Social Media Dataset Sentiment Analysis Using Machine Learning Algorithms

Harjeet Kaur[1], Dr. Rakesh Kumar[2], Parekh Sharma[3]
[1]Research Scholar, [2]Principal, [3]Assistant Professor
[1,2,3]Department of Computer Science and Engineering
Sachdeva Engineering College for Girls, Gharuan, Mohali (Punjab)

Abstract-Computer vision applications and the internet altered the way of people interacting. Currently, these services are demanded dramatically. The interaction takes place through online chats such as social media (Twitter, Facebook, Whatsapp and so on). People became dependable upon the user generated content. Therefore, sentiment analysis is approaching under the consideration of the recognition and categorization of human expressions. It simply explains the opinions in a source text. The major applications are business companies, where the owners and managers want to review the opinion of customers to make the best interaction and relationship. For the text mining, sentiment analysis is a better way that has the tendency to tackle the computational issues in the text. In this research work, a deep survey is performed on the various categories of sentiment analysis data mining techniques. A brief description is given to the concept of sentiment analysis that gaining a lot of attention. The process of the detection, extraction and classification of the opinions are explained. The major areas, where it is applied are the political field, market intelligence, customer feedbacks, user satisfaction, movie sale production and so on.

Keywords-*SA (Sentiment Analysis); SVM (Support Vector Machine); SSTB (Stanford Sentiment Tree Bank); STS (Stanford Twitter Sentiment).*

I. INTRODUCTION

The emotions are playing a pivotal role in the interaction of human beings. Actually, the emotions outweigh IQ for better communication. Nowadays, the emerged artificial intelligence makes the sentiment analysis as key terms. Additionally, it has applications in various companies that are small or big. The companies considered sentiment analysis as a crucial mission that based on the emotions. SA (Sentiment Analysis) is also used in other application fields to improve the tendency of communication among customers and for the recommendation systems. The mining sentiments are obtained from public are raised for the upcoming time of the web services. The basic sentiments are social relationships (Websites), social functions (Applications and Widgets), social colonization (Shared Ids), the social context passed on websites and social commerce [1].Generally, SA is an emerging concept in the artificial intelligence which follows up the amalgamation of natural language processing and machine learning. The purpose of SA is to create the new text forms, new tasks and new language with features [2]. SA has the capability to recognize and categorize the opinions and sentiments in the particular source text. It is necessary to capture the opinion of users that helped in various situations. The internet changed the way of interaction. Now it’s done by the use of social posts, online discussion, forums, product reviews and on websites. People rely on the user given content.

The fundamental methods of analysis the sentiments from users are done by symbolic approaches and machine learning approaches. Symbolic approaches are commonly considered to make the availability among the lexical resources for users. In this way, the bag of word method is used. The other approach machine learning utilized to make a use of the training set and a set of text for the classification of data. For the classification of text, several kinds of methods are accessed such as Naïve Bayes Classifier, SVM (Support Vector Machine) Classifiers, maximum entropy and ensemble classifier [3].

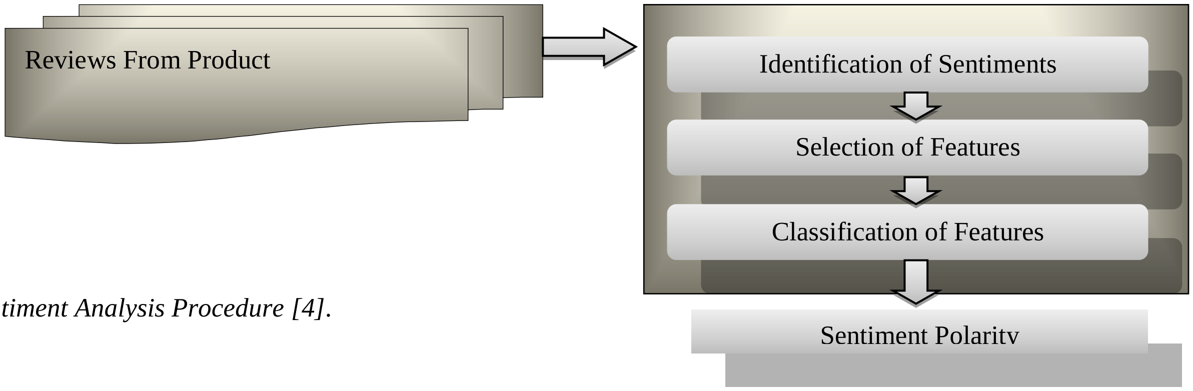

Fig.1: Sentiment Analysis Procedure [4].

In sentiment analysis several kinds of analysis are performed like document level, word level, aspect level, sentence level, concept level, phrase level, etc. Sentiment classification is the evaluation of the orientation of a particular text in two segments or classes. Binary, trinary are the classes for sentiment analysis ,which formed in the stars and thumbs [6].

A. Challenges in SA (Sentiment Analysis)

Some challenges occurred in the sentiment analysis and these are required to be removed to enhance the process of data extraction and analysis. The common challenges are mentioned in the following section.

1) *Increment Method:* The analysis of real time data is not a simple task and not applicable to be performed at one time. When data are collected, it is necessary to analyze it. The increment process upgrades the new individual data instances without accessing to the previous data and after passing time, the old data is not available for the data modification.
2) *Computation of Massive Data:* If the data is partitioned in the small tasks and processes, then the data execution worked simultaneously and there must be an occurrence of enhancement in the processing speed. Therefore, it is essential to process the massive and complex data to parallel process.
3) *Behavior:* The behavior, homophily on the social networking determined by the traces left on the social networks. It is difficult to analyze the social media and patterns to verify and validate them. The evaluation regarding this becomes more complicated.
4) *Sarcasm on Sites:* The sarcasm used to hurt and offend which commonly preferred for the comic effects. Its support to the false positive for children really brightens up a household. They never off the lights. The detection of sarcasm from the expressions and to correct the context is a difficult task and required for the sentiment analysis.
5) *Grammar Correction:* For the sentiment analysis, it is a fundamental requirement to remove all the grammar mistakes from the content to make it more effective for the analysis of expressions. The most common grammatical errors are expelled and the improvement done over the given content.

Other challenges are related to lexicons, spell checking, refinements of content, managing noise, deletion, addition and updating of lexicons in the given content [6].

The research work is partitioned into several sections. The content in sections is organized according to the relevant title. First of all, Section I, relating to the introduction details of the sentiment analysis, which also get together with the process of it and some challenges that are needed to be diminished. Section II is about the previous work performed in the same criteria. The used techniques, results and drawbacks are discussed in this section. Section III, it is created to describe the procedure to organize the datasets by sentiment analysis. Section IV is considered to explain numerous data mining methods that are utilized for the sentiment analysis. The last section V is the overall description of the survey concluded and further work.

II. LITERATURE REVIEW

Sindhu, C., et al., (2017) [7] proposed a deep research on sentiment analysis, which based on the product rating on the basis of text reviews. Due to the advancement of technology, the extreme amount of data was present on the internet. Mostly, data were generated by social media such as Facebook, twitter and whatsapp. To manage these kinds of data, the new automated analysis was proposed as known as sentiment analysis. Before the initialization of sentiment analysis, it passed through several pre-processing techniques. The determination of true, false or positive negative was completed by the open source data tool. It was known as a miner. Consequently, the relative study was done, which related to support vector machine and Naïve Bayes.

Akhtar, M.S., et al., (2017) [8] worked on feature selection procedure which depends upon the aspect based sentiment analysis. In this research work, the principal attention was on the design of cascade of feature selection and classifiers. The premeditated technique was PSO (Particle swarm optimization). The aspect based SA (Sentiment analysis) was operated in two steps. Aspect extraction and classification were the two fundamental steps on which SA was performed. Further, the features were used for the identification on the properties of various classifiers and its domain. Basically, three kinds of classifiers were used as maximum entropy, conditional random field and support vector machine. The new generated outcomes were more effective.

Pham, D.H., et al., (2018) [9] described the multiple layers of data representation for the aspect based SA. In general terms, sentiment analysis was an automated process to discover the sentiment ideas about a specific product from the customer textual comments. In this research work, a dynamic multilayer framework was described to represent the customer reviews. It was predicted that, the entire sentiment for a product was consists of the sentiments of an aspect. Each of the aspect was expressed as a relative sentence and it was composed of words. The multilayer framework was gradually converted into neural network basically to obtain a model for prediction of product ratings. Subsequently, the learning approaches were included the word embedding and vector models along with BPNN (Back propagation neural networks). This proposed model accessed the aspect ratings and its relative weights. The experiment was completed with a huge data set of hotel domain review.

Al-Smadi, M., et al., (2017) [10] proposed a deep recurrent neural network with support vector machine for the sentiment analysis. State of the art techniques and supervised learning methods were represented in this work. The techniques and machine learning methods were associated to solve the issues of aspect based SA. The research work was performed on Arabic hotel reviews. For research, the focus was on the reference dataset of Arabic hotel's reviews. The outcomes proven that support vector machine was better as compared to recurrent neural networks. When focused at the implementation time for both training and testing the recurrent neural networks implementation time was quick and performed given tasks in seconds.

Geetha, M., et al., (2017) [11] explained the relation of customer sentiments and online customer ratings. A relationship was established between the customer sentiments on online reviews and ratings for hotels. The sentiments of customer composed of emotions on reviews. The sentiments could be positive or negative. The study articulated the sentiments in the customer sentiment polarity. The topmost determination was to search the more consistency among customer rating and real customer feelings for hotels. The customer polarity represents the significant alterations in customer rating. It was proven from experiment that when the premium hotels were compared, then the managers of budget hotels must train and enhances the performance of staff along with the services.

Table.1: Different Techniques And Drawbacks

Author Name	Year	Technique Used	Dataset Used	Issues /Gap
Sindhu, C., Vyas, D. V et al.,	2017	SVM and Naïve Bayes	Product reviews	False Positive high
Akhtar, M. S., Gupta, D., Ekbal, A. et al.,	2017	PSO: Particle Swarm Optimization	Amazon Dataset	Sequence Labeling problem
Pham, D. H., & Le, A. C.	2018	Back propagation algorithm	Hotels crawled	Polarity classification, subjectivity classification, and opinion spam detection
Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y et al.,	2017	Supervised Learning Approach	Arabic Hotel Public review	Multilevel and error rate high
Geetha, M., Singha, P., & Sinha, S	2017	Naïve Bayes`	Customer Reviews (Hotel)	-

III. DATASETS

The selection of data sets in SA is a difficult issue. The basic source of data gathering is completed from the product review data. The product reviews are considered as a crucial information for the business holders or company owners. The decisions about the business are occupied on the basis of product review and customer feedbacks given on the business site. Firstly, they analyze the reviews and after that, they take some serious decisions and make decision that are beneficial for all users and employees. The reviews are in other forms like a review sites and it applied mostly in stock markets. Social sites and election results are also seen on the online sites [4].

To create a dataset by sentiment analysis, first of all, create a ChaSM basically for the different corpora through two specific domains as a movie review and twitter posts. SA of small text messages as a single sentence and twitter messages are challenging due to the restrictions over the information which normally present in the small messages. The reviews of movie dataset were utilized in SSTB (Stanford Sentiment Treebank). It's composed of fine segments labels, particularly for 215,154 phrases in the trees of 11,855 sentences. In the research, the focal point was the predictions about the sentiments to finish a sentence. In the second corpus, the use of STS (Standford Twitter Sentiment) corpus operated in 2009. The main training set consists of 1.6 million tweets which were the automatic labels and must be true or false with the use of emotions (Noisy labels). In the experiment, only use of a single training data which was composed of 80K tweets. Eventually, a construction of development set was performed that choose the 16000 tweets from the training set. The detailed description about the corpus is shown in the table 2. [12].

Table.2: Dataset Obtained By Sentiment Analysis

Sentiment Analysis Datasets	Data Sets	Tweets	Classes Or Segments
Stanford Sentiment Tree bank (SSTB)	Train	8545	5
	Dev	1100	5
	Test	2210	5
Stanford Twitter Sentiment (STS)	Train	80000	2
	Dev	16000	2
	Test	500	3

A. Dataset Process

Before initializing the sentiment analysis on data sets, First of all, the requirement is to well organize the format of text and relevant features to extract. To acquire this, the following steps are essential in the consideration.

1) *Data Gathering:* The data are extracted from various online sites such as social sites like Twitter, Facebook and e-commerce. The requirement to access its data is to create an account on the required online site.
2) *Pre-processing of Data:* The pre-processing of data is carried out to perform the crucial tasks in the sentiment analysis for the collection of data sets. The concepts which are covered under the preprocessing are case conservation, stop word removal, minimizing punctuations, stemming, lemmatization and correction of spellings.
3) *Feature Extraction:* This is the third step in the process of SA, which performed after the collection of data and its preprocessing. In this step, the essential and unique features of content are captured and make them according to the process of SA. It includes the frequency, speech tagging words, etc.
4) *Training and Classifiers:* This is the last step in the process which is done on the extracted features and in this step, the machine learning algorithms are applied to generate the correct and effective datasets of the acquired information. The training set is used to train the classifiers and the common classifiers of machine learning are Naïve Bayes classifier, SVM (Support Vector Machine) and Decision trees [13].

IV. DATA MINING TECHNIQUES IN SENTIMENT ANALYSIS

The extreme use of the computer and the internet gives plenty of data which needs to be managed correctly. Generally, data mining is a collaborated group of computer science and statistics that mainly acquired to extract the information from the patterns. It is a process of extraction that is in hidden form and a vital component in the data management. The other big advantage is to correct the mistakes from the huge relational datasets. For the sentiment analysis, data mining performed a major task to complete the process of it. It introduced several kinds of techniques that performed for the efficient completion of the sentiment analysis process [14].

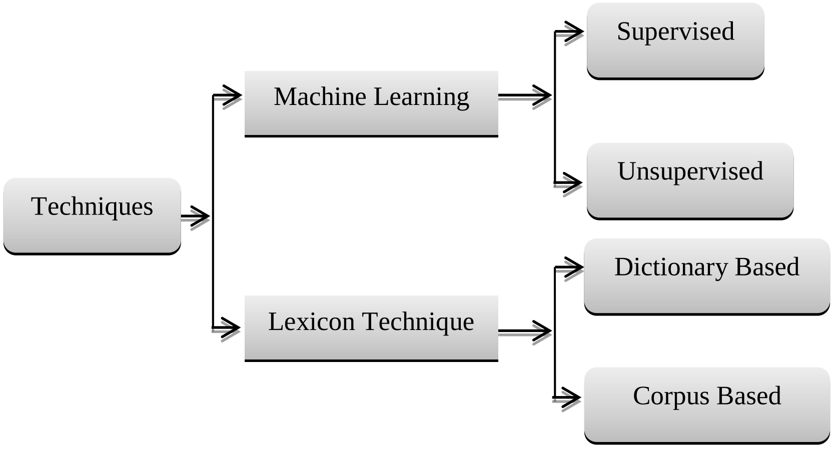

Fig.2:Different Techniques in Sentiment Analysis [4].

A. Lexicon Based Technique

Lexicon based technique is an opinion based method that easily determined the sentiment of a specific text. It worked with the evaluation of different number of positive and negative words in a particular text. There is a condition that is always essential to be considered. If the high rate of positive numbers in the text, then it assigned to a positive score, whereas if the high rate of negative numbers then it is allocated to the negative score. A neutral score is made on the condition, if both are of same rate such as the same number of positive and negative.

The lexicon based technique is sub-categorized in various other techniques such as :-

- *Dictionary Based Technique:* It is a collection of tiny words that commonly called as orientations. The synonyms and antonyms are searched in this method. The group of words is commonly gained until the new words are put in it.
- *Corpus Based Technique:* This is relied on the huge corpora which referred for the both syntactic and semantic patterns of the words. The words that are obtained from it are specific context and occupied a huge data set.

The corpus based technique is sub-partitioned in other small methods which are used for the sentiment analysis. The names of nested techniques are statistical and semantic.

B. Machine Learning Technique

Machine learning is used in each field in this era, it played out a major role in every field such as science, networking, medical and so on. For the sentiment analysis, the machine learning technique accessed to perform some specific tasks that are crucial to complete the process of analysis of emotions also. The sentiment analysis begins with the group of data sets which has the labeled data. The data sets are predefined by the use of some natural language processing. Next to it, the features are relevant for the sentiment analysis. The learning methods are:-

- *Supervised Learning:* The supervised learning methods are SVM (Support Vector Machines), Naïve Bayes classifier and decision trees.
- *Unsupervised Learning:* The unsupervised learning methods are similar as the lexicon based method. It included the learning patterns in input in case of no value of output. For example,K mean is an example of unsupervised learning method.

C. Hybrid Technique

The hybrid technique is a collaboration of both machine learning and lexicon based technique. It is proven by the developers and researchers that the combined method gives a better performance. It includes the further three sentiment analysis methods.

- *Word Level Analysis:* The most used technique for the sentiment analysis. In this method a powerful encoding is completed between the sentiment words.
- *Sentence Level Analysis:* In this method, the various levels of text are analyzed. A rule based method is performed for the identification and verification of the sentence. To describe the negative perspective, the words like no, not and never are used. In this method, the negative expressions are also explained by the use of verbs such as problem and stop.
- *Feature level analysis:* It is considered as an intelligent method of analysis that performs well and overweight the review process of other methods. In this method, the identification of features is done and the orientation score searched out.
- *Document level Analysis:* This method is preferred to describe the each document in a class or more. The positive and negative labels are taking place. The positive label shows the positive opinions, whereas the negative label described the negative opinions of users [13] [15].

V. CONCLUSION AND FUTURE SCOPE

The sentiment evolution and reviews are flourishing because the extreme growth in e-commerce that specifically tried to express the analysis of the opinions. In this paper, a comprehensive study and review is done to describe the several aspects of sentiment analysis. The most preferred techniques of data mining are discussed. These techniques are lexicon based, machine learning based and a combination of both. It includes the further categories of these techniques that played out a pivotal role to perform the basic tasks of sentiment analysis. These methods are naïve Bayes, support vector machines, particle swarm optimization, decision trees and so on. The other major work is related to the datasets which is generated using SA. Subsequently, the overview of latest work is described which is performed in the SA. The simple and easy opinions are given to sort out the challenges occurred under the analysis procedure. In the upcoming time, the SA must be more effective to specify and detect the election results and reviews from online sites. The major social sites like Facebook and Twitter give access to SA to predict about the future of their work.

REFERENCES

[1]. Cambria, E. (2016). Affective computing and sentiment analysis. IEEE Intelligent Systems, 31(2), 102-107.

[2]. Bhattacharyya, P. (2013). Sentiment Analysis. First international conference on emerging trends and applications in computer science.

[3]. Neethu, M. S., & Rajasree, R. (2013, July). Sentiment analysis in twitter using machine learning techniques. In Computing, Communications and Networking Technologies (ICCCNT), 2013 Fourth International Conference on (pp. 1-5). IEEE.

[4]. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

[5]. Ravi, K., & Ravi, V. (2015). A survey on opinion mining and sentiment analysis: tasks, approaches and applications. Knowledge-Based Systems, 89, 14-46.

[6]. Patil, H. P., & Atique, M. (2015, December). Sentiment analysis for social media: a survey. In Information Science and Security (ICISS), 2015 2nd International Conference on (pp. 1-4). IEEE.

[7]. Sindhu, C., Vyas, D. V., & Pradyoth, K. (2017, April). Sentiment analysis based product rating using textual reviews. In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of (Vol. 2, pp. 727-731). IEEE.

[8]. Akhtar, M. S., Gupta, D., Ekbal, A., & Bhattacharyya, P. (2017). Feature selection and ensemble construction: A two-step method for aspect based sentiment analysis. Knowledge-Based Systems, 125, 116-135.

[9]. Pham, D. H., & Le, A. C. (2018). Learning multiple layers of knowledge representation for aspect based sentiment analysis. Data & Knowledge Engineering, 114, 26-39.

[10]. Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., & Gupta, B. (2017). Deep Recurrent neural network vs. support vector machine for aspect-based sentiment analysis of Arabic hotels’ reviews. Journal of Computational Science.

[11]. Geetha, M., Singha, P., & Sinha, S. (2017). Relationship between customer sentiment and online customer ratings for hotels-An empirical analysis. Tourism Management, 61, 43-54.

[12]. dos Santos, C., & Gatti, M. (2014). Deep convolutional neural networks for sentiment analysis of short texts. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers (pp. 69-78).

[13]. Jain, A. P., & Dandannavar, P. (2016, July). Application of machine learning techniques to sentiment analysis. In Applied and Theoretical Computing and Communication Technology (iCATccT), 2016 2nd International Conference on (pp. 628-632). IEEE.

[14]. Agarwal, S. (2013, December). Data mining: Data mining concepts and techniques. In Machine Intelligence and Research Advancement (ICMIRA), 2013 International Conference on (pp. 203-207). IEEE.

[15]. Raghuvanshi, N., & Patil, J. M. (2016, March). A brief review on sentiment analysis. In Electrical, Electronics, and Optimization Techniques (ICEEOT), International Conference on (pp. 2827-2831). IEEE.