Advances and Applications in Mathematical Sciences
Volume 21, Issue 8, June 2022, Pages 4575-4584

AN INVESTIGATION OF MACHINE LEARNING ALGORITHM IN SENTIMENT ANALYSIS

A. SATHYA[1] and M. S. MYTHILI[2]

[1]Research Scholar, [2]Associate Professor
Department of Computer Science
Bishop Heber College (Autonomous)
Affiliated to Bharathidasan University
Tiruchirappalli – 620017, Tamil Nadu, India
E-mail: sathya.cs.res@bhc.edu.in
mythili.ca@bhc.edu.in

Abstract

In recent times, the use of social networking sites has been improved tremendously. Millions of people express their views and opinion through blogs or some microblogging websites. Social media is one of the handy gateways where the user can publish views and opinions. Sentiment Analysis is defined as the technique of mining data, reviewing a sentence via Natural Language Processing (NLP). It involves classifications of textual content into three phases Positive, Negative, and Neutral. Twitter is one of the most powerful social media where people can collect tweets using Twitter API based on particular keywords. The people can convey their opinions and thoughts openly in different sectors like healthcare, banking sector, sports, politics, online shopping, and tourism. In this paper, the levels, techniques, applications, and challenges in the sentiment field and a comparison of machine learning algorithms like Support Vector Machine (SVM), Naïve Bayes (NB) based on recent research problems are discussed. Finally, the sentiments are identified and reviewed that SVM leads to better accuracy compared to other algorithms.

1. Introduction

In recent days, social network sites like Facebook, Instagram, and Twitter have become very popular. People can share their opinions and feelings with everyone with the help of social media. Sentiment Analysis is one of the curious areas that can classify human expression on distinctive things. It

2020 Mathematics Subject Classification: 03B70.
Keywords: Sentiment Analysis, Machine Learning, Support Vector Machine, social media.
Received March 24, 2022; Accepted April 4, 2022

helps to make a proper decision before taking any action. Sentiment analysis is mostly considered in areas of application like healthcare, financial sector, sports, politics, hospitality etc. [1]. Social Media reviews and Sentiment Analysis helps

- In the market, to know the customer feedback to improve the productivity. In travel divisions, to measure traveler satisfaction and preferences [2]. In healthcare, pandemic widespread issues are observed and to keep the open alarm and secure.
- In Education, the students and mentors can feed the opinion through online media to improve the education system. Social media provides opportunities to improve methods for students and learning institutions to give better education. In Politics, the politicians can predict the output of the election result also check the current conclusion, and can check the status of the restricted party [3].

Multiple languages opinion with the different geographical areas on social media increases the complication of Sentiment analysis in the levels of accuracy and consistency [4]. Issues like noisy text, missing punctuation, sarcasm, spam reviews, identifying interrogative sentences in a critical document are the biggest challenges of sentiment analysis [15].

To overcome such complications, different sorts of machine learning approaches like Naïve Bayes, Support Vector Machine (SVM), and Regression are applied to classify the opinion. Most of the information accessible on the web is unstructured. So different preprocessing steps are connected to bring the information in structure format before applying sentiment analysis. This also involves selecting the proper data by removing useless data, noisy words, etc. Sentiment Analysis is also known as opinion mining, review mining, since the synonyms are similar but sometimes it may slightly vary depending upon the levels of the problem.

2. Related Works

In this section, the related studies in the field of sentiment analysis using machine learning techniques are discussed.

Abdul Mohaimin Rahat et al. [1] compared machine learning algorithms

like Naïve Bayes and support vector machines for sentiment classification of airline customer reviews. First, the collection of datasets using Twitter API is done. Then preprocess the review text from the trained model and the presented algorithm with real-world datasets gives better accuracy. Nikhil Kumar Singh [2] surveyed the summary of supervised sentiment analysis and preprocessing techniques are applied finally SVM proves better performance with levels of accuracy and consistency.

Huma Parveen et al. [5] discussed preprocessing techniques and Naïve Bayes algorithm also applied to find an opinion or emotions of people for better accuracy. Rajkumar S. Jagdale et al. [11] discussed the extraction and detection of sentiment from the text also applied the classification methods and shows the accurate measurements of SVM and Naïve Bayes. Vishal A Kharde et al. [15] surveyed the machine learning approaches and a lexicon-based approach with Twitter data and probabilistic classifiers like Naïve-Bayes and SVM are applied for better result. K. Rajarajeshwari et al. [16] have reviewed and extract opinions from customers and words alignment model, double propagation, shallow semantic parsing, conditional random field methods are used for better accuracy. Rajeswara Rao et al. [18] explained the election prediction result using machine learning algorithms. Tokenization has been applied to convert lengthy text into simple tokens. Finally, the support vector machine proved the highest accuracy than Naïve Bayes. Rupinder Kaur et al. [19] focused on the reaction of Twitter users for the Union Budget of India. Text preprocessing is applied for removing some unnecessary words and symbols to make the data at the standard level. They have suggested a custom dictionary of words to be created for better sentiment analysis in the future. Rajesh Bose et al. [20] discussed the Gujarat legislative elections 2017 to predict the result by public opinion by extracting the tweet data. Data has been collected by Twitter streaming API then removing all hashtags, URLs by preprocessing all data to give the standard output. Parallel DOT tool was used or deep learning to predict the accuracy of the result. N H Abd Rahim et al. [21] discussed about the data preprocessing techniques like Tokenization, Normalization, and Lemmatization to make data at a standard level. Finally, the SVM classification algorithm proved one of the highest accuracy performances. B. Edukondalu et al. [22] discussed the use of JIO tweets as a data stream. ADWIN sliding window algorithm detects

and verifies the changes in tweeted words to verify the sentiment orientation of the tweets. Pankaj Verma et al. [23] discussed about the data extraction, data preprocessing methods to remove all pointless data. Finally, it was observed that individuals appreciate government approaches based on the assumption scores of the polarity. Thus, machine learning and deep learning can be applied for sentiment classification as a future improvement. Swarupa Kulkarni et al. [24] discussed the three levels of sentiment like document, sentence, and aspect to extract opinion or sentiment from tweets. They mainly focused tweets on the English language only. Preprocessing is the major task to convert all unstructured into structured data and reduce irrelevant data. Reena G. Bhati [25] explained lexicon Based and machine learning approaches with different types of datasets. The Text level data can be categorized the emotions by Natural Language Processing method for better improvement.

Methodology

Levels of Sentiment Analysis

Sentiment analysis is classified into three levels such as document-level, the Sentence-level and the aspect level or phrase-based.

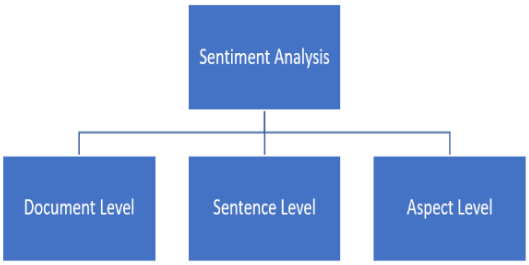

Figure 1. Levels of Sentiment Analysis.

Document Level

Like its name, it analyzes and classifies only the document. Here the whole document should be analyze and express the sentiment as Positive and Negative. If any irrelevant review is analyzed, it should be eliminated before processing [11] [15]. For example "I bought a dashboard camera yesterday".

The quality is good. I simply like it." Here the text "Good" and "like" conveys a positive attitude.

Sentence Level

This level analyzes and determines the sentence is positive, negative, or neutral opinion. For example "The dashboard camera is cool". In this sentence the text "cool" conveys a positive sentiment.

Aspect Level:

It is also known as Feature level. It particularly focuses on attributes or components of a data or service [11] [15]. For example, "The dashboard camera display is good, but life is very short". Here the attribute "display" gives positive sentiment but "life of camera" conveys negative sentiment.

Sentiment Classification Techniques

Sentiment classification can be divided into three categories. They are Machine Learning, Lexicon–based and Hybrid approaches.

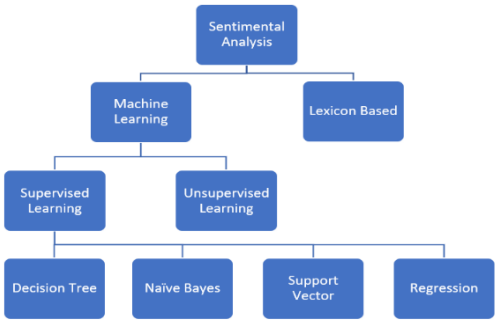

Figure 2. Classification of Sentiment Analysis.

Supervised Learning

Supervised Learning use labeled training documents. A pre-defined training data is used to predict the class of a document. From figure (2) the supervised model can be segregated as decision tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), and Regression.

1. Decision Tree: A decision tree is a decision support tool that contains conditional statements using a tree-like graph. Due to its lowest cross-validation error, high stability, and accuracy will be classified [3] [6].

2. Naïve Bayes: Naive Bayes is a collection of classification algorithms based on the Bayes Theorem. It is also known as Naïve Bayes or independence Bayes. Due to its speed and nature of simplicity, it has highly recommended for classification [13].

3. Support Vector Machine: SVM is one of the best ways of text classification using the hyperplane. It is defined as both input and output format. Output is either positive or negative [9]. SVM classifier is very expensive moreover the execution speed will be slow mode [14].

4. Regression: Figure (3) shows it is a statistical method to model the relationship between the dependent (target) and independent (predictor) variable with one or more independent variables.

Figure 3. Regression.

Unsupervised Learning

Unsupervised Learning does not depend on any domain or topic of training data. It overcome the difficulty of collecting and creating labeled training data.

Lexicon-Based Approach

A lexicon-based approach [15] is an approach encompasses a high classification speed to identify the polarity as positive, negative, or neutral [7]. Lexicons are nothing but a cluster also classified as a Dictionary-based and corpus-based approach [10] [15]. The dictionary approach makes use of an existing dictionary but corpus-based deals with the probability of occurrence of a sentiment word in combination with a positive or negative set of words [8].

5. Comparative Analysis and discussion on Machine Learning Algorithm

Author Name	SVM %	NB %
Abdul Mohaimin (2019)	82.48%	76.56%
Nikhil Kumar Singh (2018)	83.00%	77.00%
Joylin Priya Pinto (2019)	82.00%	74.56%
Rajkumar S Jagdale (2019)	93.54%	98.17%
Vishal A Kharde (2019)	76.68%	74.56%
Dr. D Rajeswara Rao (2020)	79.8%	77.40%

Table 2. Comparison of algorithms in terms of accuracy.

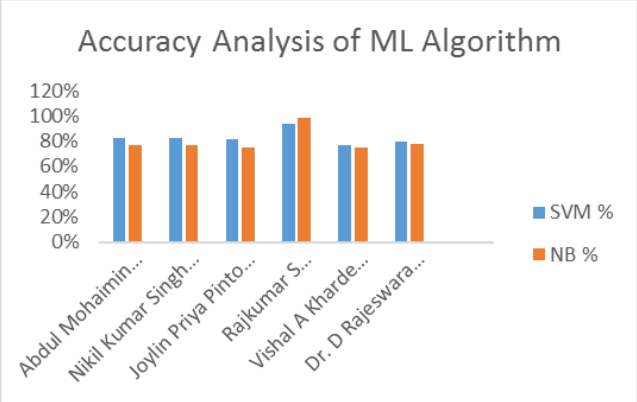

Figure 6. Performance comparison of Machine Learning Algorithms in Sentiment Analysis.

In this paper, bar maps are used to represent the output to emphasize accuracy. From Fig (6) Support Vector Machine leads the advanced accurateness than Naïve Bayes algorithm among analyzation of various researchers.

Table 3. The accuracy level of the SVM classifier.

SVM Classifier	Accuracy
SVM Polynomial Kernel	87.00%
SVM RBF Kernel	91.00%

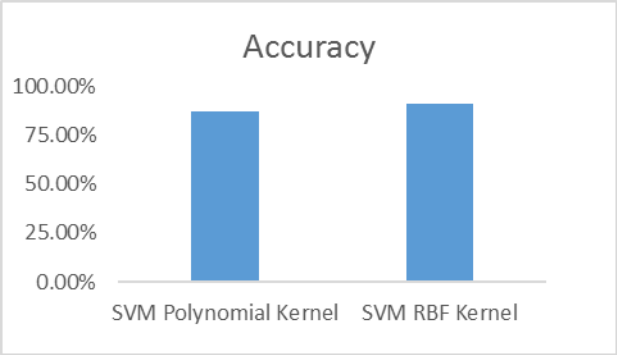

Figure 7. Graphical view of SVM classifier.

Figure (7) shows the exhibition of the classifier involving the two kinds of parts in SVM. Given Table 2. RBF kernel shows higher accuracy than the polynomial kernel [21]. As it can be seen, the Support vector machine comes out with the best result compared to another classifier among various researchers under different sentiment problems [1]. SVM is the best approach with all complex problems in real-world sentiment [9]. Our analysis shows that the SVM is the best model which minimizes the risk, takes less time for computation to give a better result.

6. Conclusion

An investigation has been made on Sentiment Analysis, looking at it from different dimensions like politics, injustice, inhumanity, international affairs, economy, natural disaster, terrorism. This paper summarizes the levels of analysis, techniques, applications, research issues, and compares the machine learning algorithms under Sentiment Analysis. Research results show that the classification algorithm such as SVM leads the highest accuracy and among SVM classifiers RBF Kernel proves better. Subsequently, Sentiment Analysis remains a promising zone of research for future demand.

References

[1] Abdul Mohaimin Rahat, Abul Kahir and Abu Kaisar Mohammad Masum, Comparison of naïve Bayes and SVM algorithm base on sentiment analysis using review dataset, IEEE 23rd Nov. (2019), 266-270.

[2] Nikhil Kumar Singh, Deepak Singh Tomar and Arun Kumar Sangaiah, Sentiment analysis: a review and comparative analysis over social media, Journal of Ambient Intelligence and Humanized Computing, Springer (2018), 97-117.

[3] Mohd Zeeshan Ansari, M. B. Aziza, M. O. Siddiqui, H. Mehraa and K. P. Singha, Analysis of political sentiment orientations on twitter, International Conference on Computational Intelligence and Data Science (ICCIDS 2019) Procedia Computer Science 167 (2020), 1821-1828.

[4] Floradel S. Relucio and Thelma D. Palaoag, Sentiment analysis on educational posts from social media, IC4E 2018, January 11-13, 2018, San Diego, CA, USA, © 2018 Association for Computing Machinery, 99-102.

[5] Huma Parveen and Prof. Shika Pandey, Sentimental Analysis on Twitter data-set using Naïve Bayes Algorithm, International Conference on Applied and Theoretical Computing and Communication Technology (2016), 416-419.

[6] Nuha Elamin, Samin A. Talab and Ahmed Khalid, Sentiment analysis with supervised learning techniques, Indian Journal of Science and Technology B(03) (2020), 249-268.

[7] T. Nikil Prakash and A. Aloysius, Applications, Approaches, and Challenges in Sentiment Analysis (AACSA), IRJMETS 02(07) (2020), 910-915.

[8] R. Cynthia Monica Priya and Dr. J. G. R. Sathiaseelan, An explorative study on sentimental analysis, 19th Oct 2017, WCCCT, 140-142.

[9] Joylin Priya Pinto and T. Vijaya Murari, Real-time Sentiment Analysis of Political Twitter Data Using Machine Learning Approach, IRJET 06 04th Apr (2019), 4124-4129.

[10] R. Aishwarya, A. Ashwatha, C. Deepthi and Beschi Raja, A Novel Adaptable Approach for sentiment Analysis, IJSCSEIT (2019), 254-263.

[11] Rajkumar S. Jagdale, Vishal S. Shirsat and Sachin N. Desh Mukh, Sentiment analysis on product reviews using machine learning techniques, Advances in Intelligent Systems and Computing, Springer (2019), 639-647.

[12] Najma Sultana, Pintu Kumar and Monika Rani Patra, Sentiment analysis for product review, International Journal of Soft Computing 1913-1919.

[13] Ankur Goel, Jyoti Gautam and Sitesh Kumar, Real-time sentiment analysis of tweets using naïve bayes, 2nd International Conference on NGCT Dehradun, India, (2016), 257-261.

[14] Monika Kabir, Mir Md. Jahangir Kabir, Xu Shuxiang and Bodrunnessa Badhon, An empirical research on sentiment analysis using machine learning approaches, International Journal of Computer Applications, (2019).

[15] Vishal A. Kharde and S. S. Sonawane, Sentiment analysis of twitter data: A survey of techniques, International Journal of Computer Applications 139(11) (2016), 1-10.

[16] K. Rajarajeshwari, A. Jenifer Jyothi Mary and Dr. L. Arockiam, Survey on aspect level opinion target from online reviews, IJITME 2 (2016), 1-10.

[17] B. Nagajothi and Dr. R. Jemima Priyadarsini, Sentiment analysis on twitter dataset using R language, IJTSRD 3(6) (2019), 199-204.

[18] D. Rajeswara Rao, S. Usha, S. Sri Krishna, M. Sai Ramya, G. Sri Charan and U. Jeevan, Result prediction for political parties using twitter sentiment analysis, International Journal of Computer Engineering and Technology (IJCET) 11(4) (2020), 1-6.

[19] Rupinder Kaur, Rajvir Kaur, Manpreet Singh and Dr. Sandeep Ranjan, Twitter sentiment analysis of the indian union budget, International Journal of Advanced Science and Technology 29(4s) (2020), 2282-2288.

[20] Rajesh Bose, Raktim Kumar Dey, Sandip Roy and Debabrata Sardar, Analyzing political sentiment using twitter data, Information and Communication Technology for Intelligent Systems, Springer (2018), 427-426.

[21] N. H. Abd Rahim and S. H. Mohd Rafie, Sentiment analysis of social media data in vaccination, International Journal of Emerging Trends in Engineering Research 8(9) (2020), 5259-5264.

[22] B. Edukondalu and P. Neelima, Sentiment analysis on social media network, International Journal on Future Revolution in computer science and communication Engineering, IJFRCSCE 6(2) (2020), 01-08.

[23] Pankaj Verma and Sanjay Jamwal, Mining public opinion on Indian government policies using R, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, 9(3) (2020), 1310-131.

[24] Swarupa Kulkarni and Priyanka Kedar, A survey on twitter sentiment analysis, Open Access International Journal of Science and Engineering, OAIJSE 5(4) (2020), 22-25.

[25] Reena G. Bhati, A survey on sentiment analysis algorithms and datasets, Review of Computer Engineering Research 6(2) (2019), 84-91.