

SENTIMENT ANALYSIS USING MACHINE LEARNING TECHNIQUES ON TWITTER: A CRITICAL REVIEW

SUDHEER KUMAR SINGH¹, PRABHAT VERMA, AND PANKAJ KUMAR

ABSTRACT. Social web for Social applications are developing an exponential rate over the internet with the growing Internet communities. Users are increasing every second to use the social applications on Social platforms like Twitter, Facebook, etc. Users are sharing their feeling or opinion about any person, product in the form of images or text on the social networks. Sentiment analysis has gained a lot of popularity in the research field of Natural language processing (NLP). Through it, the hidden sentiment in the text can be well extracted. This can assist companies, organizations, or users to make a useful conclusion to achieve their objective. The millions of data on the social network which has shared by the users will get certainly brings more opportunities and challenges to the sentiment analysis. In this paper, we will analyze the various Existing methods, techniques, and approaches for sentiment analysis, like Support Vector Machine, Naive Bayes, and KNN. In this paper, we have adopted a step by step literature review process to identify areas well focused by researchers for sentiment analysis in different fields. We have also tried to identify the best-used sentiment analysis techniques to find research gaps from the previous research, which can further be extended in future research work. We will also discuss and explore new development trends of sentiment analysis.

¹*corresponding author*

2010 *Mathematics Subject Classification.* 68T09, 62R07.

Key words and phrases. Sentiment analysis, natural language processing, Machine learning, support vector Machine, KNN.

1. INTRODUCTION

In the current scenario, Social networking websites such as Face book and Twitter are providing a platform for the social user's to communicate and post their perception of all types of subjects, articles, and events. Social network users are increasing day by day on the social media platform. Presently, Twitter has more than 700 million users. Users are sharing nearly 350 million tweets per day [2]. Recently, the Social network has materialized into extensive easily access reasonable data sources in a variety of fields [3]. Twitter as a social network platform is used for sharing information related to various fields like commercial enterprise and trading, political issues, and promotion of commodities and its services. The sentiments are a wonderful approach to extract the feeling of text in real-time Environments [4]. As we know, online social networking websites are growing as new communication platforms that give huge freedom for people to communicate and share their feeling. Peoples on a social network share the various points of view related to any subject or product with their friends, relatives, and followers, etc. users can post their views on special social issues belonging form national or International in the form of text, audio, video, photos, and post[1]. User opinions can be shared in different forms on social networking sites but from these forms, texts are still a popular and most easy way to communicate on social media sites. Sentiment can be expressed in terms of feeling, attitude, opinion, and emotions. As we know nowadays the web models are increasing so rapidly. Users express their sentiment on social networking websites or blogs. Some time users gave ratings and reviews on their interesting products or topics. Due to the exponential growth of peoples on social media in a fixed interval of time, Social data volumes are increasing. So, there is an urgent requirement to analyze the foundation of the sentiment and extract it from the text data. The objective of this research paper is to analyze different sentiment machine learning classification algorithms and detect the most usable algorithm for predicting the sentiment of the text data.

In Sentiment Analysis, subjective data is used. Sentiment analysis is carried out to extract three different feelings such as negative, neutral, and positive of the text data.

Firstly after cleaning and pre-processing of data, we extract the features using techniques such as 1, 2, or N-Gram Modelling methods. The sentiment of textual

data can be divided into three groups, positive, negative, and neutral using different Machine learning techniques. In this paper, we focused on different machine learning techniques on sentiment classification and its drawbacks, so we can further explore the new techniques using existing techniques and find the research area or research Gap, so current research can be further extended. We will explore the new trend in sentiment analysis by doing a critical review of different machine learning algorithms for sentiment classification.

2. LITERATURE REVIEW

The Authors analyzed in-depth for the study of the different classifiers when raising the issues of binary classification like positive or negative. Authors analyzed the methods or techniques about the sentiment analysis and found that SVM Model performs not so poor as comparative to the Naive Bayes, whenever performing the sentiment analysis especially positive test. Both methods and models are concerning about the fact and figure related to sentiment, whenever there is the counting of both positive and negative [5]. In social networking websites such as Twitter have used to provide the decision whether the tweets received the reply whenever it's raised a question. For example, is it possible to estimates the stock sales price of a business organization for the general evaluation? As discussed before, Naive Bayes was providing the prediction measure of any subject related discussed questions in feel likely that is the cost in the undue assess of period [6]. Authors had an opinion to reduce the research set to a satisfactory order size that is reliable with our sighting in the field of sentiment classification findings and Naive Bayes look like to jointly their better strictness by preparing the data set for consume 20% of the dataset which has participated to make a decision [7]. If by changing the preparation dataset sentiment classification algorithms like SVM, KNN and Naive Bayes reported the precise work in percentage while these are working on the same dataset [8].

3. ANALYSIS OF SENTIMENT CLASSIFICATION TECHNIQUES

In the last few decades, an emerging and famous research area was born to know as sentiment analysis. The opinion is collected in the form of text by several social networking users or individual people who are reviewed and analyzed using machine learning classifier techniques of sentiment analysis [9].

The feature vector has a very important role in the sentiment analysis of text data. Generally, classifiers are used to give prediction in the form of binary values 0 or 1. Some other classifier has also developed by the scientists to classify the data into more than two values such as positive and negative, negative or neutral, and maybe all three. The feature vector is extracted from the text or tweets on twitter. SVM Machine, Naive Bayes, and ensemble classifiers designed to use for extraction of feature vector from the text or data. The classification of text is required to resolve sentiment analysis issues. In this research review paper, we focused on the most common and useful classifier method which is categorized by machine learning and lexicon-based approaches [10]. Classifi-



FIGURE 1. Process diagram of sentiment analysis

cation techniques are used to classified text into groups by using classifiers like machine learning models or methods. Machine learning techniques are categorized mainly into two types: Supervised and unsupervised Machine learning. Unsupervised Machine learning techniques focused on that dataset in which objective is not clear, so we consider clustering is one important feature to solve the

real problem. Supervised learning techniques focused on the data set in which label data is used. Whenever researchers developed the machine learning model mainly supervised learning then labels of data have been created and provide to the model for making the data useful and categorized into training data and test data [10]. Generally, a set of particular features extraction and findings in such a way that sentiment classification can be extracted and it leads to the success of both the discussed machine learning techniques. As we know and seen most of the research paper discussed mainly three sentiment classification techniques for example Support Vector machines, Naive Bayes, and Maximum Entropy [9].

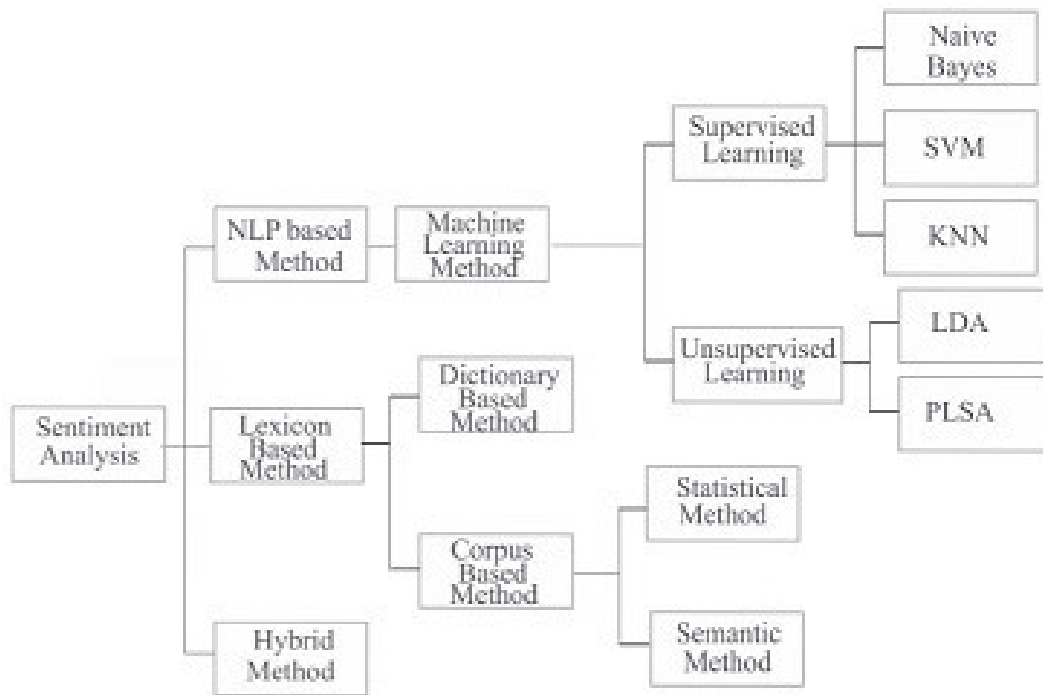


FIGURE 2. Machine learning Algorithms for sentiment classification

3.1. Naive Bayes. Naive Bayes model known as classifier is used to classifying the features of text through feature vector. It comes under supervised machine learning techniques. In this model, we consider features of the text are independent of each other and all are equally likely. Analyze these entire features vector exclusively [8]. Mathematical definition of conditional probability is given

for model as in following equation 3.1:

$$(3.1) \quad P\left(\frac{C}{D}\right) = P\left(\frac{D}{C}\right) P(C)/P(D),$$

A feature vector defines and represented by C such as $A = a_1, a_2, \dots, a_n$ and $C = a_1, a_2, \dots, a_n$ and the class label is defined and represented by C. Classification of distinct types of independent feature vector like positive and negative is effectively carried out by using the Naive Bayes model. This model cannot define a relationship among these feature vectors [9].

3.2. Support Vector Machine (SVM). SVM is used for solving the classification problems and it is a supervised machine learning algorithms. Generally it takes data in the form of positive and negative points, classified it using a line called a hyper line that line divides the whole reason into two separate parts. Points lie on either side of the regions are known as support vectors [8]. The main goal of this model is to separate the region into two different sets and justifying a hyper line in a hyper plane in which N number of features vector is used. To minimization of error, we maximize the distance of the positive and negative points of each part. SVM classifier is used for the classification of a huge margin [10]. SVM used a discriminative function and it is represented by mathematical question given as in the following equation 3.2.

$$(3.2) \quad F(x) = \text{sign}(w^2x + b).$$

In the equation Number 3.2, we define the feature, weights and bias vector by x, w and b respectively. by x, weights vector by w and the bias vector [8]. We have analyzed different machine learning techniques and conclude the following useful details which are shown in figure 3.

3.3. K-Nearest Neighbors. KNN is a supervised machine learning Algorithm used in both cases such as regression and classification [7]. We know that this model can be easily implemented on the given prepared datasets. it is based on distance functions or similarity measures [9]. We use the labelled data as input in KNN Algorithm. This model can be trained for learning the pattern of the dataset for predicting the desired result as required. It gives new unlabeled data as an output [8].

Reference No.	Supervised Machine Learning	Unsupervised Machine learning	Reinforcement
5.	— Training data — Inputs and outputs data. — Labelled data — Clear end goal — Classification.	— It required inputs data. — No clear end goal — Clustering	— Improve based on Reward system during training
6.	— Regression (linear, non linear) — Decision tree — Naive Bayes — K-nearest neighbours	— Data Compression — K-mean etc.	— Models learn through trial and error to achieve the maximum possible reward
7.	— SVM etc.		— To build Policy — Markov decision process etc.

FIGURE 3. Comparison of various sentimental analysis and Machine learning approaches

4. CONCLUSION AND FUTURE WORK

In this paper, we analyzed different sentiment classification techniques and their working on various datasets and fields. Our discussion and reviews are based only on textual data and extracted from the social networking sites. We found in social networking sites that most of the texts are shorthand's, ambiguous and their meaning is not so clear. Step by step analysis of Machine learning techniques on social data retrieved for pre-processing, clustering, and further classification .we got different parameters where specific machine learning algorithms works and sometimes we merge more than two algorithms to solve real-life complex problems. In the future we have a lot of scope in the filled with sentiment analysis for further research, there are several challenging problems in the field of sentiment analysis exists that have not solved yet. Most researchers are only focused on tweets retrieved from twitter social networking sites but few are only focused on their replies for predicting the user's behaviour. We will be developed a social computing model for the prediction of human or social networking user's behaviours or (sentiments) in the future considering few machine learning techniques as a reference model.

REFERENCES

- [1] K. SAILUNAZA, R. ALHAJJ: *Emotion and sentiment analysis from Twitter text*, Journal of Computational Science, **36** (2019), art.id. 101003.
- [2] G. ADHITHYA BALAJI, R. KHILAR: *Twitter Sentiment Analysis: Comparison of Techniques*, Test Engineering & Management, (2020), 3223-3225.
- [3] B. LUO, J. ZENG, J. DUAN: *Emotion space model for classifying opinions in stock message board*, Expert Systems with Applications, Elsevier, **44** (2016), 138–146.
- [4] S.M. LIU, J.-H. CHEN: *A multi-label classification based approach for sentiment classification*, Expert Systems with Applications, Elsevier, **42** (2015), 1083–1093.
- [5] P. TYAGI, R.C. TRIPATHI: *A Review towards the Sentiment Analysis Techniques for the Analysis of Twitter Data*, International conference on advanced computing and software engineering (ICACSE-2019).
- [6] I. CHATURVEDIA, E. CAMBRIA, R.E. WELSCHB, F. HERRERA: *Distinguishing between facts and opinions for sentiment analysis: Survey and challenges*, Information Fusion, **44** (2018), 65–77.
- [7] S. SOHN, W. KIM, D.C. COMEAU, W.J. WILBUR: *Optimal Training Sets for Bayesian Prediction of MeSH® Assignment*, Journal of the American Medical Informatics Association, **15**(4) (2008), 546-553.
- [8] Y.-N. CHEN, C.-A. LU, C.-Y. HUANG :*Anti-Spam Filter Based on Naive Bayes, SVM, and KNN model*, 2009 AI TERMPROJECT.
- [9] D. GAMAL, M. ALFONSE, E.-S.M. EL-HORBATY, A.-B.M. SALEM: *Analysis of Machine Learning Algorithms for Opinion Mining in Different Domains*, Machine Learning & Knowledge Extraction, **1** (2018), 224-234.
- [10] M.V. MANTYLA, D. GRAZIOTIN, M. KUUTILA: *The evolution of sentiment analysis: A review of research topics, venues, and top cited papers*, Computer Science Review, **27** (2018), 16-32.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, DR. A.P.J. ABDUL KALAM TECHNICAL UNIVERSITY & INTEGRAL UNIVERSITY, LUCKNOW, UTTAR PRADESH, INDIA

Email address: sudheerhbtisomvansi@gmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING,, HARCOURT BUTLER TECHNOLOGICAL UNIVERSITY, KANPUR, UTTAR PRADESH, INDIA

Email address: pvluk@yahoo.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, SHRI RAMSWAROOP GROUP OF PROFESSIONAL COLLEGE, LUCKNOW, UTTAR PRADESH, INDIA

Email address: pk79jan@gmail.com