Proceedings of the 4th International Conference on Computing and Data Science (CONF-CDS 2022)
DOI: 10.54254/2755-2721/2/20220586

Sentimental Analysis on Product Reviews Using Support Vector Machine and Naïve Bayes

Chunduru Anilkumar[1], Sathishkumar V E[2,*], Seepana Kanchana[1] and Sasapu Bharath Kumar[1]

[1] Dept of Information Technology, GMR Institute of Technology, Rajam, Andhra Pradesh-532127
[2] Department of Industrial Engineering, Hanyang University, 222 Wangsimini-ro, Seongdong-gu, Seoul, 04763, Republic of Korea

srisathishkumarve@gmail.com

Abstract. People nowadays use internet platforms to exchange ideas, share opinions, and learn online. Huge amounts of data are being poured into social media in the form of tweets, blogs, and updates on articles and items, among other things. The data is all unorganized and unprocessed. It is necessary to arrange and examine it. It takes a long time to analyze and process information using traditional methods and is impossible to analyze each and every sentence. So, there is a need to have a better approach. It can be done through sentimental analysis which extracts the opinion of a user in a piece of text data. This sentimental analysis will predict the polarity of the sentence, whether the given sentence is positive or a negative one. The sentimental analysis can be achieved through three approaches namely lexicon based, machine learning based and hybrid approach. This sentimental analysis is a part of NLP. This project aims to perform sentimental analysis using machine learning techniques and few natural language processing techniques on a product reviews dataset.

Keywords: sentimental analysis, machine learning, nlp, naive bayes, svm.

1. Introduction

Nowadays, numerous customer reviews are provided for almost everything that is available on ecommerce websites like Amazon and Flipkart.User reviews on products may be included in reviews, with the goal of assisting other users in their purchasing decisions. There are a lot of reviews, making it tough for a customer to read them all and make a selection. It is difficult for the consumer to distinguish the product reviews if they read some of the evaluations. On the other hand, Consumers rely on user reviews for important information. They can, however, improve or degrade a product's or website's reputation dependent on their credibility. Embedding social intelligence from massive online comments is a time-consuming task for any society or person. These issues prompted the development of Sentiment Analysis, a social analysis method for automatically extracting, analyzing, and summarizing user-generated data. Machine learning is making the machines to predict on their own based on the data feed to it. The effectiveness of using machine learning approaches to solve the sentiment classification problem is studied in this work. In contrast to unsupervised learning, which does not require prior training, whereas supervised learning deals with the labelled data.Labelled data specifies the output label and attributes

DOI: 10.54254/2755-2721/2/20220586

for the data. Machine Learning model learns from the training data and use it for the new data to map to the output label. This type of machine learning has more accurate results since each of the classifiers is trained on a set of representative data. Instead, it measures how far a word is inclined toward positive and negative in order to mine the data. But this needs a human labor for making datasets. Reinforcement learning is all about making decisions sequentially. This research examines supervised machine learning algorithms to understand sentiment analysis. Sentiment analysis (SA) collects online documents such as tweets, Facebook status updates, product reviews, blogs, and other social media platforms. Customer's attitudes, opinions, and emotions can be better understood by using online documents. Sentiment analysis is a technique for detecting emotional expressions in natural language texts.

2. Literature Review

An in-depth look at sentiment analysis techniques based on recent research, followed by a look at machine learning. Because of the high dimensionality of the data, it necessitates special preprocessing and feature extraction in order to increase classification accuracy. This research also addresses issues such as excessive simplicity in identifying, in general, multiple languages posts on social media with geographic treatment [1]. The results of this study show that well-trained supervised machine learning techniques can classify SA polarities quite effectively. We would use different tools to determine unfair ratings, such as the Statistical Analysis System (SAS) or the Python machine learning toolkit (scikit-learn), and then use these techniques to evaluate our work performance [2]. It shows high performance compared to traditional based approach Sentimental analysis that completed on each object analysis and then categorized using machine learning procedures NB & SVM[3]. The authors estimated that the review intensities ranged from -9 to +9. The feature extractor in this study is Latent Semantic Analysis, while the classifier is the SVM algorithm [4]. Cleaning up the data that has been crawled and in order to get the best results, all special characters (such as ":/.,'#$*&-) are removed. Then create a csv file with the crawled content. As this model classifies the reviews based on sentimental analysis by returning the +1 for positive sentence and returns the -1 for the negative sentence [5]. Then used SVM and Naïve Bayes for the sentimental analysis which are Machine learning approaches. Naïve Bayes performance (84.02%) is better compared to SVM performance (80.2%) as Naïve Bayes used for the time saving [6]. This model demonstrates that highly effective outcomes for product aspect extraction may be achieved by combining these hypotheses. A technique based on graphs for detecting implicit characteristics in reviews [7]. For determining sentiment accuracy Naive Bayes, Maximum Entropy, SVM, CNN and Long Short Term Memory algorithms are used. Finally showed the experimental results of Machine Learning approaches and their result analysis on a dataset created based on a questionnaire.CNN and SVM shows the best results [8]. Hybrid approach is used at the sentence level. The main purpose of sentiment analysis is to establish the attitude of a given sentence towards tenses, topics, and paragraphs or document. Twitter data sets and performed hybrid approach to the data sets [9]. The detailed information on the four different levels of sentiment and the optimization of machine learning classifiers for sentiment prediction. One is Naive Bayes, which needs small dataset for training on text classification and is quite fast in learning [10].

3. Methodology

Methodology was organized in below following steps as shown in the figure.

DOI: 10.54254/2755-2721/2/20220586

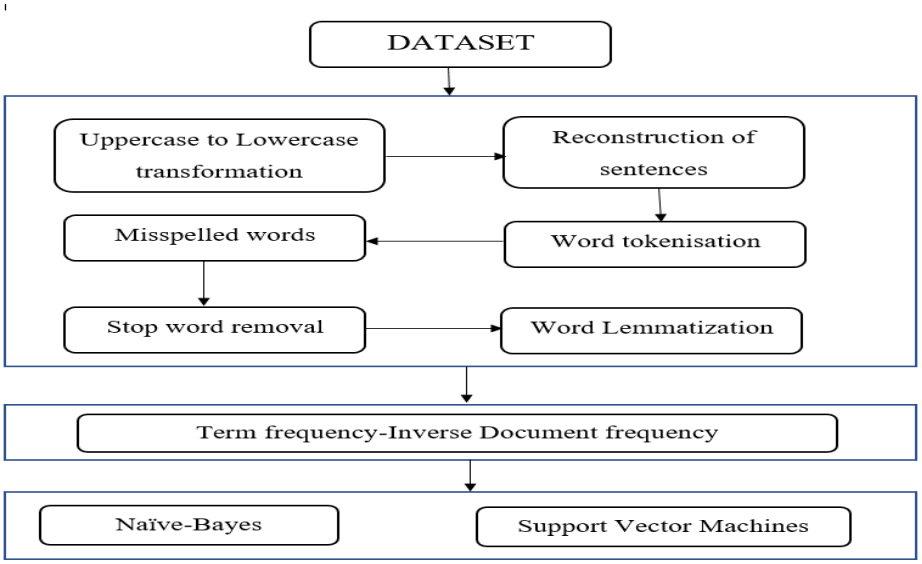

Figure 1. Flow Diagram.

3.1. Data Pre-Processing

Social media community has its own special slang language as per their convenience to post message where reviews contains many symbols, misright words, sarcastic sentences. It means dataset is unstructured, these large number of words that are not needed for determining the sentiment a summarizing the opinions. Hence, Pre-processing of the data is required in Sentimental Analysis. By cleaning and organizing the data, with the right pre-processing procedures, classification accuracy can be increased.Thepre-processing contains the following steps:

1. Transforming the text to lower case:

Eg: I am GOOD at Sports-> i am good at sports.

*We have used string lower () function.

2. Reconstructing the sentence:

Eg: i'llà i will, we've à we have, removing urls, removing symbols such as @.

*Here we have used python regular expression module're'.

3. Word Tokenization: Every piece of text is broken into set of words

Eg: {I, am, enjoying, the, taste, of, the, food, very, well}

*Here we have used word_tokenize () in nltk library.

4. Misspelled Words: In English grammar, reviewing all the words in a sentence and mapping the incorrectly spelled words to almost identical terms.

Eg: calendar->calendar

*For handling misspelled words we have used SpellChecker () module in python.

5. Removal of stop words: Stop words are words that are not used to express an emotion or feeling but are used as a connector or articles in English. We have manually written the stop words list and eliminated them from the given reviews.

Eg: and, with, of, the, a, there, they…etc.

6. Word Lemmatization: Producing the root word for the given set of words. Here the root word is the actual word in the grammar of the language.

Eg: loving, loved, lovely à love

*Here we have used wordnetlemmatizer () available in nltk library.

3.2. Feature Extraction

Extracting features from the data is called Feature Extraction.

DOI: 10.54254/2755-2721/2/20220586

Bag of words and TF-IDF are the two feature extraction techniques are used to extract the features from the data.

3.3. Bag of Words

It is a method used to extract the features or information from the text documents. It converts random text into fixed length vectors by counting how many times a word is repeating or appears. This is called vectorization. The disadvantage of BOW is we lose contextual information. Which means BOW just describes only what words are occurring in the document but not where they occur. It is simple and inexpensive to compute. The main advantage of using BOW as a feature extraction technique is its better when the contextual information is not relevant. As we can't put direct text into the machine learning model, so we first convert the text into bag of words.

3.4. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF): It is used in the search engines. It finds the meaning of sentences consisting of words and cancel outs the Bag of Words technique.

IDF- Inverse document frequency: It determines the importance of a term. It represents the significance of a word in a corpus of documents.

This paper is broken into seven sections listed below. Section I provides the introduction to this paper, Section II comprises a Literature Review (study) of previous works, and Section III represents a system design. In Section IV the methodology is described. The Results are included in Section V. Section VI contains the conclusion.

4. Methodology

Classification (Naive Bayes and SVM):

Naive Bayes: The supervised learning algorithm Nave Bayes, which uses Bayes theorem to predict the occurrence of any event. It is a probabilistic classifier which classifies based o probability of an object. It depends on Bayes theorem, so it is called as Naive Bayes theorem.

Bayes theorem:

$$P\left(\frac{B}{A}\right) = \frac{P\left(\frac{A}{B}\right) * P(B)}{P(A)} \tag{1}$$

This algorithm will find out the probability of class either it is positive or negative based on the given sentence.

$$P\left(\frac{positive}{sentence}\right) = \frac{P\left(\frac{sentence}{positive}\right) * P(positive)}{P(sentence)} \tag{2}$$

$$P\left(\frac{negative}{sentence}\right) = \frac{P\left(\frac{sentence}{negative}\right) * P(negative)}{P(sentence)} \tag{3}$$

Here probability of positive and negative sentence can be calculated as follows and denominator is same ignore it.

$$P(positive) = \frac{Total\ no\ of\ positive\ sentences}{Total\ no\ of\ sentences} \tag{4}$$

$$P(negative) = \frac{Total\ no\ of\ negative\ sentences}{Total\ no\ of\ sentences} \tag{5}$$

which leads to make the probability Zero. In order to avoid. Then we calculate the individual probability of every word in a sentence.

DOI: 10.54254/2755-2721/2/20220586

$$P\left(\frac{word}{class}\right) = \frac{No\ of\ times\ the\ word\ appear\ in\ that\ class}{Total\ no\ of\ words\ present\ in\ that\ class} \quad (6)$$

The equation becomes zero if a word from the new sentence does not exist in the class within the training set. the entire equation is nullified. In order to address this issue. Use Laplace Smoothing:

$$P\left(\frac{word}{class}\right) = \frac{Nxi, wj + \alpha}{Nwj + \alpha d}, where\ \alpha = 1 \quad (7)$$

Class for a given sentence = Max(P(positive/sentence), P(negative/sentence))

It assumes that words in a sentence are independent of each other.

4.1. SVM (Support Vector Machine)

SVM is termed as Support Vector Machines. SVM gives the best decision boundary between the vector whether the vector belongs to that particular group / category or not. So that text need to be converted into the vectors first. It means we have to encode the text into the vector. SVM draws the best line between the vectors to classify the objects. The line which is used to classify the vectors or objects is called hyper plane. This hyper plane divides the spaces into two sub spaces. As sub spaces denotes with one vector belongs to given category and another one is vectors vectors which they do not belongs to it. The parallel lines which are drawn from the hyper plane is called Marginal Planes.

The marginal planes distances will be calculated as d= |d1-d2|/||w||

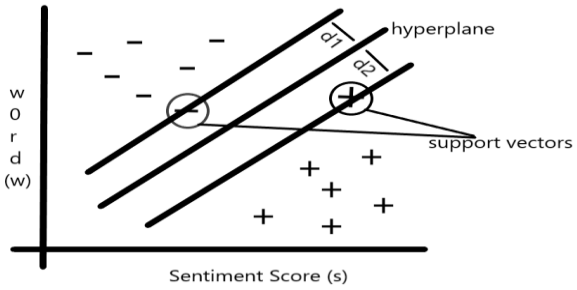

Figure 2. Sentimental Classification using SVM.

Where the distances between the hyper planes and the two marginal planes are, respectively, d1 and d2.

5. Results

The performance metrics are calculated using the Confusion Matrix. . Generally, Positive and Negative values are used to describe predicted values. True and False are used to describe actual values.

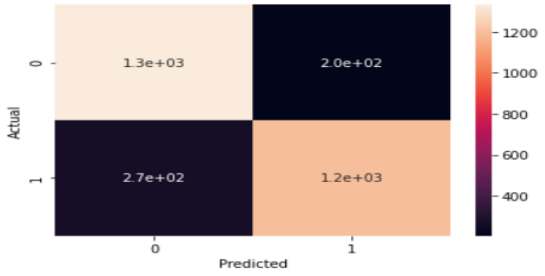

Figure 3. Confusion Matrix for the Naive Bayes.

DOI: 10.54254/2755-2721/2/20220586

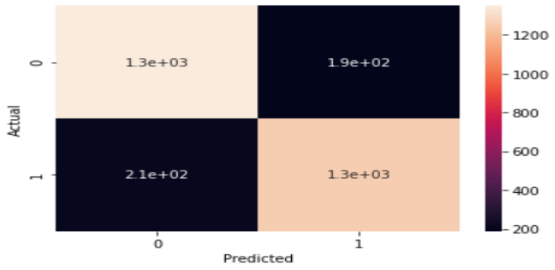

Figure 4. Confusion Matrix for the Support Vector Machine.

Classification reports:

Table 1. Naive Bayes.

	Precision	Recall	F1-Score	Support
0	0.83	0.87	0.85	1537
1	0.85	0.82	0.84	1463
Accuracy			0.84	3000
Macro Avg	0.84	0.84	0.84	3000
Weighted avg	0.84	0.84	0.84	3000

Table 2. Support Vector Machine.

	Precision	Recall	F1-Score	Support
0	0.87	0.88	0.87	1537
1	0.87	0.86	0.86	1463
Accuracy			0.87	3000
Macro Avg	0.87	0.87	0.87	3000
Weighted avg	0.87	0.87	0.87	3000

6. Conclusion

Two algorithms namely SVM and Naïve Bayes is implemented on the dataset. Performance metrics like Accuracies are calculated for both the algorithms. SVM model gave the better performance than the Naive Bayes. We have taken two datasets, where the first dataset which is having the high length in the sentence took time when compared to the other dataset whose length of the sentence is less. Naive Bayes model is considered as a time efficient as the model takes the less time for the training the model. Support Vector Machine (SVM) is a memory efficient as the only stores the support vectors data points only. Therefore, SVM uses when the user has less idea on the data.

References

[1] Singh, N. K., Tomar, D. S., &Sangaiah, A. K. (2020). Sentiment analysis: a review and comparative analysis over social media. Journal of Ambient Intelligence and Humanized Computing, 11(1), 97-117.

[2] Elmurngi, E. I., &Gherbi, A. (2018). Unfair reviews detection on amazon reviews using sentiment analysis with supervised learning techniques. J. Comput. Sci., 14(5), 714-726.

[3] Nawaz, U., Ali, A., Raza, U. A., &Shehzadi, K. (2021). A Survey: Sentimental Analysis on Product Reviews Using (MLT) Machine Learning Techniques and Approaches. International Journal, 10(2).

DOI: 10.54254/2755-2721/2/20220586

[4] Zhang, Wei, Sui-xi Kong, and Yan-chun Zhu. "Sentiment classification and computing for online reviews by a hybrid SVM and LSA based approach." Cluster Computing, Volume 22, No. 5, 2019, pp. 12619-12632

[5] Bhatt, A., Patel, A., Chheda, H., &Gawande, K. (2015). Amazon review classification and sentiment analysis. International Journal of Computer Science and Information Technologies, 6(6), 5107-5110.

[6] Chauhan, M., &Yadav, D. (2015). Sentimental analysis of product based reviews using machine learning approaches. Journal of Network Communications and Emerging Technologies (JNCET), 5(2), 19-25.

[7] Bagheri, A., Saraee, M., & De Jong, F. (2013). Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. Knowledge-Based Systems, 52, 201-213.

[8] Kumar, S., Gahalawat, M., Roy, P. P., Dogra, D. P., & Kim, B. G. (2020). Exploring impact of age and gender on sentiment analysis using machine learning. Electronics, 9(2), 374.

[9] Appel, O., Chiclana, F., Carter, J., & Fujita, H. (2016). A hybrid approach to the sentiment analysis problem at the sentence level. Knowledge-Based Systems, 108, 110-124.

[10] Singh, J., Singh, G., & Singh, R. “Optimization of sentiment analysis using machine learning classifiers”, Human-centric Computing and information Sciences, Volume 7, No 1, 32, 2017.