Supervised Sentiment Analysis Algorithms

Marianela Denegri Coria; Juan Carlos Morales Arevalo; Jorge Luis Hilario-Rivas; Jorge Rubén Hilario-Cárdenas and Justina Isabel Prado-Juscamaita

[1]Universidad de La Frontera. Departamento de Psicología-CEPEC. Chile.
marianela.denegri@ufrontera.cl
https://orcid.org/0000-0001-7954-3697
[2]UniversidadTecnológica del Perú. Lima, Perú.
C21046@utp.edu.pe
https://orcid.org/0000-0002-1834-6451
[3]Universidad Nacional de Ucayali, Perú.
Dr@jorgeluishilario.com
https://orcid.org/0000-0003-1283-5630
[4]Universidad Nacional Hermilio Valdizán, Huánuco – Perú.
jrhilario@unheval.edu.pe
https://orcid.org/0000-0001-6627-6489
[5]Universidad Nacional Hermilio Valdizán, Huánuco - Perú
jprado@unheval.edu.pe
https://orcid.org/0000-0002-6558-4233

Abstract
Sentiment analysis is used to analyse customer sentiment by the process of using natural language processing, text analysis, and statistics. A good customer survey understands the sentiment of their customers—what, how and why they're saying it. Sentiment dataset can be found mainly in tweets, comments and reviews. Sentiment Analysis understands emotions with the help of software, and it is playing an inevitable role in today's workplaces. Sentiment analysis for opinion mining has become an emerging area where more research and innovations are done. Sentiment or opinion analysis based on a domain is done using several algorithms. Machine learning is a concept among this area. In this, the main focus is on the supervised sentiment analysis or opinion mining algorithms. Supervised learning is a division coming under machine learning. Different methods of supervised learning and sentiment analysis algorithms are considered and their mode of functioning is studied. Main focus of this paper is on the recent trends of research and studies for sentiment classification, taking into consideration the accuracy of different algorithmic techniques that can be implemented for accurate prediction in sentiment Analysis.
Keywords Sentiment analysis, opinion mining, machine learning, supervised learning.

I. Introduction

Sentiments can be split into true positive, truenegative, false positive and false negative. Opinions are expressed by people in different ways. Irony, sarcasm and implied meanings can mislead sentiment analysis. A better way of prediction of sentiments is only through context: analysis, that is knowing the collective meaning of a group of words.

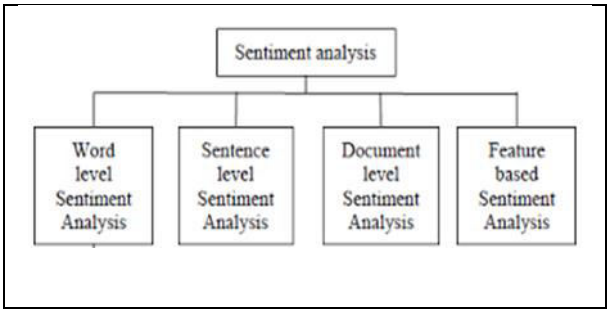

Figure 1: Levels of classification in Sentiment Analysis

Sentiment analysis, also referred to as opinion mining, is the package of classifying different domains based on the polarity. [16] Analysis and prediction on polarity can be done in word level, sentencelevel, document level or feature level as shown in figure 1. Polarity may be positive, negative or neutral. Domains may be of different kinds including reviews, tweets, opinion on social media, blogs etc. Large companies in different sectors like business, insurance, politics, news are investing huge amount of money and time for getting a conclusive opinion on their investment based on reviews and comments. [12] [20]

The process of analysing Sentiments is done using algorithms. Text analysis and natural language processing are the main techniques that are used to classify words as either positive, negative, or neutral. Generally, all sentiment analysis algorithms use natural language processing (NLP) to decide the polarity of input data. Word spotting is the common technique implemented by sentiment analysis algorithms. Sample dataset is scanned for positive and negative words like 'good', 'bad', 'interested', and 'not satisfied'. Different algorithms have different methods to identify the polarity. The algorithm that is applied should have a correlation with the domain and type of sentiment analysis that we are intending. [6]

This article is concentrating on the SA related fields that use SA techniques for many real world applications. Frequently used algorithms are Naïve Bayes and SVM, which are expected to give more accurate predictions. [1] [19] The effect of entropy based category coverage difference enhances the selection feature techniques. [2] The various ML based algorithms are taken into account and studied in detail with the Amazon reviews. [3] Impact of SA in social media is taken into study. A study was done on lexicon based models which is found to be more efficient in social media. [3][4] There are many challenges faced by sentiment analysis techniques in identifying correct polarity.NLP and text analysis methods are used to find out the correct subjective information. [22]

The analysts use different methods for getting this conclusive output based on polarity classification. A broad classification of sentiment analysis techniques can be done as in Figure 2.

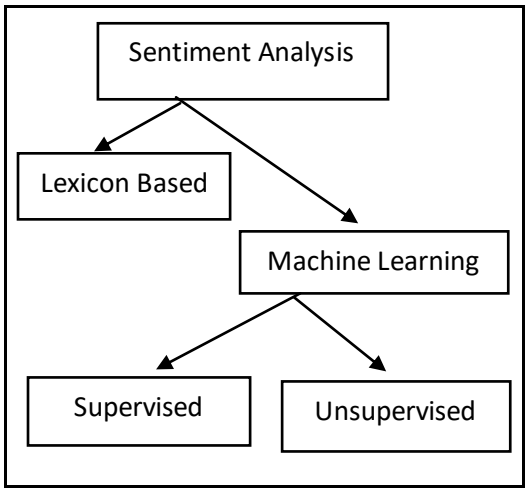

Figure 2: Categorization of techniques used for opinion mining/Sentiment Analysis

Types of sentiment analysis

Sentiment analysis can be classified into different categories based on the focus areas. [21]

1. Fine grained sentiment analysis

In this model, classification is based on the polarity. The polarity categories include positive, negative and neutral. The above classification can be further divided into very positive and very negative.

2. Emotion detection

This model detects emotions or feelings. It can be text as well as emoji's/ expressing feelings like sad, happy, anger, frustration etc.

3. Aspect based

Aspect based models focus on the polarity of a particular feature or specific area. For example, "the screen of the mobile is not so good". The polarity should be measured based on the aspect. In the above example, polarity classification is done on the aspect 'screen'.

4. Multilingual classification

For analysing multilingual classification models, algorithms should detect the language automatically and then should identify the polarity.

The rest of the article is categorized in the following way. Section 2 gives an overview of machine learning concepts. Section 3 describes the supervised learning methods. Section 4 lists a comparative study on selected supervisory modelssection 5 contains results and discussions and section 6 conclusion.

II. Lexicon Based-Methods

In **lexicon-based models** a text message is represented as a BOW (Bag of Words). Lexicon models work based on clustering and scoring algorithms. They rely on external lexical sources that rely on polarity of each term. Here the polarity of a textual content depends on the polarity of individual phrases it comprises. In general cases conjunctions, adverbs and punctuations are used for decomposition.

Dictionary-based sentiment analysis is a computational approach that identifies the feeling of a text. Generally, sentiment analysis has a binary classification, but we can extract more feelings from a sentence too, like fear, sadness, anger etc. The dictionary model works based

on a predefined set of dictionary of words. Corpus linguistics which is the study of real life languages stored in prebuilt computerized databases (Corpora or corpuses). This method uses corpus data to arrive at a particular hypothesis. Corpus linguistics is based on a predefined set of seed words, positive and negative.

III. Machine Learning approach

Machine Learning (ML) provides the systems with the capacity to learn from experience and produce output rather than explicitly programming. [9] ML techniques and algorithms are classified as supervised and unsupervised models. While dealing with sentiment analysis problems, from the research done, it is found that a better and accurate result is produced by supervisory models. [8] Machine Learning is a tool that can be used efficiently to classify the polarity of texts as positive, negative or neutral and beyond that without the need of a predefined set of rules. This increases the flexibility in prediction which in turn helps in a more accurate prediction. Sentiment analysis models can be used efficiently by using machine learning techniques, which helps the model to identify or decide the polarity considering the facts like irony, sarcasm etc.

Un supervised model

In this method, there will be no trainer and no prior data set training. Unsupervised learning method groups and sorts information according to similarities, patterns and differences without prior training of data.

There are two stages in this model. The first stage is clustering where inherent grouping of the data will be discovered. For example, customers can be classified based on the purchase behaviour. The next stage is association where hidden rules are extracted. For example, if a customer buys X he also tends to buy Y.

In unsupervised algorithms, the user first selects the number of classes. Then using clustering algorithm dataset is edited and evaluated. Finally, classification is done and after that evaluation will be performed.

IV. Supervisory model

This model works in the presence of the supervisor to control the whole process. First, machines will be trained with a set of data. After training, machines will be fed with a new set of data. Supervisory model learns from the existing data and applies knowledge to the new set of data. [7] [13]

Supervised methods contain a training process. In this training process or algorithm the training data and its related outputs or labels are taken for final analysis or prediction. Supervised techniques always focus on getting an association between the data samples which are taken as input and its relative output. To start with a supervised learning process, we need two data sets.

For instance, A − Input data − Output data

After deciding the input and relative possible outputs, the next step is to apply a mapping algorithm, b=f(a). The main aim of this mapping algorithm is to narrow down the mapping function so that, when a new data(a) comes we can easily predict the output(b), based on the training data sample instance we have taken.

Decision tree, Random Forest, KNN, Logistic Regression, SVM, Naïve Bayes etc. are some examples of supervised ML algorithms.

The task areas under supervised ML algorithms can be broadly classified into two areas classification and regression. Both are used for predicting the dependent attribute value from data input variables. The difference is that, the value will be numerical for regression and in a categorized manner for classification

Moreover, classification models are applied when the output variable is a category. Regression methods are used when the output is a real value. [10]

Regression

Regression models are mainly used when the expected prediction is real and continuous. Different types of regression models are there, as shown in figure 3.

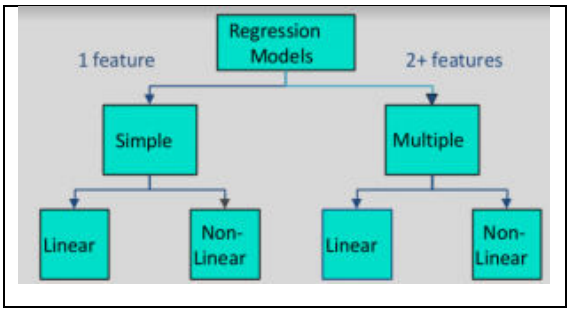

Figure 3: Levels of classification in Regression model

The most commonly used and the simplest one is the linear method. Linear regression is statistically defined as a linear approach to define the relationship between dependent and independent variables.

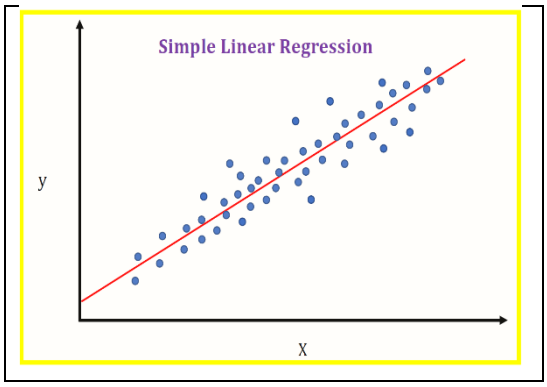

Figure 4: Simple Linear regression

Classification

Classification problems are defined when we need to predict the outcome under different labels. Forexample, it can be age, gender, area, colour of an object, typeetc. The output is dependent on the training data set used by the supervisory model.

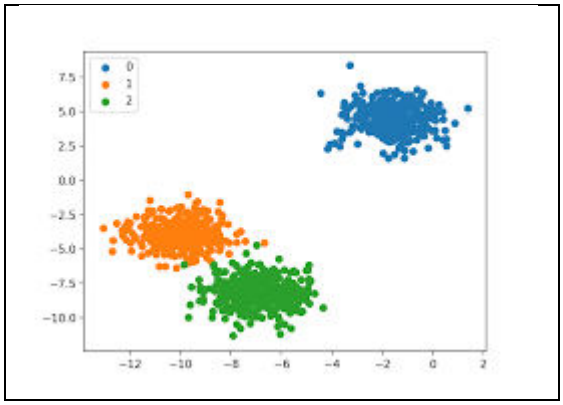

Figure 5: Example for classification

In sentiment analysis depending upon the polarity classification and class labels different types of classifications can be used, like binary, multi class and multilabel. Common way of implementation is binary classification, as in sentiment analysis polarity classification is normally done as positive and negative. Decision tree, Support Vector Machine and Naïve Bayes are the algorithms that are commonly being implemented and having good accuracy values on prediction results.

1. Decision tree classifier

Decision trees are supervised methods. They have to be trained on some sample data. Given a set of documents, the algorithm will calculate the correlation of a word with a particular label. In this classifier, hierarchical decomposition of data is done based on the presence or absence of a keyword. [11] Decomposition continues until the leaf node contains a predefined set of minimum number of nodes.

The categorization of data in a decision tree starts with the calculation of Entropy. It is the measure of randomness in a dataset. The formula for calculating entropy is

$$\sum_{i=1}^{k} P(value_i)\, log_2\big(P(value_i)\big)$$

Decision tree concept works by splitting data into root node, decision node and leaf node. Root node contains the domain data set. Based on decision classifiers it will be split into decision nodes under various labels. Splitting will continue by narrowing down the labels and finally will reach the leaf node, were the actual prediction is formulated.

2. Linear classifier

A. Support vector machine (SVM)

This method is applied on text data where texts are scattered. Few features are irrelevant but they seem to be relevant when they coexist. [17] SVM is a kind of feed-me machine learning algorithm. We can implement SVM for both classification and regression challenges. Classification in SVM is done using a hyperplane, which helps to divide data into different categories. Hyperplane is constructed by SVM with the help of “Kernels”. Kernels are mathematical functions which can be linear as well as nonlinear.

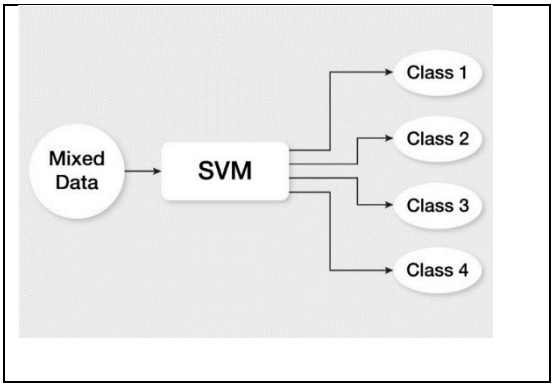

Figure 6: SVM illustration

B. SVM Classification

In SVM all data values are represented as points in a plane corresponding to certain coordinates. Classification is done on these points by creating a decision line which is normally called a hyperplane. The points which are close to the hyperplane are considered for the evaluation process in SVM algorithms. These points are called support vectors. Themargin, which is normally the area between the support vectors and hyperplane, should be measured and compared. Algorithm extracts the hyperplane with maximum margin value as the optimal hyperplane.

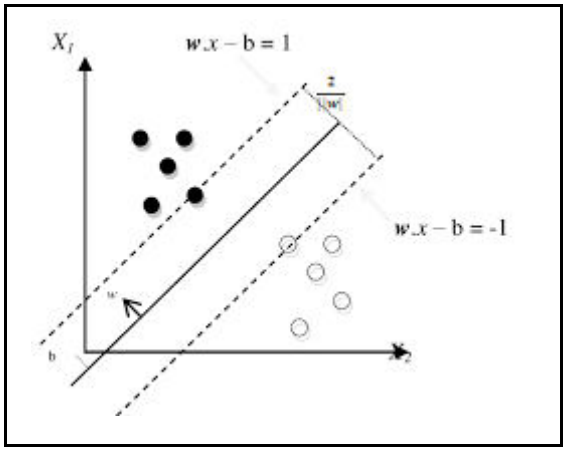

Figure 7: Linear classification in SVM

SVM can be linear as well as nonlinear. If we can divide the entire dataset into two categories by clearly drawing a decision boundary or hyperplane it is called linear SVM.In those case, we have to implement a linear SVM classifier. If data samples are scattered in such a way that we cannot categorize them with one hyperplane, it's a type of nonlinear SVM classifier.

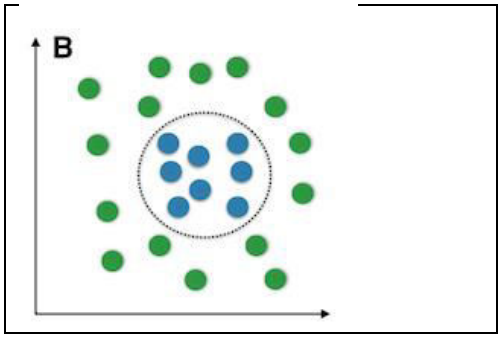

Figure 8. Non-linear classification example

In the case of nonlinear classification problems, we convert the data into a linearly separable data in a higher dimension, by adding one more dimension z=x^2+y^2.

C. Neural network

NN or neural networks in NLP is implemented when some analysis has to be done by synthesizing the textual data. Basic unit of this model is neurons. A weight will be associated with each neuron. Based on the inputs and weights of the neurons, output is formulated. Conversion techniques like word embedding and tokenization are used and numerical equivalent input of corresponding words are figured out. All the rest procedures are based on these neurons which are quantified values and can be used effectively in artificial intelligence for making a correct prediction.

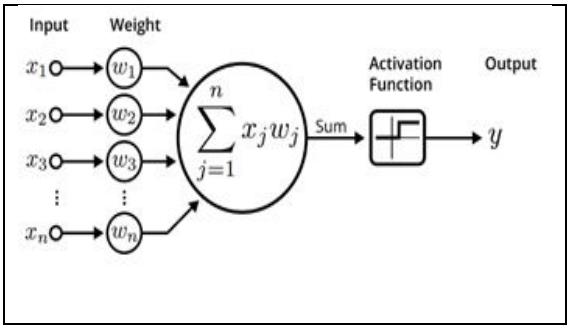

Figure 9. Flow diagram of NN procedure

3. Rule based classifier.

Rule-based approaches always tend to focus on pattern-matching or parsing. Rule based predictions have low precision and high recall. Here, the classifier makes use of the effect of if then rules for classification. If condition, then conclusion is sorted out. The overall process goes through three stages.

First feature extraction and rule learning isperformed. After that relevant words of opinion are extracted from the set and finally prediction on orientation of polarity will be done. Pattern matching is done by the help of a pre trained data which is divided into two categories, positive and negative. This dataset library is called lexicons. After feature extraction each token will be matched with lexicons and marked as positive or negative. The overall polarity is decided by finding out the maximum case. Forexample,like, if the polarity is greater than zero the sentence may be categorized as positive. If it is less than 0 as negative.

4. Probabilistic classifier

A. Naive Bayes

In the naive model, a cluster of algorithms based on Bayes theorem is defined. The output predicts which is independent and has equal features. It helps in classifying data and figures out probabilities of different attributes of data. Application is implemented by splitting text into words. Each word is considered as independent and equally important to work with. The word counts will be converted to probabilities. Naive Bayes model works on these probabilities to produce the output. [15]

B. Bayesian network

This is a probabilistic graphical model, which is applied in the areas like prediction, automation of insights, reasoning and decision making under uncertainty. A graph is made up of nodes and directs links between them. Each node represents a variable. Links show

relationships between nodes. Bayesian network works basically on two components. (1) The structure of the input variable set and the set of dependent and independent assertions. (2) The conditional probability that is defined between the root and child nodes.
Here an unstructured data is transformed to structured data and key terms are extracted. [18] After that text classification will be done and inference is produced. It's a more suitable method for applying in domains where the values are uncertain.

5. Maximum entropy

Maximum entropy model chooses the best from a number of probability distributions. All unwanted assumptions are removed and the one with maximum entropy value is chosen as the best. [14]

It relies on the principle that there should be a uniformity in the probability distribution even in the case of absence of pre-defined knowledge set. The complete methodology constitutes a series of steps like, segmentation, negationhandling, feature selection and finally model evaluation. Maximum Entropy method chooses the distribution with maximum uniformity. Distribution with maximum uniformity is assumed to have maximum entropy.

V. Better performing algorithms

From the various supervisory models analysed four methods are selected and compared for accuracy, training speed and prediction speed.

Table 1: Supervised algorithms

Method	**Average predictive accuracy**	**Training speed**	**Prediction Speed**
Naïve Bayes	Lower	Fast	Fast
SVM	Average	NA	NA
Decision Tree	Lower	Fast	Fast
Maximum Entropy	Higher	NA	NA
Linear Regression	Lower	Fast	Fast
Neural Networks	Higher	Slow	Fast

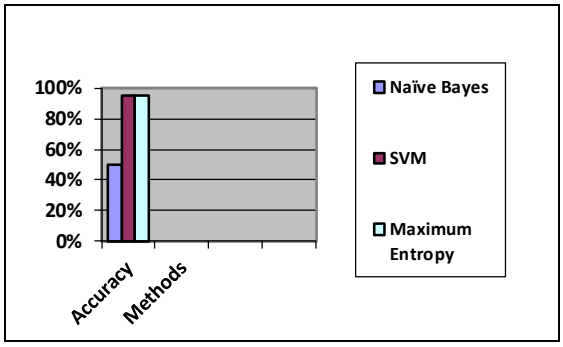

Figure 10. Accuracy of supervised algorithms in terms of percentage

VI. Results and discussions

In this paper different techniques for identifying the polarity in a dataset and measuring the sentiment of a text is examined in detail. Out of the observation done the general steps in finding the polarity of the dataset can be summarized as follows, irrespective of the method or technique implemented.

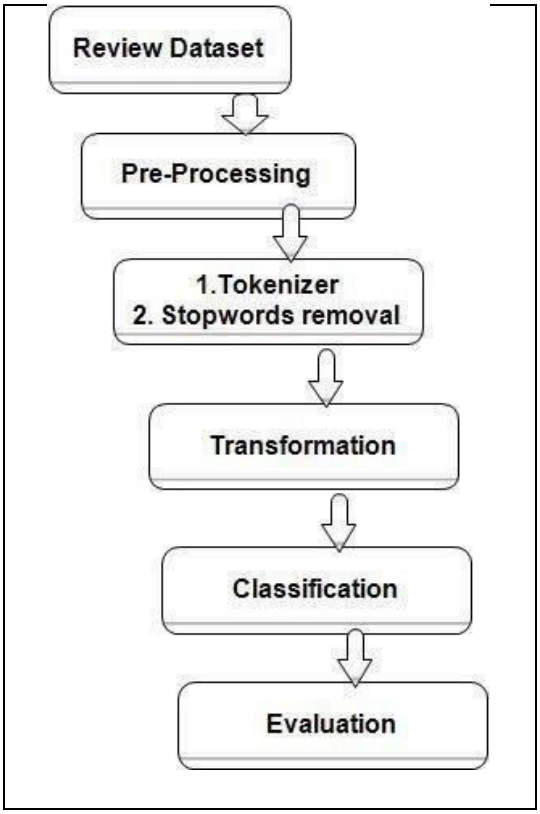

Figure 11. Different phases in polarity detection.

Differences mainly affects two phases, transformation and classification. The final stage, evaluation, that is measuring the accuracy of the prediction gives a better comparison over the different techniques. The process of sentiment classification starts with pre-processing the data, after that extracting the necessary features relevant for predicting the polarity followed by transforming it for classification and last stage is to classify it under different polarities or labels. After the completion of these processes, accuracy can be determined with the help of Precision, recall and if needed F-measure.

Main techniques that are found relevant and most commonly used can be categorized as follows under supervised methods. Tools are divided into two sets based on the nature of the dataset and prediction to be done, continuousand categorical.

Table 2: Tools used by Supervised Algorithms

Method	Tools used by Supervised Algorithms
Continuous	• Regression 1. Linear 2. Polynomial • Decision Trees • Maximum Entropy
Categorical	• Classification o KNN o Trees o Naive Bayes o SVM

VII. Conclusion

In this article the importance of sentiment analysis and various methods to make a conclusive output on opinion mining is analysed. Various supervised learning algorithms are studied and their mode of working is examined. It is found that implementation of supervised algorithms depends on the focus area and type of conclusive opinion needed. A study of accuracy is done on various supervised algorithms. SVM and Maximum Entropy method are the two methods that have highest accuracy. Among this SVM is used more publicly as it seems to give the highest accuracy on product reviews.

VIII. References

[1] WalaaMedhat ,HodaKorashi, "Sentiment Analysis Algorithms and Applications A Survey", Ain Shams Engineering Journal, Volume 5, Issue 4 December 2014

[2] Christine Largeron, Christophe Moulin, Mathias Gery,"Entropy Based Feature Selection for Text Categorisation"

[3] Abhilasha Singh Rathor, Amit Agarwal, PritiDimri,"Comparative Study Of Machine Learning Approaches For Amazon Reviews", International Conference On Computational Intelligence And Data Science, 2018

[4] Deepak Das,"Social Media Sentiment Analysis Using Machine Learning Part 2", September 22, 2019

[5] Jason Brownie, "How to Compare Machine Learning Algorithm in Python with Scikit-Learn", Machine Learning Mastery, June 1 2016
[6] Jeyakumar, Balajee, MA Saleem Durai, and Daphne Lopez. "Case Studies in Amalgamation of Deep Learning and Big Data." In HCI Challenges and Privacy Preservation in Big Data Security, pp. 159-174. IGI Global, 2018.
[7] Victor Roman,"Supervised Learning Basics of Classification and Main Algorithms", 2019
[8] Dipak R Kawade, Dr.Kavita S, Oza, "Sentiment Analysis Machine Learning Approach", International Journal of Engineering and Technology
[9] Sultana, H Parveen; Shrivastava, Nirvishi; Dominic, Dhanapal Durai; Nalini, N; Balajee, J.M. Comparison of Machine Learning Algorithms to Build Optimized Network Intrusion Detection System, Journal of Computational and Theoretical Nanoscience, Volume 16, Numbers 5-6, May 2019, pp. 2541-2549(9).
[10] Priyavrat, A J Singh, Sentiment Analysis," A Comparative Study Of Supervised Machine Learning Algorithms Using Rapid Miner", International Journal For Research In Applied Science And Engineering Technology.
[11] Jaspreet Singh, Gurvindersingh, Rajinder Singh," Optimization of Sentiment Analysis Using Machine Learning Classifier", Human Centric Computing and Information Sciences
[12] Vidisha M Pradhan, Jay Vala, Prem Balani,"A Survey on Sentiment Analysis Algorithms for Opinion Mining", International Journal of Computer Application, Volume 133 January 2016
[13]Sangharshjit S Kamble, Prof.A.R.Itkikar,"Study of Supervised Machine Learning Approaches for Sentiment Analysis", International Research Journal of Engineering and Technology, Volume 5 Issue 4 April 2018
[14] NipunMehra, Shashikant Khandelwal, Priyank Patel," Sentiment Identification Using Maximum Entropy Analysis of Movie Reviews"
[15] Amit Gupte, Sourabh Joshi, Pratik Gadgul, Akshay Kadam, A Gupte, "Comparative Study of Classification Algorithms Used In Sentiment Analysis."
[16] Vidisha M Pradhan, J Ay Vala, Prem Balani,"A Survey on Sentiment Analysis Algorithms for Opinion Mining", International Journal of Computer Application, Volume 133 Number 9 January 2016
[17] Sherin Mariam John, K Kartheeban,"Sentiment Scoring and Performance Metrics Examination of Various Supervised Classifiers", International Journal of Innovative Technology and Exploring Engineering Volume 9, Issue 2S2, December 2019
[18] Min Zhao,TanmingChen,DachengQu,HongQu,"METSP:A Maximum Entropy Classifier Based Text Mining Tool For Transporter Substrate Identification With Semi Structured Text", Biomed Research International, Volume 2015 ,October 2015
[19] WaalaMedhath,AhmadHassan,HodaKoreshy ,"Sentiment Analysis Algorithm and application a Survey",AinSahms Engineering Journal, Volume 5, Issue 4, December 2014, Pages 1093-1113
[20]MeghaJoshi,AyeshaShaikh,PurviPrajapatu,Vishwavala,"A Survey on sentiment Analysis",International Journal of Computer Applications,April 2017
[21]SyedSaood Zia1, Sana Fatima2, IdrisMala3, , M. Sadiq Ali khan, , M. Naseem5, , Bhagwan Das6," A Survey on Sentiment Analysis, Classification and Applications", International Journal of Pure and Applied Mathematics, Volume 119 No. 10 2018, 1203-1211

[22] DoaaMohey El Din Mohammed Hussein,"A Survey on sentiment analysis challenges", Journal of King Saud University - Engineering Sciences, Volume 30, Issue 4, October 2018, Pages 330-338.