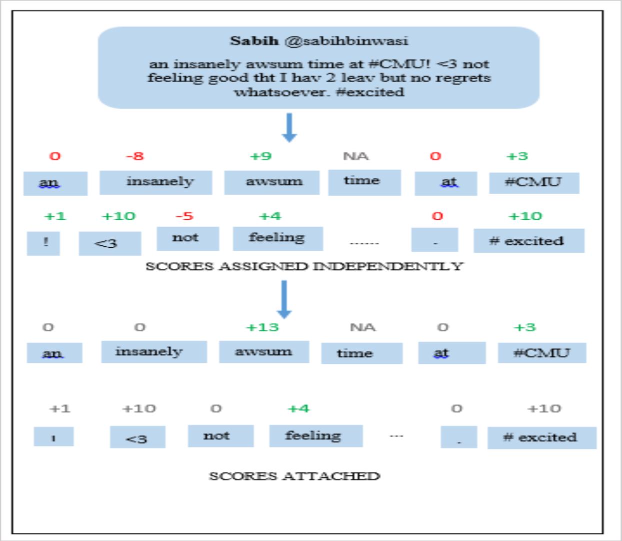

Figure.5 Lexical Feature Set

The lexical based feature set are subdivided with the following features:

Handling Negations: As shown in the figure above "good" is labelled with positive type of score (+8). However, in case of "not feeling good", the "good" must be labelled or assigned a score i.e. negative. This creates an issue of negation context. The tokens "insanely" and "awsum" denotes a negative and a positive score as shown respectively. If one considers "insanely awsum", then one cannot used them in independent form but it intensifies the sentiment expressed by "awsum" which denotes intensifier context problem.

1.5 Data Mining Extraction Process

The vital elements of a classic data mining system [11] [12]are as follows and shown in Figure 1.2.

(1) Knowledge Base: For entire data mining process, the knowledge base is very helpful. It may be search or evaluate the patterns for results. To generate more reliable and accurate results the inputs can be gained by the data mining engine from the knowledge base.

(2) Graphical User Interface: This module intercommunicates between client and data mini framework. This module provides the better utilization of the framework effectively and productively extracting the real complexity of the procedure.

(3) Pattern Evaluation Modules: By utilizing a threshold value the measurements of interestingness of the pattern are done by pattern evaluation module. To focusing on the search for finding interesting patterns this module associates with the data mining engine.

(4) Data Mining Engine: It is the core part of any data mining system. Numerous functions of data mining additionally classification, clustering, time-series analysis, association, characterization, prediction and so forth are performed by various modules of mining engine.

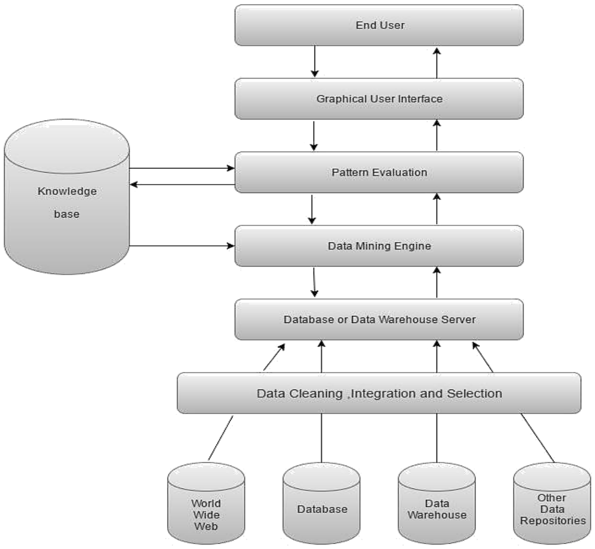

Figure.6 Architecture of Data Mining

(5) Database or Data Warehouse Server: It contains the realistic data that is to be processed. Hence, considering the data mining request of the user then relevant data is obtained by a server to process the request. First data required cleaning and integration. After that, a larger number of data more than required gathered from various data sources from which the data of interest should be chosen and passed to the server. As a component of cleaning, integration and selection of data various techniques may be performed.

(6) Database, Data Warehouse, World Wide Web, or other Information Repositories: These are the real sources of data. For effective data mining, we require large volumes of historical data. A large amount of data is usually stored by organizations in databases. Warehouses may contain many databases, content records, and different sorts of data archives spreadsheets.

III. RELATED WORK

Balachandran et al [1] outlined the today's scenario about choosing the right institute is the most challenging task. To get the estimated idea about the particular institute every student surf over the social network sites for the reviews, ratings about the particular institution. But it is difficult to analyze the statistical aspect from the reviews. In this Aspect based Sentiment Analysis is directly implemented on the reviews which gives us negative and the positive reviews of the particular institution. The various techniques are used for the aspect identification such as NLP-based technique, Machine Learning based (ML), unsupervised approach, Dictionary based, Corpus based. The best possible result analytic is given by the NLP and the ML classifier to classify each aspect into their respective category. Mohammad et al. [2] worked on the multimodal sentiment analysis which is related to the images, text audio and video those are posted by the users. There is a lack of proper method for this type of analysis so the proposed work reviewed the different approaches related to this work which helps the future researchers to use effective approach. The gaps in the approaches and their opportunities are discussed in detail. Hagge Marvin et al. [3] has described the opinion of the consumers by the micro blogging service i.e. Twitter. Twitter is social media on which each and every person expresses their views. Consumer opinions are the base of analyzing the consumer perception about the individual product. On this consumer view do aspect based sentiment analysis by Part-of-speech tagging and, parsing dependency from Natural language processing in return its extracts the positive, neutral, negative aspects from the tweets. In proposed approach software toolkit is designed in such a way that it first extracts the tweets, then filter afterwards analyze the sentiments polarity and then displays the result. In this people can rent out their homes to one another over the web platform. The aspects in consideration are airbnb, place, time, home, day, people, night, view, apartment, room, for analysis. Results are displayed via graphs. In future it will works on the reviews of Airbnb website. Pannala, Nipuna Upeka, et al. [4] described the existing work of the opinion mining are done on the word level, not on the sentence level. This finds the explicitly expressed opinions. The proposed paper works on the trained data set that analyzes and gives the positive, neutral and negative reviews for different products. The Aspect based sentiment analysis (ABSA) works on the different aspects of the entity and which in returns shows the polarity. To implement ABSA machine learning (ML), and Natural language (NL) techniques are used. The dataset used in this proposed paper have 1654 aspect category annotations in the training dataset and 845 aspect category annotations in the test dataset. The performance of software is measured by SVM and logistics regression algorithm. KeumheeKang et al. [5] proposed a novel method for identifying the users with depressive moods by analyzing their daily tweets for a long period of time. They exploit all media types of tweets, i.e., images and emoticons as well as texts. To assess the validity of the proposed method, two types of experiments were performed: 1) the proposed multimodal analysis was tested with a number of tweets, and its performance was compared to SentiStrength; 2) it was applied to classify 45 users' mental states as depressive and non-depressive ones. The experimental results confirmed that the proposed multimodal analysis method has the higher accuracy than existing methods and it can predict individuals' moods more efficiently. Wei Gao, et al [6] used suboptimal approach for sentiment classification. The previously introduced approach does not solve the problem of quantification but the proposed approach solves it by using learning algorithm. The result shows that the quantification gives better class frequency for effective sentiment analysis.

Table.1 Existing Scheduling Model.

Author's Name	Year	Methodology Used	Proposed Work
Balachandran et al.	2017	Machine Learning (ML), unsupervised approach, Dictionary based, Corpus based approaches.	Outlined the today's scenario about choosing the right institute is the most challenging task. Sentiment Analysis is directly implemented on the reviews which gives us negative and the positive reviews of the particular institution
Hagge Marvin et al.	2017	Twitter Sentiment Analysis	Described the opinion of the consumers by the micro blogging service i.e. Twitter.
Wei Gao, et al	2016	Learning Algorithm	Used suboptimal approach for sentiment classification.
M. Abdul-Mageed et al.	2011	Standard Arabic data for sentiment analysis	Worked on the standard Arabic data for sentiment analysis. In this work data set is collected and then automatic classification step is started in which tokenization is done on the data.
MalharAnjaria et al.	2014	Machine Learning Approach	Introduced the concept of supervised learning for sentiment analysis on social networking blogs.

Rongrong et al. [7] has worked on the visual sentiment analysis approaches. In this a survey is presented which defines the different techniques used for the visual sentiment analysis. In this type of analysis images are used to predict the sentiments of the person. The survey basically focused on the cutting edge methods that are used in image analysis process. This survey describes the new platform for researcher because mainly research is done on the text but visual sentiment ontology is a new concept for doing something different. Fir effective visual sentiment analysis the concept of deep learning is useful in it. MalharAnjaria et al. [8] introduced the concept of supervised learning for sentiment analysis on social networking blogs. The proposed work is done by using the different machine learning approach like SVM, nave Bayes and artificial neural networks. These are implemented on the unigram, bigram and hybrid features. The data collected from the tweeter is related to the U.S election and Karnataka election to predict the users opinion related to their polls. The Support Vector Machine gives the better accuracy. R. Socher, et al. [10] worked on the sentence level prediction of label distribution by using a new approach which is based on the sentence level prediction in recursive auto encoders. The proposed work is done on the sentiment lexica and sentiment shifting rules. The dataset used in this work is personal user stories which annotated with multiple labels and aggregated from multinomial distribution which captures the emotional reactions. M. Abdul-Mageed et al. [11] worked on the standard Arabic data for sentiment analysis. In this work data set is collected and then automatic classification step is started in which tokenization is done on the data. The two stage classification process is performed on the data set. The results of the proposed approach show the effectiveness of the approach.

V. CONCLUSION

In this study, the concept of Support Vector Machines (SVM) is used for classification of algorithm with binary classification process. Such type of method helps in analysing different feature vectors with an assigned class in order to identify the relation dependency between a sentiment and each of the feature. Here, each of the vector is considered as a point of data in vector dimensional space that equals to the size of feature-set. The SVM helps in identifying the vector dimension based hyperplane which divides the class into two types. One is the considered as "best" i.e. defined as a good type of separation gained by the hyperplane having the large distance to the point nearest to the training data type of any kind of class known as functional margin. In general, if the margin is large then the classifier error gets reduced. When the new form of tweet i.e.

un-labelled is fed into the system, it helps in extracting the feature vector same as that of labelled tweets we have shown that our proposed approach of improving the performance of hybrid SVM and KNN for aspect-based sentiment analysis by using unigram features. It was also shown, in direct comparison that hybrid approach considerably better. In the results it can be an inspiration to use additional data. In thesis analysis, review of Indian welfare scheme by its aspects and several aspects of sentiment analysis are existing. The breadth of this study can lead to more general view and better understanding of sentiment analysis. In the experiment KNN with SVM shows effective precision and accuracy than other approaches.

VI. REFERENCES

[1] L. Balachandran and A. Kirupananda, “Online reviews evaluation system for higher education institution: An aspect based sentiment analysis tool,” *2017 11th Int. Conf. Software, Knowledge, Inf. Manag. Appl.*, pp. 1–7, 2017.

[2] M. A. Ullah, M. M. Islam, N. B. Azman, and Z. M. Zaki, “An overview of Multimodal Sentiment Analysis research: Opportunities and Difficulties,” *2017 IEEE Int. Conf. Imaging, Vis. Pattern Recognition, icIVPR 2017*, 2017.

[3] M. Hagge, M. Von Hoffen, J. H. Betzing, and J. Becker, “Design and implementation of a toolkit for the aspect-based sentiment analysis of tweets,” *Proc. - 2017 IEEE 19th Conf. Bus. Informatics, CBI 2017*, vol. 1, pp. 379–387, 2017.

[4] N. U. Pannala, “Supervised Learning Based Approach to Aspect Based Sentiment Analysis,” 2016.

[5] K. Kang, C. Yoon, and E. Y. Kim, “Identifying depressive users in Twitter using multimodal analysis,” *2016 Int. Conf. Big Data Smart Comput. BigComp 2016*, pp. 231–238, 2016.

[6] W. Gao and F. Sebastiani, “From classification to quantification in tweet sentiment analysis,” *Soc. Netw. Anal. Min.*, vol. 6, no. 1, 2016.

[7] R. Ji, D. Cao, Y. Zhou, and F. Chen, “Survey of visual sentiment prediction for social media analysis,” *Front. Comput. Sci.*, vol. 10, no. 4, pp. 602–611, 2016.

[8] M. Anjaria and R. M. R. Guddeti, “A novel sentiment analysis of social networks using supervised learning,” *Soc. Netw. Anal. Min.*, vol. 4, no. 1, pp. 1–15, 2014.

[9]Y. Lin, J. Zhang, X. Wang, and A. Zhou, “An Information Theoretic Approach to Sentiment Polarity Classification,” pp. 35–40, 2012.

[10] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, and C. D. Manning, “Semi-Supervised Recursive Autoencoders for Predicting Sentiment Distributions,” *EMNLP 2011 - Conf. Empir. Methods Nat. Lang. Process. Proc. Conf.*, no. ii, pp. 151–161, 2011.

[11] M. Abdul-Mageed, M. T. Diab, and M. Korayem, “Subjectivity and Sentiment Analysis of Modern Standard Arabic,” *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol.*, vol. 27, no. 1, pp. 587–591, 2011.

[12] P. D. Turney, “Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews,” *Proc. 40th Annu. Meet. Assoc. Comput. Linguist.*, no. July, pp. 417–424, 2002.