# A SURVEY PAPER ON TEXT MINING ON WEB DATA USING MACHINE LEARNING TECHNIQUE

Kshama Singh
M.tech Scholar
Department of Computer Science & Engineering
LNCTE
Bhopal, India

Sitaram Patel
Assistant Professor
Department of Computer Science & Engineering
LNCTE
Bhopal, India

*Abstract*— **In today's e-commerce market where online shopping and tourism is fastly growing so it very important to analyze such huge amount of large data present in web. So it is very important to create a method which classify the web data. Sentiment analysis is a method to classify the web data such as product reviews, views in to various polarities such a positive, negative or neutral. Text mining generally refers to the process of extracting valuable information from unstructured text. In this survey of text mining, several text mining techniques and its applications in various fields have been discussed.**

*Keywords*—**web data, text mining, data mining, visualization, R, text mining techniques.**

## I.INTRODUCTION

Sentiment Classification techniques is roughly divided into Lexicon primarily based approach, Machine Learning approach and hybrid approach. The Machine Learning Approach (ML) applies the renowned metric capacity unit algorithms and it uses linguistic options. The Lexicon-based Approach depends on a sentiment lexicon. Lexicon could be a assortment of illustrious and precompiled sentiment terms. it's once more divided into dictionary-based approach and corpus primarily based approach that use linguistics or applied math strategies to seek out sentiment polarity of the text.

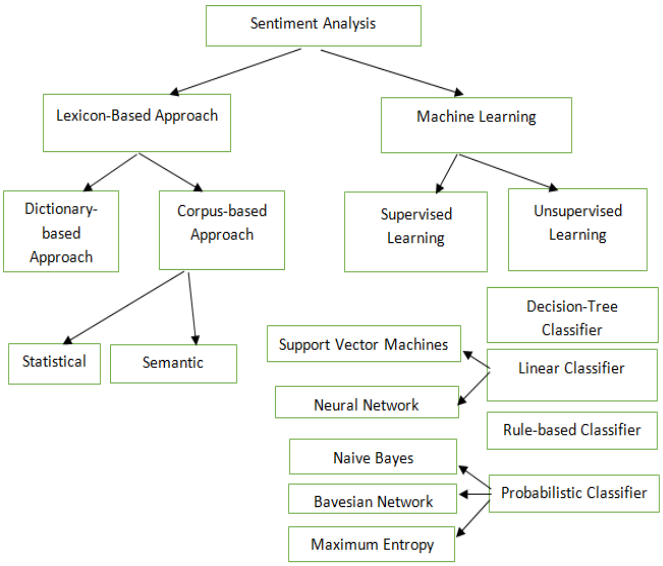


Figure 1. Machine Learning Approaches

### Machine Learning Approach

In Machine learning we can created a model and these model can classify the reviews based on there learning techniques. In these we can train the model onto two ways

1. Unsupervised learning:
In these they do not want a supervisor to set thier label to get the desired output, in these the model learn by itself and then classify the reviews.
2. supervised learning:
In these we can select the labels and thier is a supervisor who guide the model and selection in label though which they can get the desired output. In these we can divide the process into two phase:
1. Training Set
2. Test Set.
A various number of machine learning techniques are developed to classify the tweets into categories. The Machine learning techniques like Naive Bayes (NB), most entropy (ME), and support vector machines (SVM) have achieved nice success in sentiment analysis.

In these a machine learning first pre-process the dataset based on the algorithm which is using in supervised learning and after that we can trained the model by giving training dataset and the model is start learning and after the learning of the model is completed we can test the model performance on test dataset and then we can compute the various performance measures like accuracy, f1 score.

## II. LITERATURE REVIEW

According to ( Rachana Bandana et al., 2018) [1], Human disposition has always influenced by others suggestion and reviews. People are always eager to know other's reviews for their profit but, every website contains a very large amount of review text, the average human reader will have trouble in identifying relevant sites, extracting and abstracting the reviews so they cannot reach to the right decision in less time that is why automated sentiment analysis systems are required. In the proposed approach, heterogeneous features such as machine learning based and Lexicon based features and supervised learning algorithms like Naive Bayes (NB) and Linear Support Vector Machine (LSVM) used to build the system model.

Since early 2000, sentiment analysis is the most vigorous research areas of Data Mining and Natural Language Processing (NLP). Human disposition has always influenced by others suggestion and reviews. That is why our reviews are very much influenced by other's reviews, and whenever we need to make a decision, we often seek out other's reviews. When we need to check reviews, we will started trying to find reviews in many digital platforms such as social media , reviews site, forum discussions, blogs, and micro blogs, though every website contains a very large amount of review text, the average human reader will have trouble in identifying relevant sites, extracting and abstracting the reviews so we cannot reach to the right decision. This is not only true for any individual business person but also for organizations, companies, political parties that is why people need automated smart sentiment analysis system which can accurately give correct sentiment and relevant information in less time for their benefits.

Mine people's review and feelings toward any subject matter of interest, which is the task of sentiment analysis. Now a day, sentiment analysis can apply to almost all possible domains like products, services for social events and political elections, market research, social media, advertising, recommendation systems, email filtering, stock market prediction, upcoming movie reviews sentiment prediction, book reviews sentiment, etc.

According to (Paramita Ray et al., 2017) [2], In recent times, people share their opinions, ideas through social networking site, electronic media etc. Different organizations always want to find

public opinions about their products and services. Individual consumers also want to know the opinions from existing users before purchasing product. Sentiment analysis is the computational treatment of user's opinions, sentiments and subjectivity of text. In this paper they propose a framework for sentiment analysis using R software which can analyze sentiment of users on Twitter data using Twitter API. Our methodology involves collection of data from twitter, its pre-processing and followed by a lexicon based approach to analyze user's sentiment.

Customer feedback about particular products is very important for commercial organization. They can improve their product quality, services on the basis of customer opinion about their product. Twitter is a kind of micro-blogging social networking site and billions of users use it to give their opinion related to a particular topic. On the basis of opinion, sentiments can be estimated through analysis.

According to ( Rasika Wagh et al, 2018) [3], Social networking sites like twitter have millions of people share their thoughts day by day as tweets. As tweet is characteristic short and basic way of expression. So in this review paper they focused on sentiment analysis of Twitter data. The Sentiment Analysis sees as area of text data mining and NLP. The research of sentiment analysis of Twitter data can be performed in different aspects. This paper shows sentiment analysis types and techniques used to perform extraction of sentiment from tweets. In this survey paper, we have taken comparative study of different techniques and approaches of sentiment analysis having twitter as a data.

The social networking sites like Twitter, Facebook, and YouTube have obtained so much popularity now days. The area of sentiment analysis is known as opinion mining, it is under umbrella of computational linguistics and data mining. Its main aim is to detect the person's mood, behavior and opinion from text documents. With the expanded use of social networking sites, sentiment analysis techniques have started to use these sites' public data to do sentiment analysis studies in different sociological areas, such as politics, sociology, economy and finance.

Most of the data that available in social networks is unstructured. Such unstructured data is almost 80% of the data all over the world. This makes it difficult to analyze and gain valuable judgment from such data. Sentiment analysis or opinion mining is the important technique, which help in detecting opinions of people on social media data.

Opinions of others can be important when it is need to make a decision. When those decisions involve valuable resources people think about their companions' past experiences. Now a day's social media gives new tools to conveniently share ideas with peoples linked to the World Wide Web. Though sentiment analysis concentrate on polarity detection (positive, negative or neutral). Twitter is a micro blogging site which contains large number of short length utilizes for marketing, social networking. For example, political parties might be eager to know whether people support their curriculum or not. In present scenario the need to gather opinions from social networking sites and draw conclusions that what people like or dislike, has been the most important perspective. The objective of this review paper is to discuss concept of sentiment analysis of twitter tweet.

## III OBJECTIVE

The objectives for carrying out sentiment analysis can be as follows:

- Collecting review : In this dissertation we can collect the tweets review on a particular product.
- Preprocessing the text : The data which we get in real time is unstructured in nature means the does not have any structured or format and also the raw data contains very meaningful information and symbols like unwanted uri's, white space , number, punctuations, etc, So the text need to pre-processing before classification.
- Feature Extraction : Before classification using machine learning techniques, first we can create a model and train the model and during these training period the classification model collects the features from the text and based on these feature selection its classify the text into polarities.
- Classification of text : After the model is trained its been tested on data and based on feature selection from previous step the classifier model, classify the text into positive, negative and neutral classification.

## IV PROBLEM OBSERVATION

The existence and therefore the constructive stability of the trade and business depends on establishing a competitive dominance through effective and aggressive selling ways. With the prolific quantity of data that's unceasingly being created offered through the electronic media, the online users are unable to require advantage of those resources because of the shortage of acceptable tools to utilize. And many increase within the variety of internet sites puts forward a difficult task to arrange the contents of the websites to cater to the wants of the users.

Text mining becomes an solutions for organizations to analyze the text bases content such as social media posts on twitter, facebook and linkedin. Sentiment classification is a technique to classify the text expressed on social sites into various sentiment polarities.

## V CONCLUSION

Text mining generally refers to the process of extracting valuable information from unstructured text. In this survey of text mining, several text mining techniques and its applications in various fields have been discussed. A comparison of different text mining has been shown which can be further enhanced. Text mining algorithms will give us useful and structured data which can reduces time and cost. Hidden information in social network sites, bioinformatics and internet security etc. are identified using text mining is a major challenge in these fields. The advancement of web technologies has lead to a tremendous interest in the classification of text documents containing links or other information.

## REFERENCES

[01] Rachana Bandana, " Sentiment Analysis of Movie Reviews Using Heterogeneous Features " in IEEE 2018.

[02] Paramita Ray and Amlan Chakrabarti, "Twitter Sentiment Analysis for Product Review Using Lexicon Method" in 2017 International Conference on Data Management, Analytics and Innovation (ICDMAI) Zeal Education Society, Pune, India, Feb 24-26, 2017.

[03] Rasika Wagh, Payal Punde, "Survey on Sentiment Analysis using Twitter Dataset" in Proceedings of the 2nd International conference on Electronics, Communication and Aerospace Technology (ICECA 2018), IEEE.

[04] Borut Sluban, Igor Mozetič, Jasmina Smailović, Petra Kralj Novak, “Sentiment of Emojis”,Plos one (2015).

[05] Hailin Jin and Jianchao Yang, Quanzeng You and Jiebo Luo, "Robust Image Sentiment Analysis Using Progressively Trained and Domain Transferred Deep Networks",Association for the Advancement of Artificial Intelligence (2015).

[06] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau, "Sentiment Analysis of Twitter Data", NextGen Invent (NGI) Corporation(2012).

[07] Efthymios Kouloumpis, TheresaWilson and Johanna Moore, "Twitter Sentiment Analysis: The Good the Bad and the OMG!", Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media(2011).

[08] James Spencer and Gulden Uchyigit, "Sentimentor: Sentiment Analysis of Twitter Data", School of Computing, Engineering and Mathematics, University of Brighton(2014).

[09] Aarati Patil and Srinivasa Narasimha Kini, "Location Based Sentiment Analysis of Products or Events over Social Media", International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 4 Issue:7,50 -55.

[10] Vasavi Gajarla and Aditi Gupta, "Emotion Detection and Sentiment Analysis of Images", Georgia Institute of Technology(2015).

[11] Kausikaa.N and V.Uma, "Sentiment Analysis of English and Tamil Tweets using Path Length Similarity based Word Sense Disambiguation", IOSR Journal of Computer Engineering (IOSR-JCE), e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 3, Ver. I (May-Jun. 2016), PP 82-89.