ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

Sentiment Analysis with Vector Feature Extraction and Classification of Social Media Dataset

[1] Misha Jain, [2] Dr. B. K. Verma
[1][2] Department of computer science
[1][2] Chandigarh Engineering College, Landran, Mohali

***Abstract* - The paper presents a methodology used for sentiment analysis. Data to be analyzed will be extracted from social media sites like twitter. Feature extraction will be done using support vector machine. Instance selection will be done using genetic algorithm operators: Selection, crossover and mutation operators. Classification of sentiments will be done using back propagation neural network technique. Training and testing phase evaluates various performance parameters: False Rejection Rate, False Acceptance Rate and Accuracy.**

Index Terms: Sentiments, Sentiment Analysis, Genetic algorithm, Feature extraction, Back propagation neural network,. Genetic operators.

1. INTRODUCTION

Sentiment is view, opinion of a person for some product, occasion or service. Sentiment Analysis is a stimulating Text Mining and Natural Language Processing problem for automatic extraction, organization & summarization of opinions and emotions expressed in online text [1]. Sentimental analysis is great for business intelligence applications, where business analysts can analyze public sentiments about products, services, and policies [2].

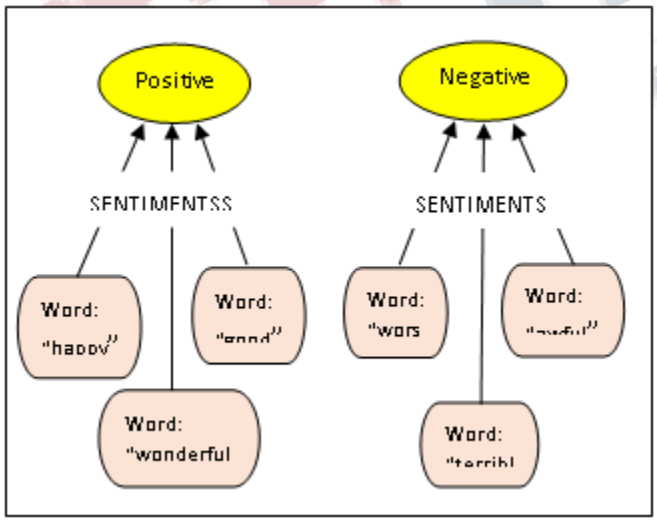

Fig 1. Sentimental Analysis

The sentiment initiate within comments, feedback or critiques and provide useful indicators for many dissimilar purposes. These opinions can be categorized either into two categories: positive and negative; or into an n-point gauge, e.g., very decent, good, acceptable, bad, very bad.

A. Sentiment Analysis Techniques

Sentiment Analysis can be done in three ways:-

- Sentiment Analysis based on Supervised Machine learning method,
- Sentiment Analysis by using Lexicon based Technique and
- Sentiment Analysis By combining the above two approaches.

We are using Supervised Machine learning method. In this method, two types of data sets are required: training dataset and test data set. An automatic classifier learns the classification truth of the document from the training set and the accuracy in classification can be evaluated using the test set.

B.Feature Extraction in Sentimental Analysis

Text Analysis is a main application field for mechanism learning processes. However the raw information, an order of symbols cannot be fed straight into the algorithms themselves as maximum of them expect arithmetical feature paths with a fixed size somewhat than the raw text forms with variable length. In imperative to address this, sickie-learn offers utilities for the most mutual ways to extract numerical structures out of texts, as follows:

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

•Tokenizing the strings and giving an integer id for each imaginable token, for example by using white-spaces & punctuation as symbolic separators [9].
•Counting the existences of tokens in each document.
•Regulating and weighting with diminishing importance tokens that occur in the majority of samples / forms.

C. Classification of Sentiment Analysis

Sentiment analysis can defined at dissimilar levels:

• *Document Level Classification*
In this procedure, sentiment is extracted from the complete review and sentiment is classified-based on the overall sentiment of opinion holder. The main goal is to classify a review negative, neutral and positive.

• *Sentence Level Classification*
This approach has two type:
a)Subjective classification of a sentence into one of binary classes: objective and subjective. This analysis distinguish sentences that express factual information from sentences to express views or opinion
b)Sentiment classification of particular sentences into binary follows: Negative and Positive. This looks for opinion itself and targets it.

• *Feature Level Classification*
This process goal is to search and extract entity features that have been commented on by the emotion holder and define whether the opinion is negative, positive and neutral. Feature similar meanings are grouped, and a featured based summary of multiple reviews is formed.

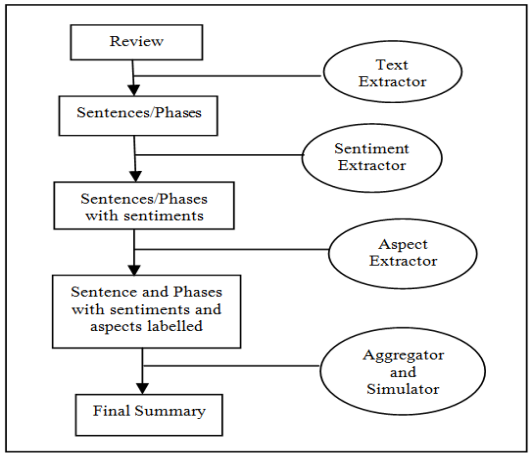

Fig 2. Classification Levels of Sentimental Analysis

II. RELATED WORK

Svetlana Kiritchenko (2014) [8] In this paper Support Vector Machine and Message-Level Task is used for detecting purposes behind political tweets, for feelings in text and to recover the sentiment lexicons by producing them from larger quantities of data, and from dissimilar kinds of statistics, such as tweets, blogs, and Facebook posts. The issue to a two-way classification task was reduced. The optimum threshold in unsupervised settings was set to escape the difficult.

Muhammad Zubair Asghar (2014) [3] used Clustering based Feature Extraction, NLP based, Machine learning. Reduction, Redundancy Removal and evaluating performance of hybrid methods of feature selection. Main issue connected with clustering based frequent feature selection methods is their domain dependency in terms of heuristics and threshold setting.

Doaa Mohey El-Din Mohamed Hussein (2016) [6] used Natural Language. Outcomes of the average of accuracy based 212 on the number of studies in each challenge. The more the 213 study in a sentiment experiment, the less the Average of accuracy rate. Facing the sentiment analysis and estimation process.

Ravendra Ratan Singh Jandail (2014) [7] In this paper Support Vector Machine is used. For the mobile phone only, it contained 16 most usable features and their connected keywords for the mobile phone. Unstructured and Ungrammatical text, a fact that tweet communications are not always accurate and Ambiguity.

Maria Pontik (2015) [10] used Aspect classes, opinion target words, and polarity classification. Achieved the best score. ABSA problem has been formalized into a righteous unified framework in which all the well-known constituents of the conveyed opinions meet a set of conditions and are linked to each other within the tuple.
Sara Rosenthal (2015) [11] Classified using SVM Moving to a well-ordered five-point scale resources moving from binary classification to ordinal regression. Problem of sentiment polarity classification and our subtasks.

Deepali Virmani (2014) [15] In this paper Sentiment Analysis was done in collaboration with opinion

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

extraction. Scores were assigned to each sentiment, word in the database. Cooperated opinion is estimated by teacher's remarks word by word and then implementing proposed algorithm. The remark related to the issue is to be analyzed by concerned organization to enhance their skills.

III. PROBLEM FORMULATION

The most difficult task is to analyze the human emotions which are very diverse. Difficulty lies in the fact that there could be mixed opinions of people expressed on social media sites like Twitter. With the creative nature of natural languages, people, might express the similar sentiments in vastly dissimilar ways.
In cross language sentiment classification based on support vector machine, only statistics were used to extract the feature words in the feature selection stage and the classifier could not adapt the target language well. To solve all these issues, new proposed technique will be implemented using techniques like genetic algorithm for instance selection and classification of sentiments will be done using back propagation neural network technique

IV. TECHNIQUES USED

A. Feature Extraction using Vectoring

Feature extraction technique is used to recover most revealing terms from amount of matrix. This study used Principle Component Analysis technique to calculate and study the Eigen vector and values to find the feature values and then to direct individual data with its principle components / Eigen Vectors [13].

B. Instance Selection using Genetic Algorithm Optimization

Instance selection is a data optimization approach. Main task of instance selection is to eliminate some malicious characteristics of a given data set. In transactions with selection of instances to optimize the size of the matrix and would easily in processing to deal with the further proceeding input. Genetic algorithm optimization is an instance based method which is used to optimize the instances of the sentiment words [14].

C. Back Propagation Neural Network for Classification

Network of Neural is a computational scheme inspired by the arrangement, dispensation technique, and knowledge ability of an organic brain. The essential dispensation rudiments of neural systems are named artificial neurons. It is a simplified arithmetical mold of the neuron.
In Back propagation neural network, the constant individual is fed into the output unit and the network is run backwards. Incoming information to the neurons is included and the consequence is multiplied by the value reserved in the left part of the unit. The consequence is transmitted to the left of the unit and collected at the input unit is derivative of the network function.

V. PROPOSED METHOD

A. Objectives

This research work will be focused to achieve the following objectives:-

- To develop an algorithm to enhance the feature extraction, selection and classification of the sentiment analysis.
- To evaluate the performance parameters of False Acceptance Rate, False Rejection rate and accuracy.
- To validate our new approach by comparing it with the existing methods.

B.Methodology

1. Dataset is uploaded with social media data and divided into three categories: positive, negative, and neutral.
2. Feature extraction algorithm is applied on the dataset.
3. Features are extracted in the form of Eigen vectors and Eigen values.
4. For instance selection, genetic algorithm is applied. Population size is selected and initialize the genetic algorithm operators are initialized.
    - Selection operator is used to initialize the data.
    - Crossover operator is used to divide the data into two categories according to Eigen values and vector range.
    - Mutation operator is applied for end movement modification.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

5. Fitness function is applied to calculate the f-value.
6. Reduce index and the best index value is obtained.
7. For classification of the sentiments, back propagation neutral network technique is applied.

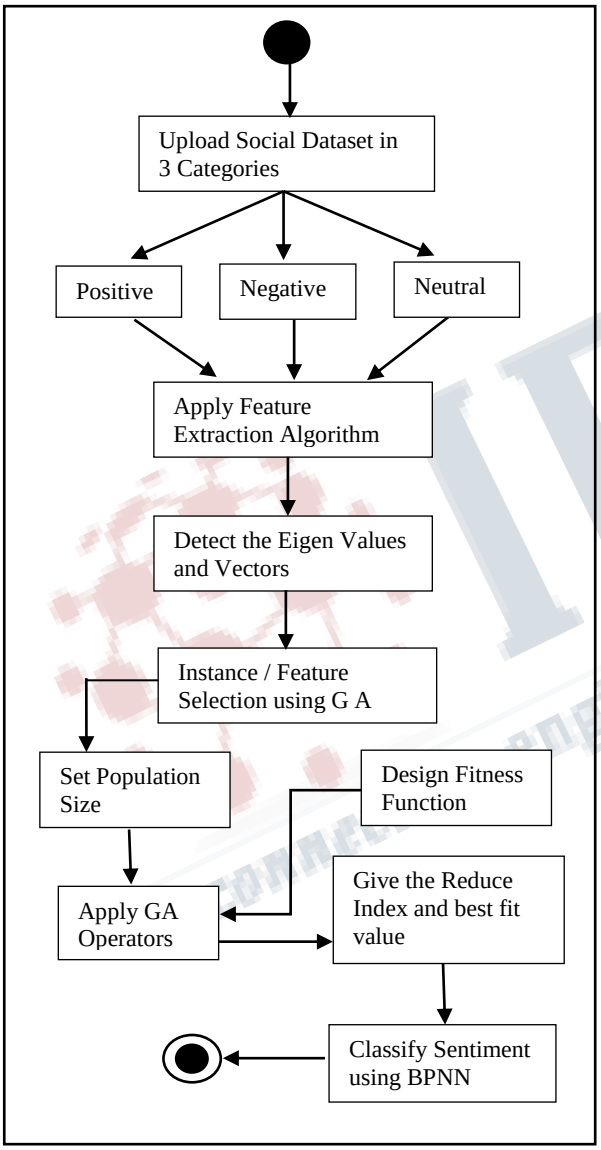

Fig 3. Flow Chart of Sentimental Analysis

VI. RESULTS AND DISCUSSIONS

A. Simulation Environment

To compare the performance of various methodologies, we modeled a Sentimental Analysis and Classification system. The experiment was conducted on a laptop equipped with 2.13GHz, i3 Intel processor, 3Gbyte RAM and 32-bit operating system. As for the simulation tool, MATLAB 2013a has been used. MATLAB stands for MATRIX LABORATORY. It is a high appearance language for technical computing. It consists of a calculation and programming environment. It is an interactive system. It has debug tools, complex data-structures.

The simulated Sentimental Analysis and Classification system consists of data set to categorize the sentiments using PCA algorithm and GA for instance selection and analyze the reviews based on classification approach BPNN.

B. Performance Results

The performance based on matching of various categories in this process. The sentences collected for sentiment detection match with three different data clusters. The maximum matching features with a category defines sentiments. Categorization is evaluated by analyzing various performance parameters like false rejection rate, false acceptance rate and accuracy.

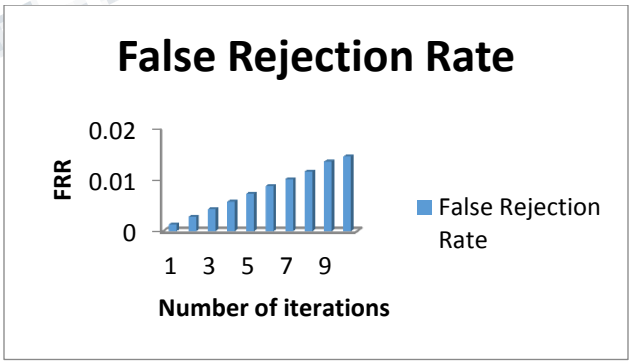

Fig 4. False Rejection Rate

The false rejection rate is used to enhance the performance of system. The FRR in above figure is stable and optimized. As a result accuracy become more than 95%. Less FRR defines the less error in system's classification.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 9, September 2017

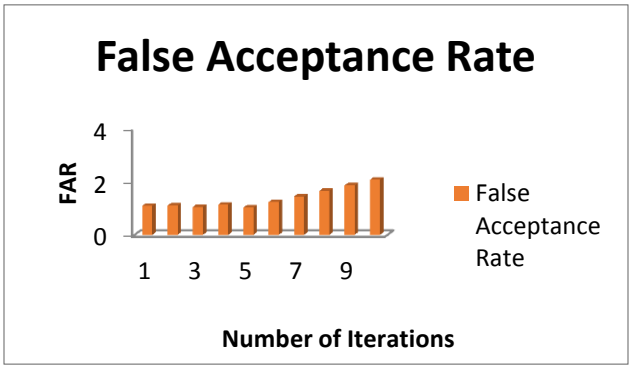

Fig 5. False Acceptance Rate

The false acceptance rate is also used to enhance the performance of system. The FAR in above figure is stable and optimized. As a result accuracy become more than 95%. Less FAR defines the less error in system's classification.

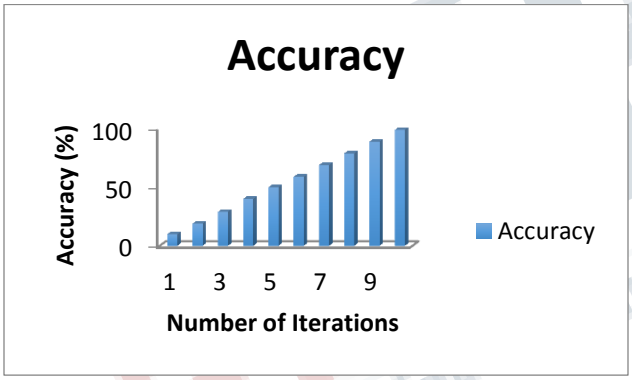

Fig 6. Accuracy

Accuracy defines the working and classification of the system. The performance of proposed work in terms of accuracy is also better than previous approach. This sentimental analysis and classification approach is working with more than 98.5% of accuracy as shown in Fig 6.

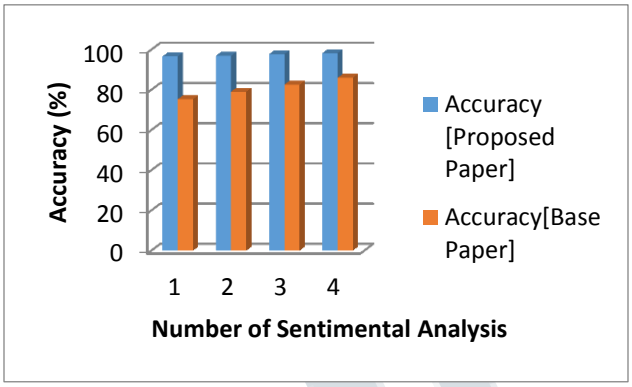

Fig 7. Comparison between Accuracy (Proposed and Existing work)

The comparison between two different algorithms named as base technique and proposed hybrid algorithm shown in the graph in terms of accuracy. The performance of proposed algorithm is better as compare to previous approach. It defines the accurate classification of sentiments as compare of other existing approach. High accuracy in all the cases shows the stable and accurately working of proposed approach.

Number of Category	Accuracy [Proposed Paper]	Accuracy [Base Paper]
1	96.7	75.38
2	97	79.12
3	97.8	82.6
4	98.3	86

Table 1. Comparison Between Accuracy (Proposed and existing work)

Sentimental	Positive	Negative	Neural	ACCURACY
1	0.4805	0.5030	0.4978	0.4938
2	0.7773	0.7513	0.7850	0.7712
3	0.6755	0.6601	0.6623	0.7960
My Approach	0.76	0.780	0.7810	0.9852

Table 2. Comparison Between Accuracy (Proposed and existing work)

VII. CONCLUSION

In conclusion, the novelty of the proposed method is shown by using techniques like genetic algorithm for instance selection and classification of sentiments will be done using back propagation neural network technique. In cross language sentiment classification based on support vector machine, only statistics were used to extract the feature words in the feature selection stage and the classifier could not adapt the target language well. To solve all these issues, new proposed technique will be implemented.

ACKNOWLEDGMENTS

The authors would like to thank the Department of Computer Science & Engineering, Chandigarh Engineering College, Landran for providing outstanding support.

REFERENCES

[1] Abdi, Herve, Lynne J. Williams, and Domininique Valentin. "Multiple factor analysis: principal component analysis for multitable and multiblock data sets." Wiley Interdisciplinary Reviews: Computational Statistics 5, no. 2 (2013): 149-179.

[2] Alessia, D., Fernando Ferri, Patrizia Grifoni, and Tiziana Guzzo. "Approaches, tools and applications for sentiment analysis implementation." International Journal of Computer Applications 125, no. 3 (2015).

[3] Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad, and Fazal Masud Kundi. "A review of feature extraction in sentiment analysis." Journal of Basic and Applied Scientific Research 4, no. 3 (2014): 181-186.

[4] Chatterjee, Arijit, and William Perrizo. "Investor classification and sentiment analysis." In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on, pp. 1177-1180. IEEE, 2016.

[5] David, Omid E., H. Jaap van den Herik, Moshe Koppel, and Nathan S. Netanyahu. "Genetic algorithms for evolving computer chess programs." IEEE Transactions on Evolutionary Computation 18, no. 5 (2014): 779-789.

[6] Hussein, Doaa Mohey El-Din Mohamed. "A survey on sentiment analysis challenges." Journal of King Saud University-Engineering Sciences (2016).

[7] Jandail, Ravendra Ratan Singh. "A proposed Novel Approach for Sentiment Analysis and Opinion Mining." International Journal of UbiComp 5, no. 1/2 (2014): 1

[8] Kiritchenko, Svetlana, Xiaodan Zhu, and Saif M. Mohammad. "Sentiment analysis of short informal texts." Journal of Artificial Intelligence Research 50 (2014): 723-762.

[9] Ma, Hongxia, Yangsen Zhang, and Zhenlei Du. "Cross-language sentiment classification based on Support Vector Machine." In Natural Computation (ICNC), 2015 11th International Conference on, pp. 507-513. IEEE, 2015.

[10] Pontiki, Maria, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. "Semeval-2014 task 4: Aspect based sentiment analysis." Proceedings of SemEval (2014): 27-35.

[11] Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. "Semeval-2015 task 10: Sentiment analysis in twitter." In Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015), pp. 451-463. 2015.

[12] Saduf, Mohd Arif Wani. "Comparative study of back propagation learning algorithms for neural networks." International Journal of Advanced Research in Computer Science and Software Engineering 3, no. 12 (2013).

[13] Sahayak, Varsha, Vijaya Shete, and Apashabi Pathan. "Sentiment Analysis on Twitter Data." International Journal of Innovative Research in Advanced Engineering (IJIRAE) 2, no. 1 (2015): 178-183.

[14] Tripathi, Gautami, and S. Naganna. "Feature selection and classification approach for sentiment analysis." Machine Learning and Applications: An International Journal 2, no. 2 (2015): 1-16

[15] Virmani, Deepali, Vikrant Malhotra, and Ridhi Tyagi. "Sentiment Analysis Using Collaborated Opinion Mining." arXiv preprint arXiv:1401.2618 (2014).