IJAITE

INTERNATIONAL JOURNAL OF ADVANCED INNOVATIVE TECHNOLOGY IN ENGINEERING

Published by Global Advanced Research Publication House Journal

Home page: www.ijaite.co.in

Sentiment Analysis Using Machine Learning Technique

[1]Shital Shivshankar Durudakar, [2]Prof. Tarun Yengantiwar

[1,2]Department of Computer Science & Engineering, V. M. Institute of Engineering & Technology, Nagpur, Maharashtra, India

[1]shital.patrakar@gmail.com, [2]yengantiwar@rediffmail.com

Article History

Received on: 25 Feb 2023
Revised on: 10 March 2023
Accepted on: 30 March 2023

Keywords: social media, Machine Learning, Sentiment Analysis, Opinion Mining

e-ISSN: 2455-6491

DOI:

Production and hosted by
www.garph.org

ABSTRACT

Various machine learning algorithms for sentiment analysis are discussed in this study. Machine learning classifiers such as Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, KNN, and deep learning classifiers were used to analyze sentiment. We notice some articles in this section that are assisting young investigators in determining the best path for further study. Various social networking sites, E-commerce sites like Amazon and social media like Facebook, Twitter, and Instagram is popular platforms for users to express their views on many topics. Sentiment analysis employs a machine learning approach and provides a precise assessment of people's feelings without the need for human intervention. The Sentiment analysis divides the text into three categories: positive, negative, and neutral. As a result, any corporation, institution, an examiner can accept the public's view and take action based on it.

1. INTRODUCTION

Sentiment classification is the way that examines texts for polarity, ranging from positive to negative using the machine learning approach. Machine learning automatically learns human sentiments. Today, social media is an integral aspect of people's lives; they utilize it to share their opinions on many topics such as politics, film ratings, and advertisements. There are numerous social media platforms available, including Twitter, Facebook, Instagram, and more. They use these social media sites to share their opinions on a variety of issues. Therefore, using the training data set, sentiment analysis analyses the text entered by any individual from a particular location. It evaluates the sentiments of that certain text by understanding the sentiment of such a user.

The application of sentiment analysis is very broad and powerful like Expedia Canada; Canadians take advantage of sentiment analysis when they notice that people are giving negative comments on the music used by their television channel. Rather than chalking over negative comments, Expedia manages to take advantage of that negative comment and air all-new soulful music on their channel.

A. Levels of Sentiment analysis

Document-level: The entire text in documents is analysed at the document level. A paper that is solely focused on one subject is covered in that level of categorization. Consumers often believe that document evaluation cannot be used to evaluate two themes or two documents. For the categorization of document-level sentiment analysis, supervised and unsupervised machine learning algorithms are applied.

2. Sentence level: Sentiment analysis at the sentence level is strongly connected to subjective categorization. The goal of phrase-level sentiment evaluation is to identify if a sentence contains expressions that are positive, negative, or neutral. Sentence-level emotion analysis employs the whole classification from document-level sentiment classification.

3. Aspect level: The Aspect level sentiment analysis is used to find out the sentiment on the Aspect of those entities. “My car has good handling but it is a little heavy” let’s take this example. In this example, there is an opinion on a car that the handling of a cat is positive but the weight of the car is negative. The competitive statement is part of an Aspect level sentiment analysis.

4. Phrase level: Opinion terms are classified at the sentence level in the passage wherever they appear. Both of these have benefits and drawbacks, with the benefit being that the precise viewpoint of the attribute is present. Nevertheless, it involves a contextual polarities issue, thus the outcome might not be correct.

5. Feature Level: Product characteristics are defined as product features. Feature-level sentiment classification is the term of the document Evaluating these characteristics for recognising sentiments. The retrieved characteristics are used to determine if a comment is good, negative, or neutral.

The following is a description of this study significant contributions.

- To conduct a critical examination of sentiment analysis in various contexts.
- To conduct a thorough examination of multiple sentiment modeling techniques based on machine learning approaches, datasets, techniques, findings and other performance metrics.
- To identify important research problems and opportunities according to previously significant finding to sentiment classification.

The review on sentiment analysis classification is designed in the following manner: Section II specifies the literature review on conventional sentimental analysis in social media. Section III describes various machine learning algorithms for sentiment analysis along with performance measures. The analysis on different types of data used and tools for sentiment analysis is given in Section IV. The research gaps and challenges of sentiment analysis using machine learning algorithms are shown in Section V. Section VI specifies the conclusion of the entire paper.

2. RELATED WORK

In [1], Tweets are classified into positive or negative comments using machine learning algorithms such as Naïve Bayes, Random Forest (RF), Support vector machine (SVM), Unigram with Sentiwordnet, and unigram with Sentiwordnet including negations are used as the input in this paper. The author derived three thousand one hundred eighty-four (3184) tweets using the tweeter API. Nine hundred fifty-four (954) positive, one thousand eighteen (1318) negatives and 145 stop words have been identified from 3184 tweets. Using. The author used features of sentiment analysis like Bag of words (BOW), Term frequency vs Inverse document frequency (TF-IDF), Unigram with Sentiwordnet, Unigram with Sentiwordnet including negation words as an input. The author gets the conclusion that all the classifiers with Unigram with Sentiwordnet and Unigram with Sentiwordnet including negation word shows higher accuracy the Bags of words (BOW) and term frequency vs Inverse document frequency (TF-IDF). Random forest algorithm with Unigram with Sentiwordnet including negation words get the highest accuracy of 95.6%.

In [2] the authors try to use a machine learning algorithm for Arabic customer feedback. They study two different types of methods which are voting and meta classifier combination. They collect data using Tweepy Application Programing Interface (API)17. There are many sarcastic and neutral tweets with positive and negative tweets. A total of 438,931 tweets were collected from 75,774 positive and 75,774 negatives. Removing all noisy data from the tweets like pictures, hashtags, retweets, and emotions; second tokenization removing non-Arabic letters, normalizing Arabic analogue letters. 10 classifiers NB, ME, LR, RR, PA, MNB, SVM, SGD, and Ada boost BNB were used to extract and discover the polarity of given tweets. The highest accuracy achieved by PA and RR is 99.96%. The lowest accuracy achieved by Ada boost, LR, and BNB is less than 60%.

In [3] uses Amazon customer review data to find out the positivity, negativity, and neutrality of customer reviews. In this, they compare two machine learning algorithms Naïve Bayes algorithm and the Support vector machine (SVM). The input is the customer review of the Amazon products. The review may be negative, positive, or neutral. Apriori algorithm is used to extract the frequently used aspects from the input dataset. Sentiwordnet is used to calculate positivity, negativity, and neutrality scores and after that, the classifier will apply. The comparison of the algorithm based on the performance can be calculated by using the Accuracy, Precision, Recall, and F-1 Measure of each classification. By the experimental result, Naïve Bayes classification is a batter accuracy than Support vector machine (SVM). The calculation was done by True positive sample (TP), False positive samples (FP), True negative samples (TN), and False-negative samples (FN).

In [4] many unsolicited email campaigns are one of the biggest threats affecting the users. The author combines both Sentimental analysis and personality recognition for analyzing the email content. They use two different datasets to validate the proposed method. The first dataset is the original dataset (CSDMC 2010 dataset) and the second dataset validation dataset (TREC 2007). CSDMC 2010 spam corpus: This is composed of 2949 email messages to carry out original experiments. TREC 2007 public corpus: - In this there are 75419 emails in which 25220 are legitimate 50199 spam emails. This method was validated in two different datasets improving the best accuracy in both cases (from 99.15% to 99.24% and 98.98% to 99.18%). Further, this method is also used for different validation like SMS and social media validation.

In [5], shows; During the pandemic the COVID19 whole world is suffering. Social media is a vast platform to share your thoughts in any situation. The author uses social media to analyze people's reactions to this situation. The author portrays the fact that how irrationally people are behaving in this situation. It would be easier for the victim to gather some structured information from social media. Two sets of datasets have been used in this paper. #Corona, #covid19, # were coronavirus mostly used for this survey. In dataset-1 there were 2,26,668 tweets used as the preliminary for dataset-2 they use the tweets which were retweeted most. To fit in the model data have been categorized in a train, validation, and test sets. To show the accuracy of unigram, bigram and trigram were performed. The accuracy of dataset 1 is 81% and the accuracy of dataset 2 is 75% using different classifiers. In the conclusion, the author came to know that social media is not useful enough to help people.

In [6], the author examines the Alzheimer's disease stigma on Twitter using machine learning techniques. Machine learning technique modeled stigmatization expressed in 31150 Alzheimer's disease-related tweets collected via tweeter API. In this 1% of the dataset is used to train a classifier of the tweet and the rest is 99% of the dataset. In this paper, the author discusses how social media outlet affects attitude bearing in other development outcomes. The retweet was removed, other tweets which are not related to Alzheimer's were removed, the keywords "all", "Alzheimer", "dementia", "memory loss", and "senility" defined the sample of analysis. Lastly, they removed the username which contains the topic name they removed. Two researcher manual coding and result are as follow 43.41% informative, 23.79% joke, 21.22% metaphorical, 19.29% organization, 24.50% ridicule.

3. Machine Learning Approach

The labelled polarity dataset, which contains 1500 positive and 1200 negative reviews, has been evaluated. 15. Each review initially goes through a data preparation stage when all the ambiguous details are eliminated. Possible characteristics are retrieved from the cleaned dataset. These properties must be changed to numerical representation due to the phrases in the papers. Using vectorization methods, textual information may be transformed into a numeric representation. By the use of vectorization, a matrix is produced in which each column corresponds to a characteristic and each row to a specific comment. The categorization method uses this matrix as an input, and the cross-validation approach is utilized to select the training and testing dataset for every fold. Graphical Representation displays the sentiment analysis strategy in a step-by-step manner as shown in figure 1.

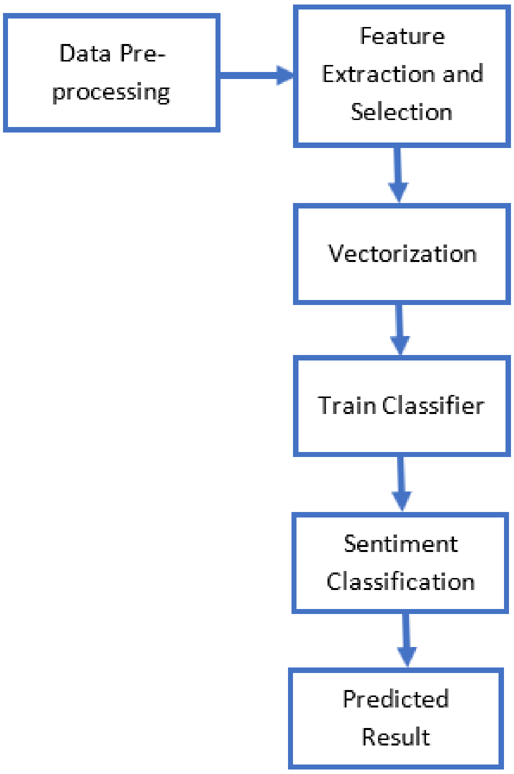

Figure 1: Steps of Sentiment Analysis

Steps Followed for Classification

Step 1: The study uses the 1500 positive and 1200 negative labelled comments from the polarities review in the dataset. A unique document is kept after every evaluation.

Step 2: There is a lot of ambiguous material in the reviews that need to be removed. First, during the data preparation stage, any special characters (such as (!@)) and extra spaces are eliminated. It has been noted that authors frequently use the same character from a term to emphasize a point or to create the review seem more current. These words include woooww and oohh. During this phase, the character redundancy is likewise done away with. Stop words refer to the majority of words utilized in English that don't express any feeling. Hence, eliminating every one of the English or other language's stop words in the second phase of preparation.

Step 3: The dataset may be used to retrieve relevant features after the third phase of cleaning it. The comment or review features are tokenized phrases. To express each review as a set of numeric values, such words must be transformed into numerical vectors.

CountVectorizer: The review is converted into a token counting vector. The review is initially tokenized, then a sparse matrix is made based on the frequency of each word.
TF-IDF: Its value reflects how significant a word is to a corpus of documents. The TF-IDF value correlates with a phrase's occurrence in a text.

Step 4: The categorization method can receive the numerical vectors as input. For categorization, several machine learning methods will be applied.

Step 5: A confusion matrix is created once the model has been trained, and it displays the percentage of positive and negative comments that were properly forecasted as well as the percentage of positive and negative comments that were incorrectly forecasted. This confusion matrix is used to compute the predictive performance for every fold, and the ultimate performance is determined by averaging all the specific efficiency values for the 10 folds. The particular precision of a given fold, therefore, may be far greater than the average of all levels of accuracy score.

Step 6: The various evaluating parameters such as precision, recall, and F-1 score are determined for classifier performance. The effectiveness assessment criteria report and the confusion matrix is generated. Lastly, the results acquired by other researchers in the research are examined to the findings found here.

Table 1: Summary of Related Work of Sentiment Analysis

Ref.	Data set	Techniques	Parameters
Elankath et. al. (2023)	Malayalam	BERT	Acc. = 88%
Soumya S. et. al. (2020)	Malayalam	NB, SVM, and RF	Acc=95.6%.
Gamal D. et. al. (2019)	Arabic	ML algorithms, 10-fold cross-validation	Acc= 99.6%

Vanaja S. et. al. (2018)	Amazon products review	NB, SVM	Acc=90.42%
			Acc=83.42%

Ezpeleta, E., I. et.al. (2020)	CSDMC2010	Spam Classifiers	Acc=99.24%
Chakraborty, K. et. al. (2020)	Covid tweets	fuzzy rule	Acc=63%

4. Application of Sentiment Analysis

A. Analyzing market survey

This entails keeping an eye on the what different innovations are being offered and what consumers are seeking out. So can adjust the business plan in light of that analysis.

B. To monitor the competitive market

To find out what their rivals are introducing or what products competitors are putting on the marketplace. to research the strategies of the opposition using popular sentiment. Among the key uses for sentiment analysis would be that.

C. Product Evaluation

To learn how customers think of the item after it has been released or to observe responses that have never been seen before. So may quickly assess a comment by searching the term for a certain characteristic of the item.

D. Analyzing social media

Individuals express their opinions on social media in a variety of contexts, including business, politics, the marketplace, and more. Users may quickly track people's attitudes from various perspectives by using sentiment analysis and a few key phrases.

E. Consumer Feedback

In every industry or firm, customer input is crucial. With sentiment analysis, a business may quickly see consumer feedback on an item and make adjustments to the item in response to the feedback.

5. Advantages of Sentiment Analysis

1. Less expensive than assistance for consumer feedback.
2. It is the quickest method of gathering data on consumer understanding.
3. Making use of sentiment analysis will make it simple to implement client suggestions.
4. It will be much simpler to pinpoint the advantages or disadvantages of other businesses or organizations.
5. The client's assessment will be highly precise.

Conclusion

Sentiment assessment is the process of gathering and evaluating sentiments, opinions, comments, Twitter posts, and emotive language in order to derive relevant knowledge. For sentiment analysis, it is necessary to organize and analyze hidden material that has been obtained from a variety of social media channels, including Instagram, Facebook, as well as other websites for social media. The use of deep learning and machine learning for sentiment analysis was described in this research. For sentiment analysis, numerous studies employed a variety of machine techniques, including decision trees, random forests, support vector machines, Adaboost, Naive Bayes, and logistic regression, among others. The deep learning framework consists of a number of effective and beneficial techniques that are used to address a variety of difficulties. In order to offer readers a complete knowledge of the enormous progress of the deep learning field of sentiment analysis, several previous works are examined in this paper.

Conflict Of Interest

The authors declare that they have no conflict of interest.

Funding Support

The author declares that they have no funding support for this study.

References

[1] Elankath, Syam & Ramamirtham, Sunitha. (2023). Sentiment analysis of Malayalam tweets using bidirectional encoder representations from transformers: a study. Indonesian Journal of Electrical Engineering and Computer Science. 29. 1817. 10.11591/ijeecs.v29.i3.pp1817-1826.

[2] Soumya, S. and K. J. I. E. Pramod (2020). "Sentiment analysis of Malayalam tweets using machine learning techniques."

[3] Gamal, D., M. Alfonse, E.-S. M. El-Horbaty and A.-B. M. J. P. C. S. Salem (2019). "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." 154: 332-340.

[4] Vanaja, S. and M. Belwal (2018). Aspect-level sentiment analysis on e-commerce data. 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE.

[5] Ezpeleta, E., I. Velez de Mendizabal, J. M. G. Hidalgo and U. J. L. J. o. t. I. Zurutuza (2020). "Novel email spam detection method using sentiment analysis and personality recognition." 28(1): 83-94.

[6] Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. J. A. S. C. Hassanien (2020). "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-A study to show how popularity is affecting accuracy in social media."97: 106754.

[7] Ahmad, M., S. Aftab, S. S. Muhammad and S. J. I. J. M. S. E. Ahmad (2017). "Machine learning techniques for sentiment analysis: A review." 8(3): 27.

[8] Arulmurugan, R., K. Sabarmathi and H. J. C. C. Anandakumar (2019). "Classification of sentence-level sentiment analysis using cloud machine learning techniques. 22(1): 1199-1209.

[9] Chaturvedi, S., V. Mishra and N. Mishra (2017). Sentiment analysis using machine learning for business intelligence. 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), IEEE.

[10] Hasan, A., S. Moin, A. Karim, S. J. M. Shamshir band and C. Applications (2018). "Machine learning-based sentiment analysis for Twitter accounts." 23(1): 11.

[11] Hassan, A. U., J. Hussain, M. Hussain, M. Sadiq and S. Lee (2017). Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression. 2017 International Conference on Information and Communication Technology Convergence (ICTC), IEEE.

[12] Kamal, A. and M. Abulaish (2013). Statistical features identification for sentiment analysis using machine learning techniques. 2013 International Symposium on Computational and Business Intelligence, IEEE.

[13] Mukhtar, N., M. A. J. I. J. o. P. R. Khan and A. Intelligence (2018). "Urdu sentiment analysis using supervised machine learning approach." 32(02): 1851001.

[14] Nasim, Z., Q. Rajput and S. Haider (2017). Sentiment analysis of student feedback using machine learning and lexicon-based approaches. 2017 international conference on research and innovation in information systems (ICRIIS), IEEE.

[15] Pang, B., and L. J. a. p. c. Lee (2004). "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Hammad, M. and M. Al-awadi (2016). Sentiment analysis for Arabic reviews in social networks using machine learning. Information technology: new generations, Springer: 131-139.

[16] Ahmad, M., S. Aftab and I. J. I. J. C. A. Ali (2017). "Sentiment analysis of tweets using SVM." 177(5): 25-29.

[17] Ahmed, E., M. A. U. Sazzad, M. T. Islam, M. Azad, S. Islam and M. H. Ali (2017). Challenges, comparative analysis, and a proposed methodology to predict sentiment from movie reviews using machine learning. 2017 International Conference on Big Data Analytics And Computational Intelligence (ICBDAC), IEEE.

[18] Devi, G & Somasundaram, Kamalakkannan. (2020). Literature Review on Sentiment Analysis in Social Media: Open Challenges toward Applications. 29. 1462-1471