# Sentiment Analysis for Polarity Detection

**Krishna Kale[1], Prof. Pramila M. Chawan[2]**

*[1]M.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India*
*[2]Associate Professor, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India*

---------------------------------------------------------------------***---------------------------------------------------------------------

**Abstract -** *Sentiment analysis is about classifying and identifying attributes of expression. There are various platforms - twitter, facebook, Imdb where people express their opinion. On this platform it is important to recognise the opinion of people. Using machine learning techniques we can segregate people's opinion as positive, negative, neutral. We can also use this to segregate users reviews on products. This will help companies to deploy efficient solutions.*

***Key Word:*** **Machine learning, Logistic Regression**

## 1. INTRODUCTION

There are many online platforms on which user's express their views. There are social platforms, movie reviews platforms, product reviews platforms. Users use their freedom of speech to express their reviews, opinion. However, they don't consider it's implication on the society. This online platforms are great medium to express their views, opinions, criticism, contentment. However same can be used to create negative impact on the society. It can be also used to spread hatred, false political propaganda, defamation, negative image of product or person. Hence it is important to segregate user's opinions in broadly three categories positive, negative or neutral. This is where sentiment analysis comes into picture. Using appropriate machine learning techniques we can identify and classify users opinion. Users reviews, opinions cannot be used directly to perform sentiment analysis. We need to use crawler, data scraper to gather users data. Perform preprocessing on the collected data and convert into a format suitable for building model using which we can identify and classify users sentiment. This model can be used on social media platforms, IMDB movie reviews and product review platforms to capture sentiment of users. This will help to identify people who are trying to spread false narrative, hatred, racism. We can take proactive measures to curb such people from spreading false narratives. It will help companies to sell their product better by analysing user's sentiment. This will help companies to deploy solutions and products based on users preferences. This will help normal users to select a particular a product based on its positive or negative reviews. So, sentiment analysis helps in broader perspective.

## 2. LITERATURE REVIEW

### 2.1 Machine learning techniques

Machine learning techniques are used to categorized text into positive, negative or neutral. In this we need two datasets namely training and testing datasets. Training dataset is needed for learning documents and testing dataset is needed for evaluation.
Two types of algorithms are there such as supervised algorithm –SVM, Naïve bayes, KNN, maximum entropy and unsupervised algorithm – Neural networks.

In Naïve bayes we calculate probabilities of categories given in a test dataset by calculating combine probabilities of words in those categories. Naïve bayes algorithm works fast at decision making. It does not require huge learning dataset before learning begins.

In SVM support vector machine we do mapping of input set into high dimensional feature space. It is a model based on statistics. It is based on minimization of structural risk. In this we compute hyper plane to segregate data set. It has high scalability and learn larger patterns because complexity does not depend on dimensionality of feature space. It has the capability to upgrade training patterns.

In K-nearest neighbour, category labels are attached to training datasets. In this method an element is classified based on its k-nearest neighbours. In this we use graph algorithms. We compute Euclidean or Manhattan distance.

In Maximum Entropy we convert labelled feature datasets to vectors using encoding. This vector is used to compute weights to each feature set which is aggregated to compute most likely label for a given feature data set. It is used to recognise parallel phrases between pairs of languages with small training data set.

In Neural Network, it comprises of neurons where the neuron is the basic input. In this weights are associated with neuron to calculate function of its input. NN works faster when training dataset contains relations.

### 2.2 Decision tree learning

This is method is based on the concepts of tree data structure. In this we calculate root to child path to compute desired value.

It is hierarchical structure in which internal node represent test attribute, branch represent outcome, leaf node represents children node. There are various decision tree algorithms ID3, C4.5 and CART.

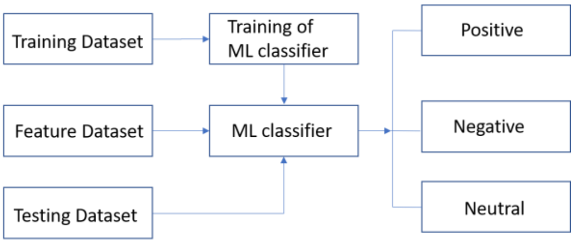
## 2.3 Information theory and coding

In this mutual information, Residual Inverse Document Frequency (RIDF), TF-IDF are used for sentiment analysis and its classification.

## 2.4 Semantic orientation approach

This approach is based on unsupervised learning. It calculates inclination of word to positive or negative clustering.

## 3. PROPOSED SYSTEM



System architecture is shown in fig. First we collect raw data and perform pre-processing on given data. We apply data inspection and data cleaning on given data. Pre-processing on given data helps to remove redundant data. There are two datasets training data set and testing dataset. In learning dataset we are given given sentences which are already classified as positive, negative or neutral. This training dataset will be feed to Logistic regression model which takes pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained logistic regression model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. This will help to identify and classify attributes holistically.

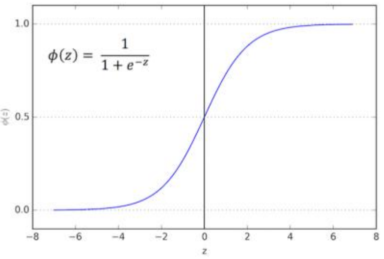
## 4. Implementation

Before feeding the data to model we perform data cleaning and preprocessing. In this we perform user handle removal , converting all the tweets into lowercase , removing all the words of having size less than three , removing all the stop words.

## 4.1 Algorithm

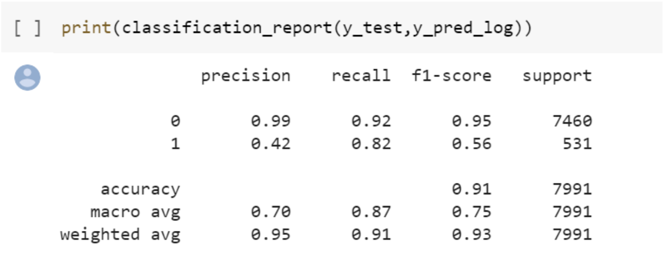
```
predicted = 1 / (1 + e^-x)
```

The logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input belonging to the default class (class 0). If the probability is > 0.5 we can take the output as a prediction for the default class (class 0), otherwise, the prediction is for the other class (class 1).



The preprocess data is feed to logistic regression model which predicts the desired label based on the predicted probability. If the predicted probability is greater than 0.5, then the sentiment is classified as positive . If the predicted probability is less than 0.5 then the sentiment is classified as negative.

## 5. Results



```
[ ]  print(classification_report(y_test,y_pred_log))

                   precision    recall  f1-score   support

               0       0.99      0.92      0.95      7460
               1       0.42      0.82      0.56       531

        accuracy                           0.91      7991
       macro avg       0.70      0.87      0.75      7991
    weighted avg       0.95      0.91      0.93      7991
```

## 6. CONCLUSION

In this Papers we learn about various machine learning algorithms. Although each algorithm is good in some aspect we can use new technique logistic regression to improve overall efficiency by identifying and classifying attributes holistically.

## REFERENCES

[1]A Nisha Jebaseeli, E.Kirubakaran, PhD., "A Survey on Sentiment Analysis of (Product) Reviews", International Journal of Computer Applications (0975 – 888) Volume 47– No.11

[2] Jalaj S. Modha, Prof & Head Gayatri S. Pandi Sandip J. Modha, "Automatic Sentiment Analysis for Unstructured Data", International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 12, ISSN: 2277 128X,.

[3] Raisa Varghese1, Jayasree M2, "A SURVEY ON SENTIMENT ANALYSIS AND OPINION MINING",

IJRET:International Journal of Research in Engineering and Technology ISSN: 2319-1163 | ISSN: 2321-7308.

[4] Arti Buche, Dr. M. B. Chandak, Akshay Zadgaonkar, "OPINION MINING AND ANALYSIS: A SURVEY", International Journal on Natural Language Computing (IJNLC) Vol. 2, No.3.

[5] Zhongwu Zhai, Bing Liu, Hua Xu and Hua Xu, "Clustering Product Features for Opinion Mining", WSDM'11

[6]Siddhi Patni, Avinash Wadhe, "Review Paper on Sentiment Analysis is – Big Challenge", International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 2, ISSN: 2321-7782 (Online), February 2014

[7] G.Vinodhini, RM.Chandrasekaran, "Sentiment Analysis and Opinion Mining: A Survey", International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 6, ISSN: 2277 128X

[8] Anderson, P., "What is Web 2.0? Ideas, technologies and implications for education", Technical report, JISC.

[9] Mishne G. and Glance N., "Predicting movie sales from blogger sentiment", In AAAI Symposium on Computational Approaches to Analyzing Weblogs (AAAI-CAAW), 2006: 155–158.

[10]Maria Tchalakova, Dale Gerdemann, Detmar Meurers, "Automatic Sentiment Classification Of Product Reviwes Using Maximal Phrases Based Analysis", Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, ACL-HLT 2011, pages 111-117, Portland, Oregon, USA 2011 Association for Computational Linguistics.

[11]Jiawen Liu, Mantosh Kumar Sarkar and GoutamChakraborty, "Feature-based Sentiment Analysis on Android App Reviews Using SAS® Text Miner and SAS® Sentiment Analysis Studio", SAS Global Forum.

[12 ]Bing Liu, "Sentiment Analysis and Opinion Mining", Morgan and Claypool Publishers, p.18-19, 27-28, 44-45, 47, 90-101.

[13] Nitin Indurkhya, Fred J. Damerau, "Handbook of Natural Language Processing", Second Edition, CRC Press.

[14] Ronen Feldman, "Techniques and Application of Sentiment Analysis", Communication of ACM, vol. 56.No.4.

[15] Ahmad Ashari, Iman Paryudi, A Min Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11.

[16] Ajayi Adebowale, Idowu S.A, Anyaehie Amarachi A., "Comparative Study of Selected Data Mining Algorithms Used For Intrusion Detection", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-3, Issue-3Sentiment Classification using Machine Learning Techniques", Proceedings of EMNLP pp. 79-86.

## BIOGRAPHIES

**Krishna Kale,** M.Tech Student, Dept of Computer Engineering and IT, VJTI College, Mumbai, Maharashtra, India.

**Prof. Pramila M. Chawan**, is working as an Associate Professor in the Computer Engineering Department of VJTI, Mumbai. She has done her B.E.(Computer Engg.) and M.E (Computer Engineering) from VJTI COE, Mumbai University. She has 27 years of teaching experience and has guided 75+ M. Tech. projects and 100+ B. Tech. projects. She has published 99 papers in the International Journals, 21 papers in the National/ International conferences/ symposiums. She has worked as an Organizing Committee member for 13 International Conferences, one National Conference and 4 AICTE workshops. She has worked as NBA coordinator of Computer Engineering Department of VJTI for 5 years. She had written proposal for VJTI under TEQIP-I in June 2004 for creating Central Computing Facility at VJTI. Rs. Eight Crore (Rs. 8,00,00,000/-) were sanctioned by the World Bank on this proposal.