WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Sentimental Analysis Using Machine Learning Algorithms

Lenka Venkata Seetha Rama Lakshmi

Student, Department of Information Technology, GMR Institute of Technology
20341A1257@gmrit.edu.in

ABSTRACT:

Nowadays, people using online platforms for exchanging ideas, sharing opinions and online learning. Social media is filling up with huge data within the kind of tweets, blogs, and updates on posts and products etc. That data is unstructured and it has to be organized and analysed. Companies and lots of others are attempting to figure onit data for several days to boost their position within the market. Sentimental Analysis plays an important role during this dataclassification.SentimentalAnalysis refers to the utilization of tongue Processing.This analysis will be done using machine learning. Machine learning is training the machines to be told from their past experiences. This study aims to be told different levels of sentiments, steps involved in sentimental analysis and machine learning algorithms for sentiment classification.

Keywords: social media, sentimental analysis, machine learning.

1.INTRODUCTION

All the social media and online web sites are dumped with huge amount of data by users. That data need to be analyzed based on its sentiment. In regards to it, every single piece of text has to be classified. This process of summarizing the opinions given in the huge opinionated user generated data is called Sentimental Analysis. Sentimental Analysis is needed in many applications which is helpful for both companies and users such as movie reviews because movie ratings will be given based on their reviews and people who want to watch the film will basically read the reviews only, Restaurant reviews has to be classified because customers all decide whether to go to the restaurant or not after checking the reviews of the food, service and their maintenance, product reviews are the best among all since every user is buying based their reviews only on many online sales systems, finally twitter data which is being used by many people who are in power to share their opinions, based on these review classification we can easily predict whether people like the politician or not.

2.LITERATURE SURVEY

In paper[1] Singh, N.K., Tomar, D.S, Sangaiah, A.K, presentedcomprehensive overview of sentiment analysis technique based on recent research .Nature of data requires specific preprocessing and feature extraction that can improve classification accuracy.Author proposed Parts of Speech tagger, Bag-of-Words, Feature hashing techniques for Feature Extraction from the texts. Then used the Support Vector Machine, Naive Bayes for Sentiment classification. SVM, NB combined with POS tagger yields the better results than with other combinations as it identifies the degree of dependency between feature value and labeled class.

In paper[2]Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh, to calculate sentiment score opinion lexicons are considered. Dictionary based approach under Lexicon based approach has been used with Machine Learning technique. The authors worked on six different products reviews taken from amazon and used Support Vector Machine, Naive Bayes machine learning techniques and concluded that machine learning techniques gives the best results on product reviews. The performance metrics considered here are accuracy, precision and F1 score.

In paper[3]Kumar, S., Gahalawat , M., Roy, P. P., Dogra, D. P., & Kim, B. G.Author explored the impact of age and gender in sentiment analysis, as this can help e-commerce retailers to market their products based on specific demographics. Demographics play an important role in deciding the marketing strategies for different products. Bag-Of-Words, Word2Vec feature extraction techniques was used. Dictionary based technique was also discussed. For determining sentiment accuracy Naive Bayes, SVM and are used.Finally showed the experimental results of Machine Learning approaches and their result analysis on a dataset created based on a questionnaire.

In paper [4]Vishal A. Kharde and S. S. Sonawane.Bigram, Unigram and features are applied as an effective feature set for Sentimental Analysis. The data is properly preprocessed data and given the experimental results on tweets from Stanford University based on 10-fold method and compared with other papers which worked on same dataset. The authors used Naive Bayes classifier categorizes tweets into subjective or objective tweets. And used the Support Vector Machine for classifying subjective tweets into positive class or negative class.

In paper [5]Harshali P. Patil and Mohammad Atique Author presented 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.Others are tree based classifiers which are J48, BFTree and OneR. J48 has an ability to work with larger training datasets

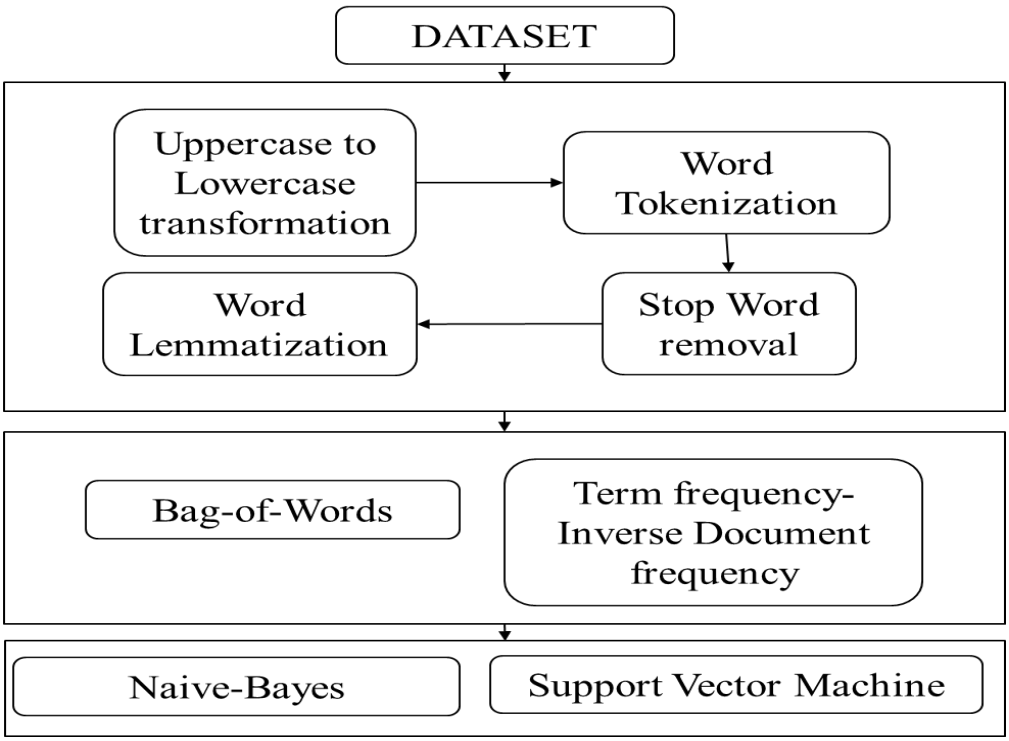
than other classifiers.Here both unigrams and bigrams are considered with J48 algorithm.

3.METHODOLOGY

The methodology we are following here is machine learning model for sentiment classification. The steps mainly involved are data preprocessing, feature extraction and classification. The dataset contains the review and its respective class i.e positive or negative statement.

The steps involved are:

1. Dataset Collection
2. Text Preprocessing
3. Feature Extraction
4. Training the Machine Learning Model
5. Predicting the class label on test set.



Methods used for classification are:

1. Naïve Bayes
2. Support Vector Machine
3. Random Forest
4. Linear Regression
5. Logistic Regression

Naive Bayes:

Basic working principle: Naive Bayes algorithm works on the probability of events. It is a classifier based on probablity. This classifier is based on Bayes Theorem

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

This algorithm will find out the probabilities of the class 'Positive , Negative' given the sentences i.e P(positive/sentence) and P(negative/sentence). As per Bayes theorem

$$P(\text{positive/sentence}) = \frac{P\left(\frac{sentence}{positive}\right) * \boldsymbol{P(positive)}}{P(sentence\ )}$$

$$P(\text{negative/sentence}) = \frac{P\left(\frac{sentence}{negative}\right) * \boldsymbol{P(negative)}}{P(sentence\ )}$$

Eg: Let the sentence be "Here the food got good taste"

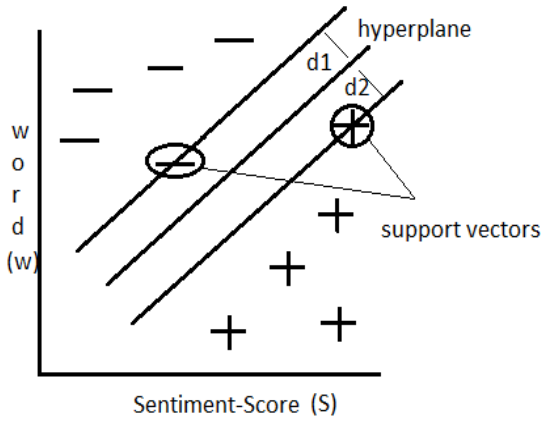
P(Here the food got good taste/class) = P(Here/class) * P(the/class) * P(food/class) * P(got/class) * P(good/class) * P(taste/class)

Support Vector Machine:

Support Vector Machine is a Binary classifier. SVM algorithm is used for both classification and regression analysis. SVM draws the graph between the word and its sentiment score. It separates both the classes by drawing a optimal hyperplane in the graph between the classes. Apart from hyperplane SVM draws two marginal planes parallel to the hyperplane so that it can easily perform the classification. Each marginal plane passes through the nearest class point. Thesepoints are known as support vectors points. The marginal planes and hyperplane are drawn such that the marginal distance d is max where d = d1 + d2. The graph is divided into two groups i.e positive and negative.

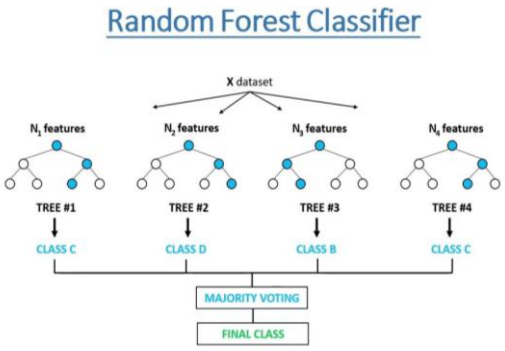
For all the positive group the score will be positive and for all negative group the score will be negative

For hyperplane, $w_{sv}s_i + b = 0$
For positive hyperplane, $w_{sv}s_i + b_1 > 0$
For negative hyperplane, $w_{sv}s_i + b_2 < 0$
W_{sv} is associated with word within the text and S_i indicate their respective sentiment class (+ve, −ve).
Advantages: with increase in memory SVM is efficient.
Disadvantages: with respect to time it is not sufficient.



Random Forest:

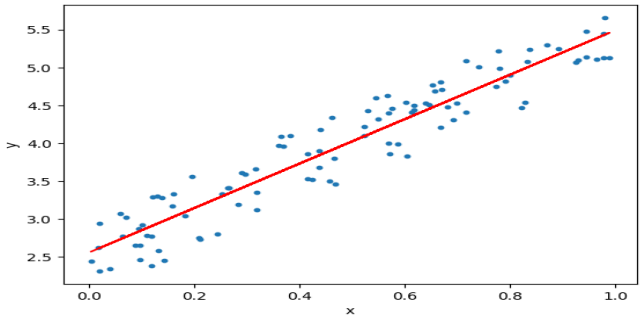
Random Forest classifier is a meta-estimator that fits a number of decision trees on various sub-samples of datasets and uses average to improve the predictive accuracy of the model and controls over-fitting. The size of sub-sample is always the same as the sample size of input sample but they can be drawn with replacement.



Linear Regression:

Linear regression is a statistical method to show relationship.It is used to predict value of Y, by which X features are given. By using machine learning, the data sets are inspected todisplay a relationship.After that along X/Y axis the relationships are placed with a straight line passes through them to predict future relationships.

Linear regression calculates how the X-axis and Y-axis values are related. Itdetermines where words and features fall on a scale from “really positive” to “really negative” and between in everywhere.



Logistic Regression:

Logistic regression is an algorithm used to predict a bipartite outcome: either something happens, or does not happen. This can be displayed as Yes/No, Pass/Fail, Alive/Dead, etc. Independent variables are examined to discover the binary outcome with the outcomes falldown into one of two categories.The dependent must be categorical whereas independent variable can be either categorical or numeric . Written like this:

P(Y=1|X) or P(Y=0|X)

It used to calculate the probability of variable *Y*which is dependent, given the probability of independent variable *X*.

4. Results And Discussions

Dataset Description:

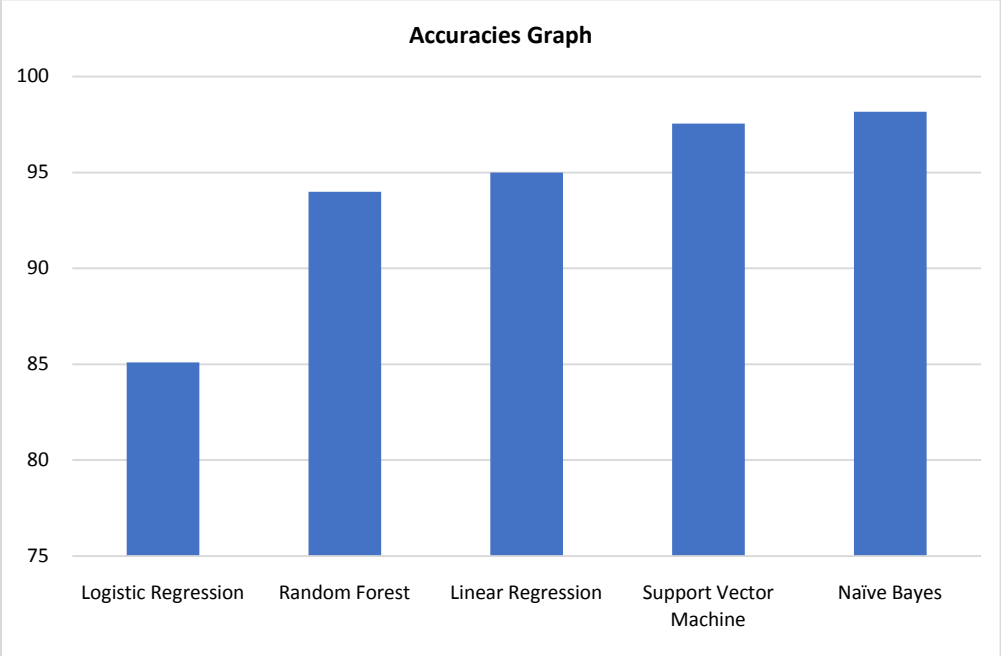
Dataset mainly consists of review id, sentence(user review), corresponding class label.

Review Id	Sentence	Class Label
01	It was such a lovely phone.	+ve
02	Battery backup was not last long.	-ve
03	Overall performance of the phone is good	+ve

Common use cases where sentimental analysis is required

- Movie Reviews
- Product Reviews
- Tweets from Twitter
- Restaurant Reviews
- You tube data

S.N0	Method	Accuracy
1.	Logistic Regression	85.1%
2.	Random Forest	94%
3.	Linear Regression	95%
4.	Support Vector Machine	97.54%
5.	Naïve Bayes	98.17%



5.Observations:

1.The various methodologies applied here are Logistic Regression , Random Forest, Linear Regression, SVM, Naïve Bayes.
2.Among all Naïve Bayes and SVM gives the more accuracies.
3.Naive Bayes model taking less time for training comparing with other classifiers.
4.With increase in size of dataset Support Vector Machine outperforms Naïve Bayes model.
5.With regards to time, Naïve Bayes is taking less time for training the model.
6.With regards to memory, Support Vector Machine is efficient.

6.Conclusion

It is noted that in present society product sales, restaurant growth, movie success keeping their performance aside all are indirectly depending on their reviews only in one way. It enlightens the need of the opinion mining or sentimental analysis for companies and users. In this paper, various approaches are mentioned and supervised algorithms in machine learning classification for the sentiments are also discussed. Machine learning approach reduces the human effort to the great extent.

In future, deep learning algorithms also need to be taken into consideration. And sarcastic sentences, slangs, symbols, misspelled words, idioms, emojis, Annotated training data, Multiple-language sentiment processing are also to be considered which can change the meaning and polarity of the entire sentence.

REFERENCES

1. Singh, N. K., Tomar, D. S., &Sangaiah, A. K., “Sentiment analysis: a review and comparative analysis over social media”, Journal of Ambient Intelligence and Humanized Computing, Vol. 11, No.1, 2020, pp. 97-117.
2. Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques", In Cognitive Informatics and Soft Computing, pp. 639-647. Springer, Singapore, 2019.
3. AnvarShathik, J.&Krishna Prasad. A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques. International Journal of Applied Engineering and Management Letters (IJAEML), 4(2), 41-77,2020.
4. Kumar, S., Gahalawat , M., Roy, P. P., Dogra, D. P., & Kim, B. G. “ Exploring Impact of Age and Gender on Sentiment Analysis Using Machine Learning”,Journal of Electronics, Volume 9, No 2, 374, 2020.
5. Harshali P. Patil and Mohammad Atique, "Sentiment Analysis for Social Media: A Survey", IEEE, 2019.
6. Vishal A. Kharde and S. S. Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", IEEE,2019.
7. Krishna, A., Akhilesh, V., Aich, A., & Hegde, C. “Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques”, In Emerging Research in Electronics, Computer Science and Technology, Springer, Singapore, 2019, pp. 687-696.
8. 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.
9.Le, B., & Nguyen, H. “Twitter sentiment analysis using machine learning techniques”, In Advanced Computational Methods for Knowledge

Engineering, 2015, pp. 279-289. Springer, Cham.
10.Arwa S. M. AlQahtani, "Product Sentiment Analysis for Amazon Reviews" international Journal of Computer Science & Information Technology (IJCSIT), vol 13, no 3, June 2021