International Journal of Engineering & Technology, 7 (2.7) (2018) 676-681

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research Paper

Sentiment Analysis of Movie Review using Machine Learning Techniques

V.Uma Ramya [1*], K. Thirupathi Rao [2]

Koneru Lakshmaiah Educational Foundation, Green Fields, Vaddeswaram, Guntur Dt., AP, India.
**Email: umaramyavigrahala@gmail.com*

Abstract

Today's online world was fully filled up with blogs, views, comments, posts through various websites and social-surfs. People were habituated with posting every incident into blogs, messed with comments like text and emotions, which are a mixed bag of sad, happy, worry, cry etc. Analysing such data was called as Sentimental Analysis. To analysis, these unordered data we use new emerged technology algorithms. Machine learning a transpire technology which is engaged with almost all the fields, where its algorithms are more powerful that give with better faultless results. In this paper, we are analyzing tweets based on movie reviews using the Multinomial Logistic Regression, Naïve Bayes, and SVM algorithms to compare score value to show the best text analysis algorithm.

Keywords: *Sentiment Analysis, Opinion Mining, Twitter Analysis, Machine Learning, Natural Language Processing.*

1. Introduction

As the world keeps on moving with its new innovations in every field, people also habituated to the technology often. Each and every field Internet is growing and showing its most prominent role in the world. Eventually, people also in their own busy world using the Internet as much they can. The Social site for both entertainment and educational purpose, running in a successful manner. Especially social networking is growing in a wide range. The social sites like Facebook, WhatsApp, Instagram Twitter etc., are tuned to everyone's daily routine. Few online Shopping sites like Alibaba, Amazon, Flipkart, Shein etc are exciting people with their own way of business strategies day by day. Feedback from all these sites, about a movie review, product details, satisfactory level, about offers they wish for us, the blog or post in all social sites, etc., all are being increased every day. Together all these are producing a wide range of data.

Generally, sentiment analysis targets to work with a user's point of view over a subject with polarity or interaction over a document or emotional review of the text [1]. The general user's point of view might be emotion, analysis, judgment, conclusion. The technique used here is that predict the content of the statement by its features in words and emojis used. For example, a positive sentiment is, "The weather is cool" on the other hand, a negative sentiment is "The weather is unbearable". The texts are supposed not to be stating any sentimentality, such as news header.

There are various sources of data from the web, sites, blogs etc., can be held to give a superior allowing of user opinion such as problem are at the between of natural language processing(NLP) and data mining research. Analysis all these data is important in business aspect for their growth, sentimental analysis has come into the picture. [1] Measuring peoples level of satisfaction is called Sentimental Analysis. Categories of sentiment analysis are,

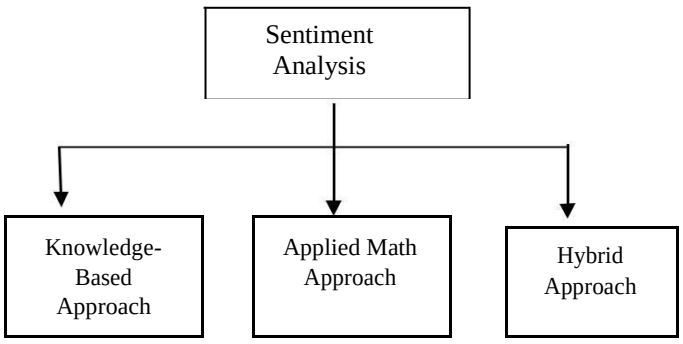


Figure1: Categories of sentiment analysis.

Knowledge-based techniques classify the text-based which is unambiguous like happy, sad, afraid, and bored. Some context bases are not likely to not involving any words influenced by related to assign available to use the words likely to understand to explicit emotions.

Applied math approach ways hold on parts of machine learning respect latent linguistics analysis, support vector machine, "bags of words" and linguistics. To extract the opinion in context and find the feature regarding that the speaker has opined, the grammatical relationships of words square measure used.

Hybrid approach max advantages on each machine learning and parts from information illustration adore ontologies and linguistics networks to notice linguistics that square measure expressed in an exceedingly refined manner, e.g., through the analysis of ideas that

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

don't expressly convey relevant data, however that square measure implicitly coupled to alternative ideas that do, therefore.

Python is both Application-oriented and Problem-oriented language which is used by thousands of people for testing, games with PyGame library etc, it provides a dynamic and automatic management for systems which supports different programming models incorporated with imperative, object-oriented, functional and procedural with a comprehensive standard library. Here we have used four python packages,

NumPy: NumPy is a Python library which adds reinforce for multi-dimensional arrays. It speedily integrates with a wide variety of databases and integrates with C/C++ also.

Scikit-learn: Scikit-learn a Machine learning library for Python which is free of cost and provides many Classification, Regression, and Clustering algorithms that include Support Vector Machine, Random Forest, Gradient Boosting, k-means, and DBSCAN. It has the ability to exchange and use information with NumPy and Spicy.

Natural Language Toolkit: Natural Language Toolkit (NLTK) is a package of libraries. It helps in writing programs for representative token and to process common language using Python which is written in English. It includes graphical demonstration and sample data which accompanied by concepts of processing tasks, plus a cookbook.

Tweepy: Tweepy is a Python library which is used to access the twitter data using Twitter API. It provides an OAuth with authentication along its access tokens. It downloads your timeline tweets and prints each one of their text.

2. Keywords

Twitter: The Tweet is a collection of "bags of word" which is presently used in a wide range. It approximately produces the range of data 200 million tweets a day. A Tweet in the user point of view to share the daily life to the public. Each and every user have their own point of view to express using these tweets. So, to analyze these texts based natural language processing techniques are used. Analysis all these data is important in business aspect for their growth [2].

Sentiment Analysis: Analysing the view in users point to predict the output with the percentage of positivity and negativity is called Sentimental Analysis. With the help of results, we can predict the ups and downs of a business point of view, to improve the quality of the film, to make more good books, most used product etc, these techniques are powerful and used at the wide range for predicting future and come to a judgment. So, sentiment analysis is important [2].

Machine Learning: Machine learning [3] is the application help for computer's self-development and adjusts automatically without human intervention. Machine learning classified into 3 different models where each has its own way of problem-solving techniques. However, to solve any problem there are two steps. In the first one allow the model to learn from existing data up to a maximum of 80% and then in the second step apply an algorithm with new data set of 20%, using model trained before. In our paper, we are using Naive Bayes, Support Vector Machine (SVM), Multinomial Logistic Regression the Machine learning algorithms that are supported by Natural Language Processing of text.

Table 1: Comparison of machine learning algorithms.

Algorithm	Advantage	Disadvantage
Naïve Bayes [5][6][7]	Easy to implement. Obtains good results in most cases even with small amount of training set. It helps to scale up for large training set also. Best algorithm for text classification and analysis.	It cannot difference between continuous features of words. It takes more runtime memory.
SVM [7][11]	Robust against overfitting, especially in high dimensional space.	Can't scale up and less preferred over the random forest
Multinomial Logistic Regression [10]	Models can be updated with new datasets and avoid overfitting.	Need huge amount of training dataset

3. Literature Survey

Jayashri khairnar and Mayura Kinikar [4] illustrated Sentiment Analysis using Machine Learning Approaches which provides a solution for classification problem and is mainly divided into two steps with training data and model based on the model to be trained from predefined rules and classifying the unrevealed data using Naive Bayes classification

$$P_{NB}(c/d) = \frac{P(c)(\pi_{i=1}^{m} P(f_i|c)\, n_i^{(d)})}{P(d)}$$

and Support Vector Machine(SVM) has been shown with highly traditional text categorization outlaying Naive Bayes theorem

$$\vec{w} := \sum_j a_j c_j \vec{d}_j \quad a_j \geq 0$$

and the evaluation of sentiment classification with confusion matrix with actual positive and negative instances. Ruchi Mehra, Mandeep Kaur Bedi, Gagandeep Singh, Raman Arora, Tannu Bala and Sunny Saxena [5] provided information of classification of Sentiment analysis data using Naive Bayes and Fuzzy algorithms which provides the data that is been collected from different social websites such as Facebook, Twitter with a list of parameters that provides the Accuracy, Recall and Precision.
Bingwei Liu, Erik Blasch, YuChen, Dan Shen and Genshe Chen [6] depicted information of sentiment analysis using Naive Bayes classifier on Big Data using the Hadoop where the problem is classified into Training Job[18-21]. Combining Job and Classify Job. It also provided information on the virtual Hadoop cluster that is built on the cloud where the results are differentiated with the cloud is weaker to Hadoop cluster.

4. Methodology

General Sentimental Analysis: A basic system that performs with a level of sentiment analysis on Twitter based on movies with following steps:

a. Fetch data from Twitter.
b. With Feature selection and Word Features, get both positive and negative tweets.
c. With the help of Training set both Positive and Negative tweets and using Classifier we can analyze tweets who are positive and negative.

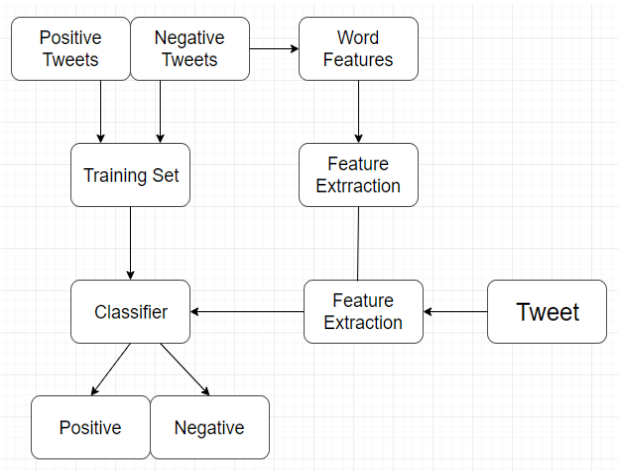


Figure 2: General Sentiment Flowchart. Source: [17]

- From above explained general sentiment analysis it says that, for the tweets retrieved we apply the word search and feature extraction to a classified set of tweets and produce the sentiment overall.
- Analysing all the tweets from twitter, our prediction model defines the sentiment (Positive, Negative, Neutral) over a "keyword" based on key search method from all the tweets.

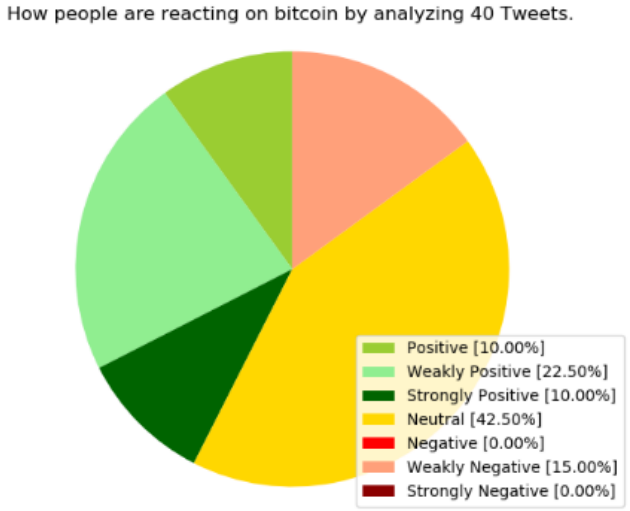


Figure 3: Description of general Sentiment Analysis

- Showing the result, users judgment for the word "bitcoin" shown in pie chart. 42% are positive, 42% are neutral, 15% are negative. Hence classified the tweets according to their sentiment.

5. Proposed Work

A model is proposed to analyze the data using tweet dataset. There are few steps that process the tweets

a. Data Preparation from Twitter
b. Pre-processing
c. Creating dataset
d. Applying Algorithms

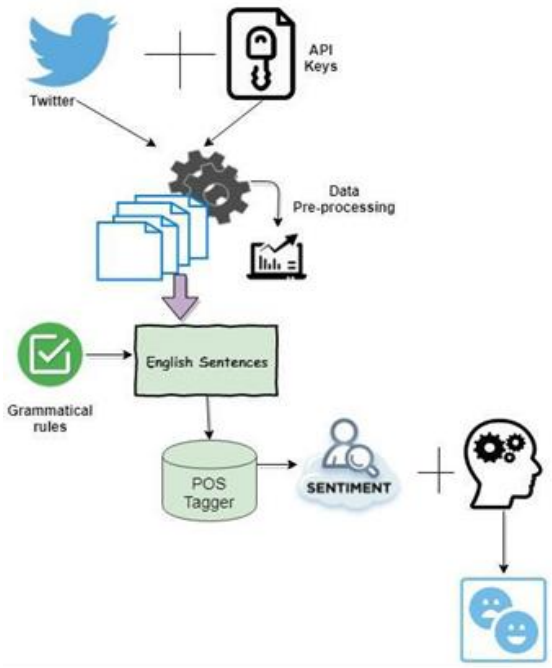


Figure 4: Description of proposed Sentiment Analysis

Data Preparation from Twitter*:* Firstly, to get data for sentiment, we need to retrieve data from Twitter. We need to access them through Twitter API. Steps as follows,

- Register with Twitter application developer's portal to get authentication for getting data from Twitter.
- After the application is created with your twitter account, we need to access the Access Tokens and Keys, they provided.
  consumer_secret: < Your Twitter Consumer Secret Key>
  consumer_key: <Your Twitter Consumer Key>
  access_token: < Your Twitter Access Token Key >
  access_secret: < Your Twitter Access Secret Key >

- Now run the code to fetch the data from Twitter.
- It creates a TEXT file into the same directory.

Pre-processing of datasets*:* A tweet is combination different opinions. It might be structured or unstructured or uses some different own languages to express the feelings in a Tweet form. For processing these tweets, we need to first classify them positive and negative tweets by pre-processing steps that are to be followed as follows,

- Eliminate all URLs, hashtags, targets.
- Correct the spellings and repeated characters.
- Substitute all emotag with their respective sentiments like happy sad, fear etc.
- Eliminate stop words
- Remove Non-Negative Tweets

Creating Datasets*:* To process data and extract the file as .csv into your directory,

- For data, pre-processing, extract the data from a text file to process it.
- It follows the all the previous steps to change the data.
- Finally, it automatically creates the .csv file into the same directory.

Applying Algorithms*:* Machine learning algorithms are the powerful algorithms that result in accuracy. In our paper, we are using Naïve Bayes, SVM, and Multinomial Logistic Regression. These three are powerful algorithms for classifying the textual data and process them with POS tagger with python language, for more accurate results.

Here we are using the TF-IDF (Frequency Inverse Documentation Frequency) method, numerical statistics that is indeed to reflect the importance of a particular word from a document.

TF-IDF Technique: It shows the reflection of the words used in the document. A weighted approach in searches of IF (Information Retrieval), Data mining. Most recommended systems in the domain of digital libraries. Collection of the TF-IDF weighted-plan is frequently used via web indexes as a focus in scoring and positioning an archive's significance given a client inquiry. [16] TF-IDF can be efficiently used for stopwords by separating in different subjects, including content rundown and characterization.

TF: Term Frequency, which measures the frequency of a particular word occurs in a report. Since each report is diverse long, it is conceivable that a term would seem significantly big circumstances in lengthy records than shorter ones.

In this manner, the term recurrence is regularly separated from the length of the report (otherwise known as. the aggregate number of terms in the record) as a method for standardization:

$$\mathrm{TF(t)} = \frac{Number\ of\ times\ term\ shows\ up\ in\ a\ report}{Total\ number\ of\ terms\ in\ the\ archive}$$

IDF: Inverse Document Frequency, shows the imperative of phrase is. While registering TF, to consider all the terms similarly critical. In any case, it is realized that specific terms, for example, "is", "of", "that", might repeat yet a tiny significance.

Along these lines, we have to overload the incessant terms while scaling up the uncommon ones, by registering the accompanying:

$$IDF(\mathrm{t}) = \log_e(\frac{Total\ number\ of\ archives}{Number\ of\ reports\ with\ term\ t\ in\ it})$$

We are going to apply this technique for three machine learning algorithm that are described with results as follows below.

The Algorithms for our proposed work defined follows:

Algorithm for retrieving Tweets

```
Begin
    While Authentication (Success) do
        if searchQuery(Found) then
                for tweetCount → 1 to reqCount do
                        Download Tweets to a Text file
        While (tweetCount < reqCount) do
            Search for new tweets
                if search found(NewTweets) then
                        Append new tweets to the Text file
                else
                        Print "No more tweets found"
        else
            Print "No Tweets Found!"
End
```

Algorithm for Data Pre-processing

```
Begin
        Define movie_name, rating, file_name
        for tweets in text file do
                extract searchQuery from a tweet and map it
                with movie_name
                remove all URL's, stopwords, usernames, hy-
                perlinks & trashy content and arrange in a
                clean format
                save the dataset into a CSV file
End
```

Algorithm for Naïve Bayes, SVM, Multinomial Regression

```
Begin
        Define Naïve_Bayes, SVM, Mnlr
        Classify_naive_bayes( train_x, train_y, test_x, test_y)
                Mnb.fit(train_x,train_y).predict(test_x)
                yHat = classify_naive_bayes(X_train,
         Y_ Train, X_test)
                Conf_mat = confusion_matrix( Y_test, yHat)

        Classify_SVM( train_x, train_y, test_x, test_y)
                svm.fit(train_x,train_y).predict(test_x)
                yHat = classify_naive_bayes(X_train,
         Y_ Train, X_test)
                Conf_mat = confusion_matrix( Y_test, yHat)

        Classify_Mnlr( train_x, train_y, test_x, test_y)
                Mnlr.fit(train_x,train_y).predict(test_x)
                yHat = classify_naive_bayes(X_train,
        Y_ Train, X_test)
                Conf_mat = confusion_matrix( Y_test, yHat)
End
```

Naïve Bayes: Naïve Bayes is generally used for text-based classification method. It is a supervised Learning technique which is used for classification problem [6], widely used when high dimensional training datasets. For example, Naïve Bayes mostly used for Spam Filtration, Sentimental Analysis, Classification new articles. With the help of Machine learning can build model fast and get quick predictions. With a well-trained dataset, it is possible to make good results with a text classifier.

Implementation using Naïve Bayes:

The following code for Naïve Bayes Algorithm performs Grid Search using algorithm. It helps to Fit the dataset and predict with X_train and Y_train.

```
The model is trained on the full development set.
The scores are computed on the full evaluation set.

              precision    recall  f1-score   support

          10       1.00      1.00      1.00        71

avg / total       1.00      1.00      1.00        71


f score is:
1.0
```

Figure 5: Score values of Naïve Bayes

The above result shows the frequency- Score, a measure of test accuracy (an harmonic value of precision and recall value). The Naïve Bayes shows the best value as 1.0. Hence to conclude that, the algorithm works perfect.

SVM: Support Vector Machine(SVM) is known as the best classifier that provides the most accurate results in speech classification problems [7]. They achieved by creating a hyperplane with maximal Euclidean distance for the nearest trained examples. Support Vector Machine hyperplane is completely resolved by a comparatively minute subset of the trained data sets which are treated as support vectors. The remaining training data sets have no access to the qualified classifier. So for the purpose of the text classification, the classifier SVMs have been used most fields.

Implementation of SVM:

Following code for Support Vector Machine performs Grid Search using algorithm. It helps to Fit the dataset and predict with X_train and Y_train

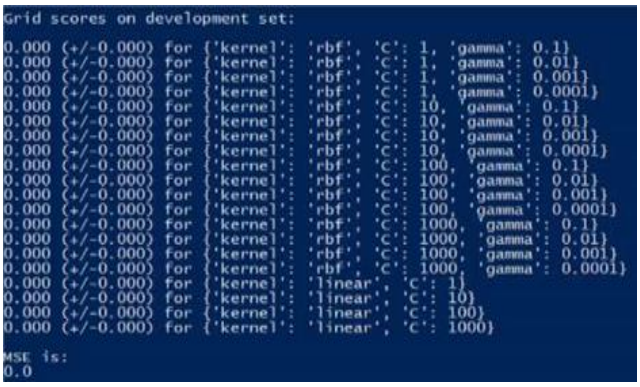


Figure 6: Score values of Support Vector Machine

The above result shows the MSE (Minimum squared error), a measure of test accuracy (an harmonic value of precision and recall value). The SVM shows the as 0.8. Hence to conclude that, the algorithm works better.

Multinomial Logistic Regression(MNLR): Multinomial logistic regression is preferred when we have the feature of different type nevertheless given that it is a regression model [10]. It is avoided when our features are highly correlated. Multinomial Logistic Regression is also called as SoftMax regression in place of ordinal logistic regression. Multinomial Logistic Regression works well on big data irrespective of different areas [16].

Implementation of MNLR:

Following code for Multinomial Logistic Regression performs Grid Search using algorithm. It helps to Fit the dataset and predict with X_train and Y_train

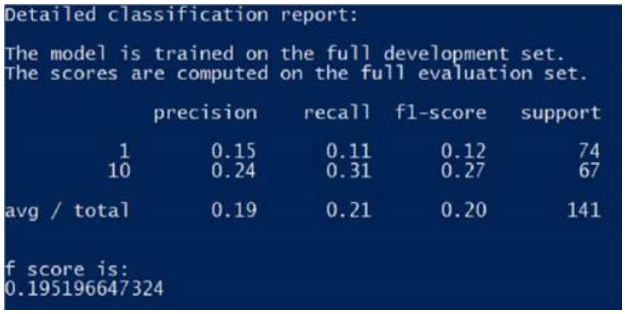


Figure 7: Score values of Multinomial Logistic Regression

The above result shows the Frequency- Score, a measure of test accuracy (a harmonic value of precision and recall value). The Multinomial Logistic Regression, shows the best value as 0.19. Hence to conclude that, the algorithm works not even better.

6. Comparison of Results

Precision*:* The value of the entire true predicted against all predicted. A measure of how frequently an assumption rating was right. For reports with tonality, precision tracks what number of those that were appraised to have tonality were evaluated accurately.

Recall: A measure of what number of reports with feeling was evaluated as nostalgic. This could be viewed as how precisely the framework decides impartiality. By and large, high review scores are exceptionally troublesome in the trial of the wide topic, as the framework is required to see ever-bigger arrangements of words and dialect.

F- Measure*:* Precision and recall can be together to procedure a single metric are known as F- measure, which is the weighted harmonic mean of precision and recall.

Table 2: Confusion matrix report.

Algorithm	Precision	Recall	F-Measure
Naïve Bayes	1.0	1.0	1.0
Support Vector Machine	1.0	0.7	0.8
Multinomial Logistic Regression	0.90	0.86	0.19

From above results, we can conclude that Naïve Bayes algorithm performs better for text analysis when compared to Support Vector Machine (SVM) and Multinomial Logistic Regression.

7. Conclusion

Sentiment Analysis is just a part of social media monitoring platform which will be beneficial for a business to improve the productivity. It shows the ones feeling or opinion over an object. We conclude by that Multinomial Naïve Bayes with Machine Learning algorithm that produces an improvised result when compared to the other classifier algorithms for language processing like Multinomial Logistic Regression and Support Vector Machine (SVM).

References

[1] B Pang, L Lee- "Opinion mining and sentiment analysis" Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135

[2] A Pak, P Paroubek- LREc, "Twitter as a corpus for sentiment analysis and opinion mining" International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta

[3] Khimar, J., Kinikar, M. (2013). "Machine Learning Algorithms for Opinion Mining and Sentiment Classification". International Journal of Scientific and Research Publications, 3(6),1-6, ISSN 2250-3153, Volume 3, Issue 6, June 2013.

[4] Jayashri Khaimar* Mayura Kinikar, "Machine Learning Algorithm for Opinion Mining and Sentiment Classification", International Journal of Scientific and Research Publications, 1 ISSN 2250-3153, Volume 3, Issue 6, June 2013

[5] Ruchi Mehral Mandeep Kaur Bedi2, Gagandeep Singh3, Raman Anaroa4, Tannu Bala5, Sunny Sazena6 Webtunix Solutions Pvt, Ltd. Mohali, "Sentimental Analysis using Fuzzy and Naïve Bayes" International Journal of Scientific and Research Publications, ISSN 2250-3153, Volume 3, Issue 6, June 2013

[6] Bingwei Liu∗, Erik Blasch†, Yu Chen‡, Dan Shen∗ and Genshe Chen, "Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier" Big Data, IEEE International Conference,2103 Electronic ISBN: 978-1-4799-1293-3, December 2013

[7] Shweta Rana, Archana Singh "Comparative analysis of sentiment orientation using SVM and Naïve Bayes techniques", Next Generation Computing Technologies(NGCT), Electronic ISBN: 978-1-5090-3257-0, 2016.

[8] Ankur Goel, Jyoti Gautam, Sitesh Kumar, "Real-time sentiment analysis of tweets using Naïve Bayes", Next Generation Computing Technologies(NGCT), Electronic ISBN: 978-1-5090-3257-0, 2016.

[9] Human Parveen Shikha Pandey "Sentiment analysis on twitter Data-set using Naïve Bayes algorithm", Applied and Theoretical Computing and Communication Technology (iCATccT), Electronic ISBN: 978-1-5090-2399-8, 2016.

[10] Vryniotis Vasils and Vasilis Vryniotis" Machine Learning Tutorial: The Multinomial Logistic Regression (Softmax Regression) May 2017.

[11] Zainuddin, Nurulhuda and A. Selamat," Sentiment Analysis Using Support Vector Machine", International Conferences on Computer, 2014.

[12] T Gunasekhar KT Rao, MT Basu "Understanding insider attack problem and scope in cloud" Circuit, Power and Computing Technologies(ICCPCT), 2015

[13] T Gunasekhar, KT Rao, "EBCM: Single encryption, multiple decryptions", International Journal of Applied Engineering Research 2014.

[14] KT Rao, PS Kiran, LSS Reddy " High-Level Architecture to provide cloud services Using Green Data Center", Advances in Wireless and Mobile Communications (AWMC), 2014.
[15] KT Rao, PS Kiran, DLSS Reddy, VK Reddy, BT Rao, "Genetic Algorithm for Energy Placement Of Virtual Machines in Cloud Environment". proc IEEE International Conference on Future Information Technology, 2012.
[16] Ramadhan WP, strip Novianty S.T, Casi Setianing S.T., M.T "Sentiment Analysis Using Multinomial Logistic Regression", International Conference on control, Electronics, Renewable Energy and communication(ICCEREC), 2017
[17] Vishal A. Kharde, S.S Sonawane " Sentiment Analysis od Twitter Data: A Survey of techniques", International Journal of Computer Applications, Volume 139-no.11April 2016.
[18] KISHORE, P.V.V., KISHORE, S.R.C. and PRASAD, M.V.D., 2013. Conglomeration of hand shapes and texture information for recognizing gestures of indian sign language using feed forward neural networks. International Journal of Engineering and Technology, 5(5), pp. 3742-3756.
[19] RAMKIRAN, D.S., MADHAV, B.T.P., PRASANTH, A.M., HARSHA, N.S., VARDHAN, V., AVINASH, K., CHAITANYA, M.N. and NAGASAI, U.S., 2015. Novel compact asymmetrical fractal aperture Notch band antenna. Leonardo Electronic Journal of Practices and Technologies, 14(27), pp. 1-12.
[20] KARTHIK, G.V.S., FATHIMA, S.Y., RAHMAN, M.Z.U., AHAMED, S.R. and LAY-EKUAKILLE, A., 2013. Efficient signal conditioning techniques for brain activity in remote health monitoring network. IEEE Sensors Journal, 13(9), pp. 3273-3283.
[21] KISHORE, P.V.V., PRASAD, M.V.D., PRASAD, C.R. and RAHUL, R., 2015. 4-Camera model for sign language recognition using elliptical fourier descriptors and ANN, International Conference on Signal Processing and Communication Engineering Systems - Proceedings of SPACES 2015, in Association with IEEE 2015, pp. 34-38.