Superupervised Enumerable Learning for Sentiment Analysis of Social Media Tweets

Zeenat Bano, Yogadhar Pandey, Bhupendra Verma

Technocrats Institute of Technology Science, Bhopal, M.P, India

Corresponding author: Zeenat Bano, Email: banozeenat7568@gmail.com

People express their views on goods, services, governments, and events via words and phrases on social media. Sentiment analysis is a technique used in natural language processing to gather positive and negative feedback on a piece of writing on social media. Academics have been forced to research sentiment analysis by exponential growth in business and e-commerce companies. This study examined emotions in two distinct domains: one for electronics product evaluations and another for film reviews. Due to the enormous volume of data gathered for analysis, it is referred to as massive data. This vast data is multidimensional and includes useless information. Furthermore, this data complicates the analysis since it reduces the effectiveness of machine learning categorization algorithms. As a consequence, the data must be cleansed and pertinent features identified. Numerous methods have been used to eliminate noise from data and identify essential components. There are two methods for determining suitable features: feature selection and feature extraction. The research is split into three stages: the first step involves converting unstructured data to structured data and choosing relevant and feature information. Second, machine learning classification algorithms are employed in conjunction with feature selection approaches, and third, machine learning algorithms are used in conjunction with ensemble learning methods.

Keywords: Social Media, Sentiment Analysis, Natural Language Processing, Machine Learning, Supervised Classification, Feature Extraction Super.

2023. In Saroj Hiranwal & Garima Mathur (eds.), *Artificial Intelligence and Communication Technologies*, 1083–1091. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.52458/978-81-955020-5-9-103

1. Introduction

Social media and e-commerce websites are viral these days. On e-commerce websites, a high volume of customer reviews has been generated. An appraisal is a statement of a person's opinion intended to assist another user in making a purchasing choice. Customer evaluations of products and services have developed into a significant source of market data. These evaluations help develop an e-commerce website's strategy. Commercial websites such as Amazon, Flipkart, and Snapdeal have large amounts of reviews. The consumer reads product reviews to choose whether or not to buy, but it's difficult for them to read all evaluations and distinguish between good and negative feedback.

Practical algorithms were required to examine this massive amount of data. Sentiment analysis is sometimes referred to as opinion mining. It is the computer study of people's views, feelings, emotions, and attitudes. As a result of these findings, a new task has been established, motivating us to develop a method for reviewing data. These are sometimes referred to as sentiments. Positive and negative emotions are included in this category. In real-world applications, sentiment analysis or opinion analysis is critical. Therefore, sentiment categorization is beneficial for developers and researchers working in information retrieval, natural language processing, and machine learning.

Everyone has internet access in today's digital environment. The number of users is growing fast due to the development of new apps and lower-cost technology. These users contribute significantly to social media's success. The proliferation of user-generated material on websites is a significant issue. Big data is a term that refers to a large volume of reviews. The massive data set contains useless information and tens of thousands of characteristics. As a result, it is critical to clean up data by removing redundant feature information and noise. These types of data are very challenging to evaluate since they lead classification techniques astray, producing unexpected findings and degrading the efficiency of classification algorithms.

Additionally, efficient methods must be used to analyze this data. Thus, many reviews present a significant challenge for sentiment analysis when identifying the polarity of a document's content: positive, negative, or neutral consumer feedback or stated opinion. Thus, polarity detection is a critical issue in sentiment analysis.

Additionally, film evaluations provide a great deal of information. With millions of reviews for a single film, personally evaluating each review will take a significant amount of time. Manual analysis may potentially have issues with incorrectly counting the number of studies due to human error. This is a complicated subject regarding the efficacy of sentiment analysis. The accuracy of sentiment analysis based on machine learning is currently very low.

Many customer reviews or a thousand reviews per product are included in the provided customer review. It is unstructured and semi-structured textual data or information. Additionally, essential and irrelevant feature information is included in the customer review data. Manual examination of review data is impractical due to the high probability of human error. Methods to extract meaningful features from multidimensional data cause machine-learning systems to perform poorly. Classified review data to enhance the effectiveness of machine learning classification algorithms. Using machine learning classification methods, how to quickly and effectively categorize user sentiment into cheerful (e.g., better) and negative (e.g., terrible) emotion.

2. Proposed Work

The purpose of this article is to extract valuable data and characteristics from a massive data collection. In this study, we classified review data sets using machine learning sentiment classification techniques. Ensemble learning techniques were utilized to enhance the performance and accuracy of machine learning algorithms used to categorize reviews as positive or negative in sentiment. We employed the preprocessing method to convert unstructured data to structured data. Additionally, relevant characteristics are selected using the feature selection techniques from product and movie review data sets. We classified product and movie review datasets using machine learning classification techniques into positive and negative sentiment. We integrated feature selection and classification techniques to achieve high accuracy and performance. Ensemble techniques have been used to enhance the performance and accuracy of classification systems. Ensemble techniques are used with classification algorithms to rapidly and accurately categorize review datasets. Additionally, ensemble techniques are utilized to boost the performance and accuracy of classification algorithms.

Initially, ensemble learning was suggested for supervised learning's categorization problem [114]. Ensemble learning's basic concept is to teach base learners (or machine learning algorithms). The researchers and scientists conducted tests to categorize the data using ensemble learning. And they have accomplished incredible things, as shown in [115]. These trained learners address the same issue as ensemble techniques. Machine learning, for example, generates a single hypothesis from the training data. The ensemble technique generates a large number of views and then aggregates them for usage. Technically speaking, the ensemble approach outperforms the individual base learner. Machine learning created models one at a time using training data. Rather than that, the ensemble approach makes many models simultaneously and then combines them to create a robust and powerful model [62].

3. Data Preprocessing

To make unstructured data information to structured data information used preprocessing technique. The preprocessing steps used tokenization, stop-word- removal, stemming word, and text transformation. The collected review data contains noise or irrelevant information. This noise must be removed by using the preprocessing technique. The preprocessed data can improve the performance of text retrieval, classification, and summarization [12]. Fig. 1 describes the data preprocessing technique.

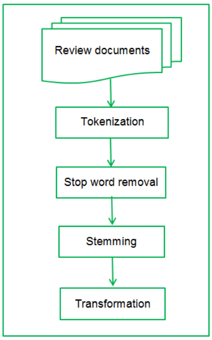

Fig1: General structure of data preprocessing steps.

3.1. Tokenization

This step performs the tokenization process. The review text split or converts in order of tokens. We can say each word is known as a token. This token has self-determining existence in each sentence of a document. Consider a sentence as “this mobile is very attractive and powerful”. This sentence converts into tokens such as “this”, “mobile”, “is”, “very”, “ attractive”, “and”, “powerful”. Natural Language Processing must follow this essential step [104].

3.2 Stop Word Removal

The stop word removal is also an important step to make relevant information about text documents [8]. The text review contains many words that are meaningless or irrelevant. These words are like “the”, “a”, “and”, “of”, “my”, “me”, “it”, “about”, “I”, “he”, “she” etc. These words do not carry any sentiment information. The natural language used this to interpret more easily of each sentence.

3.3 Stemming

Stemming is a process to find out the root of the word. It derives the actual word from their stem. The root of this word finds with the help of an algorithm. For example, the word flying, flew, flown to the stem word is "fly". The word played, plays, the root word is "play". "Porter‵‵s stemmer algorithm" is used to transform the words [105].

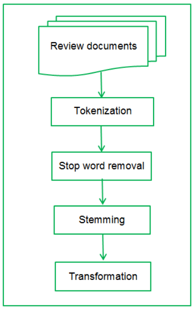

Fig. 2: General structure of data preprocessing steps.

4. Text Transformation

This process reduced the dual meaning of each word. Every word and character must be converted in the form of an upper case or lower case. The words stored in every text document by their occurrences. There are two most important approaches are presently available which is "vector space‵‵ and "bag-of- word"

5. Feature Identification

The review data collected from an online website is in the word form. This text data may contain irrelevant information. To make relevant information from irrelevant information to follow some steps which already discussed in the previous section 3.3. The relevant and meaningful attributes or features identify and select during the feature identification process from the text review dataset [7], [62], [43].

The mainly two advantages to select relevant attributes which are followed as:

- After selecting a feature, or attribute from text data, it increases the training process speed. It means this data processed quickly.

- This is also removed irrelevant texts or words from the dataset. So that increases the performance of classification algorithms.

The words, terms, or phrases represent the attributes of a product. The positive and negative sentiment classify based on attributes or features. This data convert into a structured form. Fig. 3.4 shows extract features or attributes from text documents. The product features determined in the next step. The main goal is to classify the polarity of opinion and determine that the product is good or bad. The following example discusses mobile phone text review opinions.

"This Redmi 7 Pro Mobile picture is very rich."

This opinion shows that the consumer is completely happy with the image quality of Redmi 7 Pro Mobile. The user is talking about the picture which is a feature.

6. Environmental Setup & Result Analysis

This study used the product evaluation data sets (camera, laptop, radio, television, and music) supplied by Hai et al. [27]. The review is presented as an unstructured tokenized text, with no structural information indicating the content's polarity (e.g., a 0/5 star rating). A review's average length is 746 words. Each review contains the reviewer's name, location, the product's name, the review title, the date, and the examine text. Positive reviews were classified as such, negative reviews as such, and the rest were removed due to their ambiguous polarity. Following that, each domain gets 1000 highlighted good and bad reviews on the Amazon website.

The confusion matrix was used to evaluate the experimental results and the performance of the developed method. Sentiment analysis is performed after classifiers have categorized the data sets. This emotion has many phrases or values. True positives, true negatives, false positives, and false negatives, for example, are all synonyms for the same phenomenon. Consequently, the words are beneficial for calculating or evaluating the average performance of the algorithms and classification methods.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Tables 1–4 provide experimental results regarding categorization techniques. Each classification method, including Nave Bayes, k-Nearest Neighbor, Maximum Entropy, and Support Vector Machine, was applied individually to the data set. Each classification technique combined ensemble approaches such as bagging, boosting, and random subspace for sentiment classification. As a consequence, the accuracy of all experimental tables is, on average, the highest. The table highlights the average accuracy of separate classifiers and combination classifier methods.

Table 1 summarises the average accuracy obtained via fundamental learning techniques. The Maximum Entropy classifier obtained the highest accuracy of 83.50 percent on the Camera dataset. The Maximum Entropy classifier has the best accuracy of 88.95 percent on the Laptop dataset. The Naive Bayes classifier obtained a maximum accuracy of 82.89 percent on the Music dataset. Maximum Entropy correctly identifies 84.76 percent of the Radio data set. Finally, the Support Vector Machine classifier performs admirably well on the TV dataset, with an accuracy of 84.73 percent.

Table 1: The base learner methods are achieved average accuracy

Datasets					
Methods	**Camera**	**Laptop**	**Music**	**Radio**	**TV**
NB	81.65	83.72	82.89	79.36	84.46
KNN	74.30	76.86	74.20	71.60	75.65
ME	83.50	88.95	65.43	84.76	68.87
SVM	82.49	85.60	80.14	81.38	84.73

Table 2 shows some intriguing results in terms of categorization accuracy. All ensemble techniques using base learners exhibited poor accuracy compared to base learners, as seen in Table 1. Because the bag of words transforms text information directly into spatial vectors, this might be the explanation. Many redundant and relevant characteristics and some noise may be found in these space vectors. The primary reason for this is that noisy data on laptop datasets affects the performance of maximum entropy with bagging.

Table 2: The average accuracy by combining base learner methods with the bagging method

Datasets					
Methods	**Camera**	**Laptop**	**Music**	**Radio**	**TV**
NB+Bagging	82.48	78.34	73.86	76.39	75.26
KNN+Bagging	62.25	65.83	67.24	68.80	69.59
ME+Bagging	81.68	**83.45**	81.18	80.13	76.89
SVM+Bagging	81.89	75.55	74.36	76.23	75.35

Table 2 shows the average accuracy of the base learning algorithm when utilising the bagging ensemble method. The Naive Bayes and Bagging classifier achieves a maximum accuracy of 82.48 per cent on the Camera dataset. Maximum Entropy and Bagging accurately categorise 83.45 per cent of the Laptop data set. The Maximum Entropy and Bagging classifier obtained a maximum accuracy of 81.81 per cent on the Music dataset. ME and Bagging correctly categorised 80.13 per cent of the Radio data set. On the TV dataset, the Maximum Entropy and Bagging classifier achieved the most impressive accuracy of 76.89 per cent.

Table 3 summarises the average accuracy of fundamental learning algorithms and the Boosting ensemble method. Support Vector Machine with Boosting achieves the best accuracy of 85.46 per cent on the Camera dataset. The Support Vector Machine and Boosting classifiers achieved the best accuracy of 88.58 per cent on the Laptop dataset. Support Vector Machine with Boosting achieves the best accuracy of 84.62 per cent on the Music dataset. The Support Vector Machine and Boosting classifiers on the Radio dataset achieve maximum accuracy of 86.86 per cent. Finally, SVM with Boosting classifier obtained an ultimate accuracy of 85.32 per cent on the TV dataset.

Table 3: The average accuracy is achieved by combining base learner methods with a boosting method.

Datasets					
Methods	Camera	Laptop	Music	Radio	TV
NB+Boosting	83.12	84.86	82.76	83.28	82.80
KNN+Boosting	8034	85.83	81.36	82.68	81.36
ME+Boosting	84.82	87.75	83.48	83.98	83.14
SVM+Boosting	85.46	**88.58**	84.62	86.86	85.32

The average accuracy of basic learning algorithms and the Boosting ensemble technique is shown in Table 4. The maximum accuracy achieved by Support Vector Machine with Random Subspace on the Camera dataset is 84.36 percent. On the Laptop dataset, the Support Vector Machine and Random Subspace classifier had a maximum accuracy of 87.24 percent. The Support Vector Machine and Random Subspace classifier had the most remarkable accuracy of 85.98 percent on the Music dataset. The Support Vector Machine and Random Subspace classifier achieved the maximum accuracy of 86.23 percent on the Radio dataset. Finally, the Support Vector Machine and Random Subspace classifier achieved the highest accuracy of 88.68 percent on the TV dataset.

Table 4: The average accuracy is achieved by combining base learner methods with random forest method.

	Datasets				
Methods	**Camera**	**Laptop**	**Music**	**Radio**	**TV**
NB+RS	82.80	84.29	81.96	83.80	83.98
KNN+RS	81.12	85.46	83.76	79.28	82.42
ME+RS	83.43	86.70	85.23	80.78	83.89
SVM+RS	84.36	87.24	85.98	86.23	**88.68**

Graphs in Figs. 1 to 4 demonstrate the categorization performance. The blue colour represents Naive Bayes, the maroon colour represents K-Nearest Neighbor, the green colour represents Maximum Entropy, and the purple colour represents Support Vector Machine. A Maximum Entropy method with a laptop dataset achieves the most remarkable accuracy of 88.95 percent, as shown in Fig 3

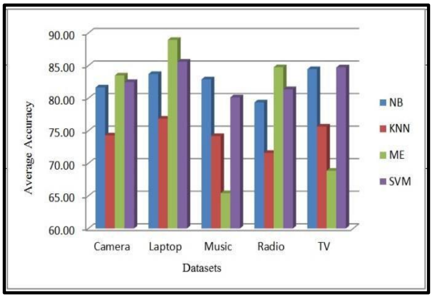

Fig. 3: The average accuracy of individual machine learning algorithms.

The most excellent accuracy is 83.45 percent, as shown in Fig. 2, which was reached by combining Maximum Entropy and Bagging classifications with a laptop dataset. The most incredible accuracy of 88.58 percent was obtained using a combination of Support Vector Machine and Boosting classifiers using a laptop dataset, as shown in Fig. 3. Finally, combining Support Vector Machine and Random Subspace classifiers with the TV dataset yielded the best accuracy of 88.68 percent, as shown in Fig. 4.

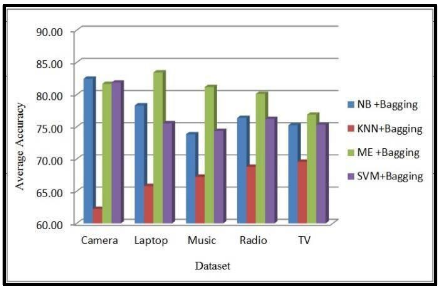

Fig. 4: The average accuracy of machine learning methods combined with bagging methods.

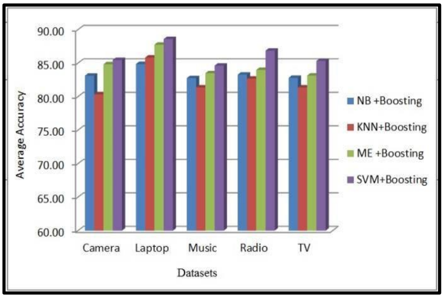

Fig 5: The average accuracy of machine learning methods combined with boosting methods.

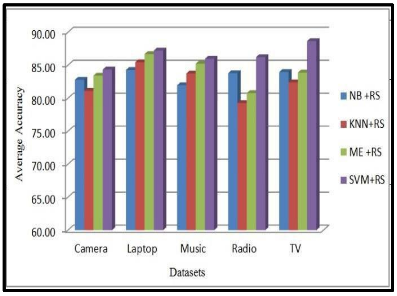

Fig. 6: The average accuracy of machine learning methods combined with random forest methods.

This experimental work conclusion is that the best performance generated by the Support Vector Machine and Maximum Entropy classifier with ensemble method.

7. Conclusion

The main goal of this paper is to improve the accuracy of machine learning classification algorithms for sentiment analysis. This research work was completed in several stages. At the first stage, the literary work on sentiment analysis by many authors was considered. The sentiment analysis table was prepared based on previous work. This table provides a detailed year-wise research work of sentiment analysis methods, contribution, and limitations or drawbacks. The preprocessing technique used to make unstructured data information to structured data information. The preprocessing steps used tokenization, stop-word-removal, stemming word, and text transformation. The feature selection methods select top features from the data set. The unigram, bigram, information gain, chi-square, and Gini index methods are used to select relevant features. The implementation results of machine learning classification algorithms have generated F-score without feature selection methods is 84.93% for Naïve Bayes, 85.29% for K-Nearest Neighbor, 86.92% for Support Vector Machine, 89.04% for Logistic Regression, and 89.97% of Random Forest on electronics product review datasets. The F-score is 85.69% for Naïve Bayes, 85.79% for K- Nearest Neighbor, 88.54 for Support Vector Machine, 90.15 for Logistic Regression, and 93.20 for Random Forest on movie review data sets. Subsequently, implementation results by machine learning. Classification algorithms combined with feature selection methods (Unigram + Bigram + IG + CS + GI). This combination generated the highest F-score with 86.28% for Naïve Bayes, 87.80% for K-Nearest Neighbor, 92.10% for Support Vector Machine, 92.84% for Logistic Regression, and 94.49% of Random Forest of electronics product review data sets. The F-score is 88.94 for Naïve Bayes, 89.92% for K- Nearest Neighbor, 92.29% is for Support Vector Machine, 93.88% are for Logistic Regression and 94.98% is for Logistic Regression on movie review data sets. Classification methods NB, KNN, ME and SVM produced maximum accuracy respectively as 84.46%, 75.65%, 88.95%, and 85.60%.Five machine learning classification methods are used in the studies mentioned above. The review datasets used Nave Bayes, k-Nearest Neighbor, Maximum Entropy, and Support Vector Machine separately and in combination. This study compared the performance of a well-trained machine learning system with ensemble techniques in classifying reviews data quickly and accurately. On the TV dataset, the Support Vector Machine with Random Subspace produced the best accuracy of 88.68. On the laptop data set, the Support Vector Machine and Boosting made the maximum accuracy of 88.58. Finally, on the laptop dataset, Maximum Entropy with Bagging produced the best accuracy of 83.45.

Reference

[1] OzasaT,WeirG,andFukuiM,“AsoftwareapplicationforassessingreadabilityintheJapaneseEFLcontext,”ThemesinScienceandTechnologyEducation,vol.3,no.1&2,2019,pp.91-118.

[2] WeirG,IshikawaS,andPoonponK,“Corporaandlanguagetechnologiesinteaching,learningandresearch,”UniversityofStrathclyde,2019.

[3] XiaR,ZongC,andLiS,“Ensembleoffeaturesetsandclassificationalgorithmsforsentimentclassification”,InformationSciences,Vol.181No.6,2011,pp.1138-1152.

[4] PrabowoR,andThelwallM,“Sentimentanalysis:A combinedapproach”,JournalofInformetrics,Vol.3,No. 2,2009,pp.143-157.

[5] TanS,andZhangJ,“AnempiricalstudyofsentimentanalysisforChinesedocuments”,ExpertSystemswithApplications,Vol.34,No.4,2018,pp.2622-2629.

[6] KhanA,BaharudinB,LeeLH,andKhanK,“Areviewofmachinelearningalgorithmsfortext-documentsclassification,”Journalof AdvancesInInformationTechnology,Vol.1,No.1, 2010,pp.4-20.

[7] MiaoQ,LiQ,andZengD,“Miningfine-grained opinions by using probabilistic models and domain knowledge”,Proceedingsofthe2010IEEE/WIC/ACMinternationalconferenceonwebintelligenceandintelligentagenttechnology–WI-IAT"10,Washington,DC,USA:IEEEComputerSociety01,2010,pp.358–365.

[8] GhaniR,andSimmonsH,“Predictingtheend-priceofonlineauctions”,InProceedingsoftheInternationalWorkshoponDataMiningandAdaptiveModellingMethodsforEconomicsandManagement,2014.

[9] AbulaishM,DojaMN,andAhmadT,“Featureandopinionminingforcustomerreviewsummarization”,InPatternRecognitionandMachineIntelligence,SpringerBerlin,Heidelberg, 2019, pp.219-224.

[10] Dave K, Lawrence S, and Pennock DM, “Mining the peanut gallery: Opinion extraction and semantic classification of product reviews,” In Proceedings of the 12th international conference on World Wide Web, ACM, 2003, pp. 519-528.

[11] Fang X, and Zhan J, “Sentiment analysis using product reviews data,” Journal of Big Data,” Vol. 2 No.1, pp. 5, 2015.

[12] Louis F, and Wojciech C, "Advances in the human side of Service Engineering", In Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics, Volume Set, AHEE Conference 19-23 July 2014.

[13] Liu B, “Sentiment analysis and opinion mining,” Morgan & Claypool Publishers, 2016.

[14] Wallin A, “Sentiment analysis of amazon reviews and perception of product features,” Doctoral dissertation, Master"s thesis, Lund University, 2019.

[15] Abbasi A, Chen H, and Salem A, “Sentiment analysis in multiple languages: Feature selection for opinion classification,” in web forums, ACM Transactions on information systems, vol. 26, No. 3, 2020, pp. 12.

[16] Tang H, Tan S, and Cheng X, “A survey on sentiment detection of reviews,” Expert system application, Vol. 36, No. 7, 2021, pp. 10760-10773

[17] Liu B, “Sentiment analysis and subjectivity,” Handbook of natural language processing, 2010 Mar, Vol. 2, No. 2019 pp. 627-66.