



International Journal of Multidisciplinary and Scientific Emerging Research (IJMSERH)

Volume 14, Issue 3, July-September 2026

Impact Factor: 9.798



Sentiment Analysis using Naive Bayes, Multinomial NB, Bernoulli NB and Logistic Regression

Dr. Pradeep KGM, Dr. Dasari Rammurthy, Dr. Mathesh G

Associate Professor & HOD. Department of Electrical and Electronics Engineering, Rajeev Institute of Technology,
Hassan, Karnataka, India

Professor & Dean (Academics), Department of Electronics and Communication Engineering, Rajeev Institute of
Technology, Hassan, Karnataka, India

Assistant Professor, Department of Electrical and Electronics Engineering, Rajeev Institute of Technology, Hassan,
Karnataka, India

ABSTRACT: Of late sentiment analysis has become an inevitable tool to know the people's opinion on Government Policies, E marketing, elections, stock markets, staff satisfaction in offices and movies review. In this work, we have conducted sentiment analysis on recently released pan Indian Hindi Movie "Pathan". Raw data Collected from Twitter (10000 tweets) and data set is prepared in order to perform sentiment analysis. In this paper sentiment is estimated using different Machine learning algorithms.i.e. Naïve Bayes, Multinomial NB, Bernoulli NB and Logistic regression. The accuracy of all algorithms is above 80%, but Logistic regression has achieved 3% more accuracy compared to other algorithms.

KEYWORDS: Sentiment Analysis, Machine Learning, Naïve Bayes, Multinomial NB, Bernoulli NB, and Logistic Regression.

I. INTRODUCTION

Recent break through's in Machine Learning and Deep Learning have enabled researchers to perform Natural Language Processing (NLP) tasks such as sentiment analysis in a more simple and effective manner. Generally, in sentiment analysis Tagging play a significant role, which involves (1) Finding the objectivity of text, (2) finding the polarity of a subjective text and (iii) Finding the strength of the subjective text [1]. In sentiment Analysis, Lexicon based and Machine learning based approaches are widely used. Each Approach has its own merits and demerits. Lexicon based approach implies the use of tokens, parts of speech (POS). In this approach polarity of each word is computed locally, then used rule-based function to calculate the overall sentiment [2-5]. Lexicon based approach is further classified into Dictionary based and Corpus based approach. In dictionary-based approach each word in the dictionary consists of a pre-defined sentiment polarity strength. In corpus-based approach sentiment can be determined by finding the co-occurrence patterns of words.

In machine learning approach labeled data is used to train classifiers such as Naïve Bayes ,Support Vector Machine (SVM),Maximum Entropy , and Long Short Memory (LSTM)(deep learning recurrent network).These Machine learning approaches effectively works in sentiment analysis and Natural Language Processing (NLP) tasks [6-7].Machine learning and deep learning approaches requires preprocessed text to prepare feature vector. Supervised learning and unsupervised learning are two different approaches in machine learning.

In Unsupervised learning, no trained data set is given, and there is no specific target output with associated input. In this approach algorithm learns through its experiences. Clustering algorithms such as K-means, Hierarchical, Gaussian mixer models, Hiden Markov models, Euclidean Distance and Self organizing maps will plat significant role in Unsupervised learning.

In Supervised learning algorithms trained data set and test data set are used. Algorithm learns from trained data set to predict future output. To evaluate performance, classifier test is required. Decision tree classifier, Rule based classifier,

Probabilistic classifier (Naïve Bayes Classifier, Maximum Entropy based Classifier), Linear classifier (Support Vector Machine (SVM), Neural Networks), Lexicon Based Approach (Dictionary based approach) play dominant role in supervised learning approach.

In this work we have effectively applied Naïve Bayes, Multinomial NB, Bernoulli NB and Logistic regression. Naive Bayes works on the principle of Bayes theorem. In the approach probability is determined, based on text whether text consists positive or negative polarity. Multinomial NB, Bernoulli NB also works on the basis of Bayes Theorem. Multinomial NB predicts tag of text such as news article and piece of Email etc. Bernoulli NB works on the idea of Bernoulli distribution and accepts only binary values (0 and 1). As a part of this work, sentiment of every sentence is expressed in terms of binary values. Logistic regression tells the sentiment of every sentence (positive or negative) based on applied data set.

II. RELATED WORKS

Researchers have already done some notable work in Sentiment Analysis. In Paper [8] authors presented sentiment analysis using binary classification by assigning positive and negative polarity to reviews. In [9] authors applied sentiment analysis on product consumer summarization. In this work Latent Sentiment Analysis (LSA) based filtering mechanism is used to classify opinions, a few of those were used to know the review summarization.

In paper [10] authors implemented rule-based mechanism on Chinese document to perform sentiment analysis using baseline and Support Vector Machine (SVM), which then determines overall polarity of the document with the help of and dependency some specific words with respect to word polarity as defined in sentiment word dictionary.

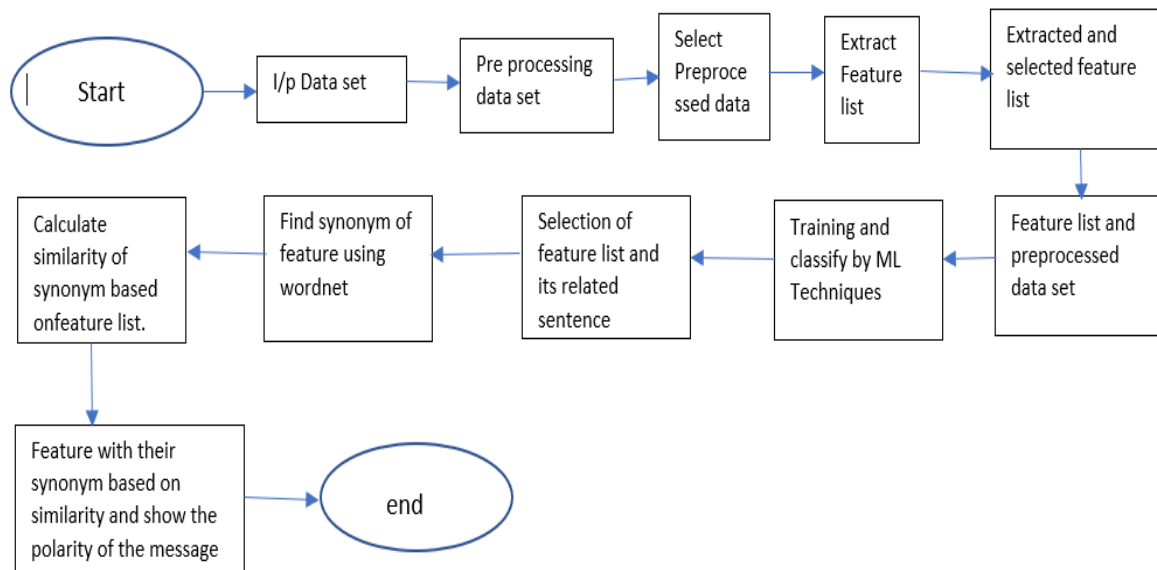
In [11] authors implemented term pattern system to extract contextual and syntactic information based on Parts of Speech (POS) and dependency relationship rules. In [11] authors implemented Volumetric, supervised and unsupervised algorithms to predict election results of four nations based on twitter messages. In this approach authors collected data from Twitter and other social media sites. In [12] authors performed sentiment analysis to predict election result of three political parties in UK in 2010. In this approach authors collected samples from twitter, and applied SVM and Naïve Bayes (NB) classifiers to classify the samples. Authors observed that NB classifier outperformed SVM without implementing any training phase and achieved a classification accuracy of 86%

III. PROPOSED WORK

In this approach raw data is collected from twitter. Labeled datasets analysis is done using unigram feature extraction technique. Preprocessor is applied on raw data for better understanding. Various machine learning algorithms were used to train the data sets with the help of feature vectors. Semantic analysis consists of huge volumes of synonyms and similarity, which helps in predicting contents emotion.

The main objective of this technique is to collect live tweets from twitter server, with the help of Python scripting language, which further perform sentiment analysis on featured data sets. Python is installed on Windows Operating system, thereafter BeautifulSoup is installed from ubuntu terminal command prompt environment. BeautifulSoup supports HTML parser and used to collect live twitter messages from twitter server.

In this work We have used SK learn to implement various machine learning algorithms. The messages are vectorized and then trained data set is prepared on 80% of the twitter messages and 20% of the twitter messages are used for testing. Figure 1 depicts flow diagram of the method to be used, apart from that the pseudocode is also presented.



The pseudo code for proposed work is given below

Input: labelled file

Output: Similarity between terms, positive and negative emotion.

Step 1: Preprocessing. Remove URL, Special Characters, convert to lower case and Special Symbols)

Step 2: Prepare feature vector

For a in words

Replace two or more words strip

If (a in stop words) Continue

Else:

Append the file return feature vector

Step 3: Extract features from feature vector list: for words in feature list

Features = words in tweets_words

Return features

Step 4: Combine preprocessing dataset and feature vector list

Pre – processed file=path name of the file stop words=file path name

Feature vector list=file path of feature vector list

Step 5: training the step 4 (Apply classifier classes)

Step 6: Find synonym and similarity of the feature vector.

IV. RESULTS AND DISCUSSION

Sentiment analysis is performed using Naïve Bayes, Multinomial NB, Bernoulli NB and logistic regression algorithms. All the algorithms showing considerable performance, but Logistic Regression has better performance compared to other companions with an accuracy rate of 88.41%. In this work We observed that Naïve Bayes and multinomial NB are showing similar performance with an accuracy rate of 85.65%. But Multinomial NB has achieved only 85.12% of accuracy rate. All the algorithms have achieved more than 80% of the accuracy. From this research we understand that all the algorithms are showing remarkable impact on Sentiment Analysis. The performance of all the algorithms is presented below.

4.1 Naive Bayes:

Accuracy: 0.8565

Confusion Matrix:

[[4342 619]
[816 4223]]

Classification Report:

	Precision	Recall	F1-score	Support
Negative	0.84	0.88	0.86	4961
Positive	0.87	0.84	0.85	5039
Accuracy	NA	NA	0.86	10000
Macro Avg	0.86	0.86	0.86	10000
Weighted Avg	0.86	0.86	0.86	10000

4.2 Multinomial NB

Accuracy: 0.8565

Confusion matrix:

[[4342 619]
[816 4223]]

Classification Report:

	Precision	Recall	F1-score	Support
Negative	0.84	0.88	0.86	4961
Positive	0.87	0.84	0.85	5039
Accuracy	NA	NA	0.86	10000
Macro Avg	0.86	0.86	0.86	10000
Weighted Avg	0.86	0.86	0.86	10000

4.3 Bernoulli NB:

Accuracy: 0.8512

Confusion Matrix:

[[4389 572]
[916 4123]]

Classification report:

	Precision	Recall	F1-score	Support
Negative	0.83	0.88	0.86	4961
Positive	0.88	0.82	0.85	5039
Accuracy	NA	NA	0.85	10000
Macro Avg	0.85	0.85	0.85	10000
Weighted Avg	0.85	0.85	0.85	10000

4.4 Logistic Regression:

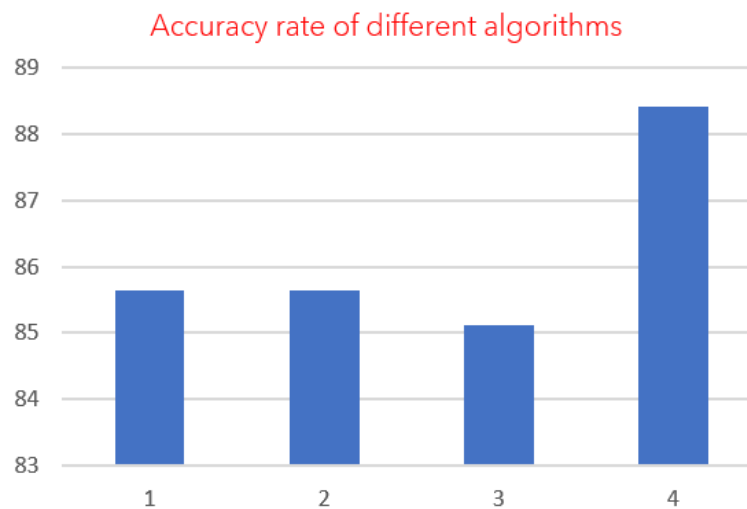
Accuracy: 0.8841

Confusion Matrix:

[[4357 604]
[555 4484]]

Classification report:

	Precision	Recall	F1-score	Support
Negative	0.89	0.88	0.88	4961
Positive	0.88	0.89	0.89	5039
Accuracy	NA	NA	0.88	10000
Macro Avg	0.88	0.88	0.88	10000
Weighted Avg	0.88	0.88	0.88	10000

**Figure.2. Accuracy rate of different algorithms****V. CONCLUSION**

We applied different Machine learning algorithms on Pathan (Hindi) movie reviews to perform sentiment analysis, raw data is collected from twitter server. The main objective of this study is to evaluate sentiment based on pre labelled twitter dataset. The logistic regression out performed all other three algorithms with an accuracy of 88.41%. Our work is carried on 50,000 review reports, out of them 80% of the reviews are used for training and remaining 20% of the review reports are used for testing.

REFERENCES

- [1] Esuli, F. Sebastiani, and V. G. Moruzzi, "A publicly available lexical Resource.pdf," in Proceedings of the 5th Conference on Language Resources and Evaluation, pp.417–422, Genoa, Italy, May 2006.
- [2] J. Ding, H. Sun, X. Wang, and X. Liu, "Entity-level sentiment analysis of issue comments," in Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering, pp. 7–13, Gothenburg, Sweden, June 2018.
- [3] S. Baccianella, A. Esuli, and F. Sebastiani, "SENTIWORDNET 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, Valletta, Malta, May 2010.
- [4] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," Proceedings of Workshop at ICLR, vol. 2013, 2013.
- [5] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," Journal of the American Society for Information Science and Technology, vol. 63, no. 1, pp. 163–173, 2012, <https://doi.org/10.1002/asi.21662>.
- [6] Singh, P. K., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A, an approach towards feature specific opinion mining and sentimental analysis across e-commerce websites, IEEE International Conference on the next Generation Information Technology, 2014, pp.329-335.

- [7] Molla, A., Biadgie, Y., & Sohn, K. A. Network based visualization of opinion mining and sentiment analysis on twitter, IEEE International conference on It Convergence and Security, 2014, pp.1 -4.
- [8] Usha, M. S., & Indra Devi, M, Analysis of sentiments using unsupervised learning techniques, IEEE International Conference on Information Communication and Embedded System, 2013, pp.241-245.
- [9] Ameer, H., & Jamoussi, S, Dynamic construction of dictionaries for sentiment classification, IEEE 13th International Conference on Data Mining Workshops, 2013, pp.896-903.
- [10] Kherwa, P., Sachdeva, A., Mahajan, D., Pande, N., & Singh, P. K, An approach towards comprehensive sentimental data analysis and opinion mining, IEEE International conference on Advance Computing Conference, 2014, pp.606-61.
- [11] S. Ahmed, K. Jaidka, and M. M. Skoric, "Tweets and votes: a four-country comparison of volumetric and sentiment analysis approaches cl-scisumm: the computational linguistics scientific summarization task view project a literature review framework for multi-document summarization of research papers view



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



International Journal of Multidisciplinary and Scientific Emerging Research (IJMSEH)

Impact Factor: 9.274