# Sentiment Analysis using various Machine Learning Techniques: A Survey

**Tanishq Mehta[1], Varsha Khandekar[2], Rachana Munot[3],Kathan Jain[4], Ruchita Bhamare[5]**

*[#]Department Of Information Technology,*
*Sinhgad College(SKNCOE),Pune,India*
*[1]tanishqmehta13@gmail.com,[2]varsha.khandekar@gmail.com,[3]rachanamunot2424@gmail.com,*
*[4]kathanjain77@gmail.com,[5]bhamareruchita20@gmail.com*

***Abstract***

*Sentiment analysis is the way of categorizing the opinions and understanding the emotions expressed in the piece of text computationally. Sentiment analysis proved its importance with the fact that it allows different vendors to segregate the reviews described by various groups of individuals. Recent researches involve the use of sentiment analysis in various applications like multilingual web texts, movie reviews, twitter dataset, etc. The basic purpose of study is to figure out how accurately and consicely different Machine Learning algorithms justify the sentiments that are dignified through the text .There are various parameters according to which sentiments can be classified into various categories. This paper highlights the essentialities as well as the applications of Sentiment Analysis done uptil now. A comparative analysis is also shown so as to let researchers to drive through the better technique.*

Keywords— *Sentiment Analysis, Machine Learning, Multilingual,Twitter*

## I. INTRODUCTION

From the past one decade, with the rise of uers on internet, the user feedback and reviews have been increasing imensely against various reforms. The boom of search engines like Yahoo and Google has flooded users with huge amount of relevant reviews related to specific organization or activity. With the increasing use of Social web such as twitter, facebook and other social media contributes vast amount of user generated content such as customer reviews, comments, opinions. Such a huge rise in Datasets makes it difficult to study and analyze the opinion of different community about a particular Domain. Each organization wants to know the

Emotions of their customers through the product feedbacks which helps them to improve quality of their products according to the demands. This helps them to increase their productivity. Such a huge rise in Datasets makes it difficult to study and analyze the opinions of different community .The traditional way of analyzing these datasets involved a lot of manual work which makes it unreliable and untrustworthy. Thus to reduce that, a technique known as "Sentiment Analysis", is being made into picture so that the evaluation of reviews or feedbacks could be made machine dependable. Sentiment Analysis is an intellectual process of extracting user feelings and emotions.

The reviews can be of product, service,movie or it includes individual statement. Sentiment analysis is used to classify these statements as positive or negative.This information is useful for the individuals.Analyzing this bulk of content is quite time consuming and difficult. So there is need to develop a smart system which helps in classification of huge data into positive, negative and neutral category.A lot of reserch work has being held in Sentiment analysis due to marketing level competition. Right now, a look at the different strategies have been utilized for Sentiment Analysis by breaking down different techniques. The objective of this paper is to discover the intelligent system that automatically finds and distributes the content of sentiment analysis with different techniques and approaches.

The rest of the paper is organized as follows. Section 2 discusses the Data Preparation steps so as to clean the data. Section 3 presents various Sentiment Classification Techniques. Section 4 presents models and methodology as well as tools involved in researches done uptil now. Section 5 shows the comparative analysis of various techniques based on previous researches. Section 6 concludes the paper.

## II. DATA PREPARATION

### i. Data Collection

Data is gathered from various organization in which sentiments of people within that organization is required to be analyzed.

### ii. Data Preprocessing

- **Removal of Null Values**: It involves the exclusion of null/NaN values from rows and columns.
- **Lowercasing**: Characters are converted to lower case so as to reduce the difficulty in the process of matching words in the given dataset. It helps to establish font uniformity.
- **Removal of stop words**: “Stop words” are the most common words in a language like “the”, “a”, “on”, “is”, “all”. These words do not illustrate any sense in analyzing sentiments, so they are usually removed from texts.
- **Removal of Punctuations**: Punctuations are considered as irrelevant content in the dataset, since they does not describe any sentiment .
- **Removal of URL**: In certain datasets like twitter Data, links also have been seen to appear in the messages or tweets. The URLs are considered as noise and thus are required to be removed.
- **Tokenization**: The words in the sentence are extracted as tokens from the dataset.
- **Lemmatization:** It helps in dissolving the conjugational ending words and return the base form of a word.
- **Normalization**: Abbreviated content is normalized by using a dictionary to map the content to frequently used Internet slang words. For example, “gud” and “awsm” are mapped to “good” and “awesome,” respectively.

### iii. Feature Selection :

- ***Term Frequency:*** Term Frequency $tf_{t,d}$ of a report or document d is proportion of significant terms inside a given report to the total no. of words in that report.
- ***Term Frequency–Inverse Document Frequency (TF-IDF):*** TF-IDF can be applied for separating stop words in various branches of knowledge like text summarization and classification.
- ***Feature Extraction :*** Once data gets preprocessed , the features are required to be extracted. It can include parts of speech tagging, opinion words, phrases and negation.

## III. SENTIMENT CLASSIFICATION TECHNIQUES

### A. *Lexicon based approach*

Lexicon based approach is contextual and domain-specific. It is mainly used in sentiment analysis which makes use of lexicon to calculate the overall polarity or orientation of a given comment or feedback. Lexicon based approach uses a dictionary that is full of Sentiments and matches them with the opinion words present in the feedback to determine the polarity score.

In this approach, sentiments or opinion words are divided into three categories. Positive opinion words depict that the things are in favour, negative opinion words depict that the things are not in favour, and neutral opinion words don't depict anything, these words have a numerical representation that decides whether the word is positive, negative or neutral[2] . A sample in the Table I is shown wheresentiment score is assigned to the words.

TABLE I

SAMPLE OPINION WORDS IN DICTIONARY DATABASE [1]

| Sentiment | Score | Description |
|---|---|---|
| Good | +2 | Positive |
| Fast | -1 | Negative |
| Very | +50% | Intensifier |
| Slightly | -50% | Intensifier |
| ordinary | 0 | Neutral |

This method describes the opinion words which are generally labeled according to their semantic orientation as Positive, Negative and Neutral and then assign polarity scores to them after matching it with dictionary database. The range of polarity score for the sentiments is from -3 to +3. Score range 1 to 3 shows the positive words, -1 to -3 shows negative words and 0 shows the neutral words[1]. It is domain-specific which means that the opinion words in different domains represent different meaning e.g.- ‘fast’ depicts negative word in student feedback but it depicts positive nature in sports like- “He is a fast runner”[3].

There some more words used in lexicon-based approach:

I. Negation words: These are the words that reverse the polarity of the opinions. For example: “He is not a good teacher”, in this ‘good’ is a positive word with polarity taken as ‘+2’ but the presence of ‘not’ word reverses its polarity by ‘-2’. Other negation words are: no, not, never, neither.

II. Blind Negation words: Some comments consist of positive words which tells that it is positive feedback but the presence of the negative words in that makes that comment negative. For example: “His teaching skills needed to be better”, in this ‘better’ depicts a positive word with polarity ‘+2’ but the presence of ‘needed’ makes that word negative with polarity score ‘-2’. Other Blind negation words are: need, needed, better.

III. Intensifier: These are divided into two categories which depend on their polarity score. Amplifiers increase the intensity of the sentiments (example: very) and down toners decrease the intensity of sentiments (example: slightly) [1].

Lexicon-based sentiment analysis is used to evaluate the level of teaching performance from student's textual feedback comments by using the dictionaries constructed in the database [5]. There are three different methods used for evaluating the overall polarity of the comments are -

I. Term counting(TC) : It classifies the word in positive, negative and neutral by counting the words present in the comments. For example,"He is a good as well as a nice teacher", in this comment there are two opinion words 'good', 'nice' which shows that they are positive and the term count of this positive word is '+2' so it clearly shows that this comment is positive.

II. Term Score Summation(TSS) : The role of this method is to do a summation of scores of all positive words with positive and negative with negative.
Example 1, "He is not a good teacher but a nice person".
Example 2,"His knowledge about the topic is good but didn't give proper notes. "In this, there are two positive words 'good', 'nice' , so the TSS of positive words is '2+2=4', and there are two negative words 'falls', 'bad', so the TSS of negative words is '3-2=1'.

III. Average on comments(ASAC) : The main role of these methods is to calculate the average of positive and negative scores for each opinion word found in the student feedback comments. These methods consist of the inner three method accuracy, precision, recall of this method [4].

After evaluating the average polarity score of all the feedback if the average Polarity score is '>0' or '<=1', the feedback result is defined as weakly positive. If the Polarity score is>1 or <=2, the feedback result is defined as moderately positive. If the Polarity score is>2, the feedback result is defined as strongly positive. If the Polarity score is '<0' or '>= -1', the feedback result is defined as weakly negative. If Polarity score is '< (-1) 'or '>= (-2)', the feedback result is defined as moderately negative. If the Polarity score is '< (-2)', the feedback result is defined as strongly negative. If the Polarity score is '=0', the feedback result is defined as neutral[1].

### *B. Support Vector Machine*

SVM is the characterization of the instances as points in multidimensional space, drawn so that similar type of instances could be easily integrated and the different instances could be separated by a wide gap. It is a non-probabilistic classifier in which a large amount of training set is required. This approach was introduced by Vladimir vapnik, Isabelle Guyan and Bernhard Boser in 1992. SVM helps in examining the data, define decision boundaries and then implement kernels for calculations which are performed in input space[6]. SVM is generally applied for linearly separable data but for non-linear data classification, implicit mapping of inputs is done into higher dimensions [7]. The process of SVM involves categorization of data followed by mapping of support vectors and by taking reference of those support vectors a hyperplane is drawn which is an optimal solution.

### *C. Naive Bayes Classifier*

As the name suggests, it is one of the simplest and popular classifier which has remarkable performance in text classification. This classifier works by computing the posterior probability of a class by extracting the words as features from the document . The feature extraction is done with the help of methodology named as “Bag of Words”[10]. It assumes that the features are independent of one another. Bayes Theorem is used to predict the probability for a given feature as shown in Equation 1 [11]:

$$P(label/features) = P(features/label) * P(label)/P(features) \textbf{...1}$$

Here, P(label) represents the likelihood that the label has been observed. P(features|label) shows the probablity of feature set that has been classified as label. It is also known as prior probability. P(features) shows the probability for the given feature which has already been occurred. The Naive assumption that illustrates the independence of features with each other can be equated as shown in Equation 2:

$$P(label/features) = P(label) * P(f\,\mathbf{1}/label) * \textbf{..}\, P(fn/label)P(features) \textbf{...2}$$

### D. Maximum Entropy

Maximum Entropy text classifier is also known as 'MaxEnt classifier', commonly used in Natural language Processing.It belongs to the class of exponential models.It doesnt assume that the features are conditionally independent of each other.It is used when conditional independence of the features can't be assumed.The MaxEnt requires more time to train the data comapring to Naive Bayes.These models are feature-based model.Without worrying about about features overlapping, features like bi-grams and phrases can be added in MaxEnt. In classification, significance of a feature is dependent on weight vector for any class.The higher weight means the feature is a strong indicator of a class. MaxEnt is parameterized by set of weights that are used to combine joint-features generated from set of features by an encoding.In encoding each pair of feature set and label get mapped.After extracting some set of features from input,they get combined linearly and this sum is exponent.The best thing about MaxEnt is it performs better than Naive Bayes bayes because it handles overlapping. This classification algorithm regularly yields excellent outcomes, as it can manage inadequate data in a rich manner. While characterizing natural language articulations the preparation models seldomly spread all variation linguistic expressions that sign certain semantics[13].

### E. Hybrid Approach

To improve the classification performance,there are few techniques with combination of lexicon based and machine learning techniques. This helps in extending the accuracy of the system but it is complex to design. Olga , Kolchyna et. al.[15] in their study proposed the algorithm where they combined Lexicon with machine learning technique to compute the better and accurate result. They showed how lexicon score can be used as a feature in machine learning classification. It helped to improve the accuracy of model prediction by 5% [15].

## IV. METHODOLOGY AND TOOLS

Sentiment Analysis comprises of huge amount of studies and processes addressed by many researchers. But there are some studies which had their remarkable impact in the field of Sentiment Analysis. Those methodologies including tools are listed below :

### A. *Rapid Miner Tool*

Using the open-source data analytics tool Rapid Miner, simultaneous training and testing the classifiers can be done by making use of the Validation operator from Rapid Miner, whereas the performance is evaluated using the Performance operator.

All the other necessary operators are contained in the ProcessDocumentsFromFiles operator. One such field of data analytics is Learning Analytics which measures, does analysis, reports and predicts data about learners for optimizing the teaching and learning ways. With lots of recent developments in data mining, various things like computational linguistics, Machine Learning, and Natural Language Processing have come together for the purpose of automation of the unstructured data. Various performance metrics like accuracy, precision, recall, etc. are used for evaluating the performance. For the data classification, the polarity is determined for which a large amount of labeled data is used which is generated from various supervised learning algorithms. For classifying the sparse text data, the SVM classifier works best and in this classification simply by defining the rectilinear partitions in the dataset and then dividing that data into appropriate classes, and output can be achieved.

There are three models which are developed according to research which tells us that the first model consists of operators for classification purpose. The second model consists of the data pre-processing step while the third model consists of the student's feedback categorization and performing the sentiment analysis of it. For the unstructured data, in order to assign the correct polarity for the sentiments like positive, negative or neutral, the ANNIE technique is used to create RDF or OWL which is metadata.

### B. *Satisfaction and Disatisfaction Computation*

The system proposed by Sujata Rani and Parteek Kumar [8], classifies the sentiments in two categories , positive and negative and emotions into Robert Plutchik's eight categories, namely, anger, disgust, anticipation, fear, joy, sadness, surprise, and trust from which it calculates satisfaction and dissatisfaction. This classification was done with the help of NRC Emotion Lexicon so as to to associate words with positive or negative sentiment and the eight basic emotions.Their model of lexicon included annotations for around 14,182 unigram words for English.

Model was designed in such a way that every word in Lexicon had an emotion vector (E) containing a Boolean value (b) for each sentiment (s) and emotion (e) such that :

$$\vec{E}=\vec{E_e}+\vec{E_s}\ , \qquad \ldots\ldots\ldots\ldots\ldots\ldots\mathbf{3}$$

If the word obtained through the dataset matches the one in lexicon, the respective emotion vector is returned. Thus, an emotion vector is created for each comment containing the different emotions and sentiments within. A formula was derived to elaborate the parameters like satisfaction and dissatisfaction. Satisfaction and dissatisfaction are crucial parameters in education. To calculate the satisfaction a formula was derived. For example, if we take a positive comment,"The notes by Sir are awesome". This comment shows that the feedback given in the dataset is having emotions as trust and joy. Thus, the binary score is assigned to the emotions respectively as shown in Table II.

TABLE II

EMOTION CLASSIFICATION[8]

| Anger | Anticipation | Disgust | Fear | Joy | Sadness | Surprise | Trust | Negative | Positive |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 |

Therefore, in computing student satisfaction, multiplication of the sum of anticipation and trust by a constant (α = 0.6) to give these parameters more weight. The same mechanism has been employed in computing student dissatisfaction to give more weight to anger and disgust than to sadness.

The satisfaction and dissatisfaction can be calculated as :

Satisfaction = $[\alpha(TA)+(1-\alpha)(J)]/n$

Dissatisfaction = $[\alpha(AD)+(1-\alpha)(S)]/n$ ,

where TA = trust + anticipation , J = joy,

AD = anger+ disgust, S = sadness, and n = max(TA or AD, J or S).

From Table II, “TA=1+1=2”, “J=1” and “n = max(TA, J) = 2” ,thus the satisfaction for the above example can be calculated as :
[0.6(2) + 0.4(1)]/2 = 1.6/2 = 0.8. Later, to expedite analysis of student feedback about course satisfaction and teacher performance, data visualization was applied which generates sentiment and emotion word clouds as well as line graphs of changes in sentiments and emotions over time.

### *C. IBM Watson*

The system architecture by Watson is the great innovation for the computation in the new era of cognitive system where effective navigation through the huge and dynamic amount of unstructured information is performed[9]. IBM Watson is one of the finest cognitive system when Natural Language Processing comes into picture. Cognitive Systems are those which have configurational abilities such that the system can have human-like capabilites to convey and determine the ideas. Since the systems prior to this one, provided precise results when it comes to describe the sentiment score. But , “precise” is a mechanical term which is totally different from the term accuracy.

In simple words, Watson described it by an example like , suppose, if configuration of car is considered, then car can be described as “2+2” seater which mathematically be described as four seater but logically, it indicates that two seats are in front and two are in the back. This is what the difference between accuracy and precision. Watson called this concept as Shallow Natural Language Processing. Shallow natural language processing can be fairly precise within its more narrow focus, but is not very accurate[9]. Unlike other methodologies, Watson does not actually considers the individual words in the language. Rather it just understands the features of language that were used by people. Using those features, the system can determine how one text passage is linked with another text passage, with a remarkable level of accuracy under dynamic situations.

Text analysis by Watson goes in a very systematc order by initiating the process of question and topic analysis. It does that by exploring multiple interpretations and finding hundreds of answers. Hypothesis is generated and thousands of evidences are gathered. Then deep learning algorithms are applied and hypothesis with evidence scoring is done. Lakhs of scores are generated with help of these algorithms. Each reasoning algorithm generate one or more scores, indicating the range to which the potential response is contained within the question based on the specific area of focus of that algorithm. The outcome as score of each algorithm is then weighted against a statistical model that signifies how much accurately that algorithm did at organizing the inferences between two similar passages under the orientation for training the model. That statistical model can then be used to summarize a level of confidence that Watson has about the evidence that the candidate answer is inferred by the question. The system repeatedly performs this process for each of the candidate responses unless it could get some better responses stronger enough than other candidates.

## V. Comparative Analysis of Various Algorithms

Interests of researchers in the field of Sentiment Analysis is growing rapidly. The study of various algorithms is done in this section where the results of Lexicon based approach and Machine LearningTechniques like SVM, KNN and Naive Bayes are compared. In the area of Sentiment Analysis , all the researches showed their main interests in Lexicon based approach. But , machine learning techniques like KNN and Naive Bayes have also proved their accuracy when sentiment analysis of social media analysis was done.

In the study of Hanif Sudira, Alifiannisa Lawami Diar and Yova Ruldeviyani, two classifiers Naïve Bayes and KNN were used to model and classify the data [12]. The results for Naive Bayes reached best accuracy after 20-folds cross-validation. This result has been computed by analysing sentiments of customers using online payment modes like Go-Pay, Ovo, and LinkAja. On comparing with KNN, when considered same data as input, the model gave the best accuracy in 15-folds for Go-Pay and Ovo as 82.39% and 79.77% respectively. But for LinkAja reviews, accuracy was 94% in 20-folds[12] .

Although, SVM has also proved its accuracy in classifying texts into positive or negative categories. The scikit-learn library was used by developers where linear kernel was used to train the model. Regarding the movie reviews example, SVM gives higher accuracy when compared to the Naive Bayes algorithm. In the research by Shweta Rana and Archana Singh, Linear SVM showed the accuracy of "87.50%" nut Naive Bayes showed highest accuracy of "80.00%". A concept called the Porter algorithm is introduced which helps to remove suffixes from words[14]. The Rapid Miner tool is used for data analysis purposes. IR (Information Retrieval) is a typical environment which is used for reducing the overloaded information and also store and manage the files and documents. Most visible IR applications are Web search engines.

## VI. Conclusion

The main motive behind this study is to extract various methodologies proposed in the field of Sentiment Analysis uptil now.It has been indicated that change of the enormous volume of textual data from internet into important information can be valuable. Be that as it may, the undertaking of precise extraction still stays provoking. This paper contains various Machine Learning Algorithms like Lexicon Approach, Support Vector Machine, Maximum Entropy. There are some other algorithms too, but these algorithms are more emphasized due to their accuracy and efficiency.

Various methodologies proposed by researchers are also listed in this paper, which can provide a base for the new researchers who are working on this domain. Hybrid approach which involves various Machine Learning Techniques with Lexicon based approach are used and results are obtained with better accuracy. Although the accuracy levels are required to be increased more since the classification under the neutral category is still needed to be more. In Lexicon based approach, there are two main drawbacks , first one is that it inability to process acronyms, emoticons and second is that it unable to account for sentiment intensity (For example, Food here is exceptional vs food here is good) this drawback comebacks by VADER (Valence Aware Dictionary and Sentiment Reasoner) was developed by Georgia Tech computer science department which addresses above drawbacks[5]. Although, IBM Watson model improved the accuracy with the help of generalization but there are certain revelations. One of them is that, by utilizing Watson to help answer questions, one may understand that he/she are in a general sense posing an inappropriate question. Concisely speaking, at the point when Watson reacts to individual's inquiries, in any event, noting effectively, one may understand that he have to ask other, better, and progressively significant questions to help consider his business issue in a totally different manner. One may begin to think in manners that help themselves to comprehend the serious dangers and openings in one's field that never jumped out at them before. Thus these problems in the methodologies are required to be improved in the near future[9].

## REFERENCES

[1] Khin Zezawar Aung, Nyein Nyein Myo University of computer Studies,“Sentiment Analysis of Students Comment Using Lexicon based Approach” IEEE ICIS 2017, May 24-26, 2017.

[2] Paramita Ray and Amlan Chakrabarti,“Twitter Sentiment Analysis for Product Review Using Lexicon Method,” 2017 IEEE.

[3] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka ,“SentiFul: A Lexicon for Sentiment Analysis”,IEEE/T-AFFC.2011.vol 2. no 1.

[4] Nurul Fathiyah Shamsudina, Halizah Basirona, Zurina Saayaa, Ahmad Fadzli Nizam Abdul Rahmana, Mohd Hafiz Zakariaa, Nurulhalim Hassimb ,“Sentiment classification of unstructured data using lexical based techniques,”Jurnal Teknologi /2015- 113-120.

[5] Son Trinh, Luu Nguyen, Minh Vo and Phuc Do,“Lexicon-Based Sentiment Analysis of Facebook Comments in Vietnamese Language”DOI 10.1007/978-3-319-31277-4_23/2016 .

[6] Geetika Gautam, Divakar yadav, “Sentiment Analysis of Twitter Data Using Machine Learning Approaches and Semantic Analysis," 978-1-4799-5173-4/14/$31.00 ©2014 IEEE .

[7] P.Kalaivani et.al,"Sentiment Classification Of Movie Reviews by Supervised Machine Learning Approaches," Indian Journal of Computer Science and Engineering (IJCSE), ISSN : 0976-5166 Vol. 4 No.4 Aug-Sep 2013 .

[8] Sujata Rani and Parteek Kumar,"A Sentiment Analysis System to Improve Teaching and Learning," 0018-9162/17/$33.00©2017IEEE .

[9] Rob High’s, “The Era of Cognitive Systems: An Inside Look at IBM Watson and How it Works,”IBM Corp. 2012.

[10] Anuja P Jain and Asst. Prof Padma Dandannavar’s, “Application of Machine Learning Techniques to Sentiment Analysis,”978-1-5090-2399-8/16/$31.00 #2016IEEE.

[11] Walaa Medhat ,Ahmed Hassan, “Sentiment analysis algorithms and applications:A survey” Shams Engineering Journal (2014) 5, 1093–1113.

[12] Hanif Sudira, Alifiannisa Lawami Diar and Yova Ruldeviyani, "Instagram Sentiment Analysis with Naive Bayes and KNN: Exploring Customer Satisfaction of Digital Payment Services in Indonesia"978-1-7281-5347-6/19,2019IEEE.

[13] Erik Boiy,Marie-Francine Moens, "A machine learning approach to sentiment analysis in multilingual Web texts" Springer Science +Business Media, LLC 2008.

[14] Shweta Rana, Archana Singh, "Comparative Analysis of Sentiment Orientation Using SVM and Naïve Bayes Techniques" 978-1-5090-3257-0/16 ©2016 IEEE.

[15] Kolchyna, Olga & Souza, Thársis & Treleaven, Philip & Aste, Tomaso. (2015). "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination."