An Artificial Neural Network Based Sentiment Analysis System Using Optimization Technique

Gitanjali[1*], Shailja[2]

[1,2] Department of Computer Science and Engineering, Guru Nanak Dev Engineering College, Punjab, India

Corresponding Author: sharda.gitanjali440@gmail.com, Tel.: 7508018440

Abstract— Social Media is the next logical marketing arena. Currently, Social networking sites dominate the digital marketing space. In the past years, the World Wide Web (WWW) has become a huge source of user-generated content and opinionative data. Using social media, such as Twitter, facebook, etc, user shares their views, feelings in a convenient way. Social media, where millions of people express their views in their daily interaction, provides their sentiments and opinions about particular thing. A lot of work done earlier analyzes the polarity from text but the accuracy of existing technique is not acceptable because of the lack of feature optimization selection. The concept of feature optimization is used to find out the relevant data according to the sentiment classes. By using the concept of feature optimization technique, the chances of removal of irrelevant data is more and we can achieve better accuracy. In the proposed work, a classification technique named as Artificial Neural Network (ANN) along with genetic algorithm, an optimization algorithm will be used and it can train the large amount of dataset that is optimized by using Genetic Algorithm (GA) approach and can be divided into their groups according to the feature for social sentiment database.

Keywords— Data mining, Genetic algorithm, ANN,ARM.

I. INTRODUCTION

Data mining is an analysis process that extracts hidden patterns from available data sources. Instead, big data problems can be better addressed through an enhanced data analysis process. Data mining is not a separate process, but it is an important part of the data knowledge discovery (KDD) process. Data mining includes a set of specific methods and techniques that are specifically designed to extract patterns from raw data. Thus data mining cab be used to analyze the data on the social media. Social media provides a platform for the sharing and witnessing of interactive information. The growth of social media data has interested researchers from various disciplines to uncover the hidden knowledge by applying intelligent data analysis techniques. In this work, data mining technique is used to analyze the Sentiments. Sentiment analysis (or more specifically, sentiment polarity analysis) can be defined as the mapping of text to one of the labels (elements) taken out of a pre-defined set or placing it on the continuum from one end to the other. The elements of the pre-defined set are usually 'negative', 'positive' and 'Neutral'. In this study firstly the genetic algorithm will be used as a feature optimization technique to achieve better accuracy and then Artificial Neural Network will be used as a classifier to train the system for categorizing the polarity of the text.

The main objective of this work will be to optimize the features using genetic algorithm as an optimization algorithm and then to categorize the polarity of the text using Artificial Neural Network and to train the system on the source of the features extracted for the classification and to test the texts.

Rest of the paper is organized as follows, Section II contains the introduction to sentiment analysis, Section III contains the introduction to Genetic Algorithm, Section IV contains the introduction to Artificial Neural Network, Section V contains the brief introduction about Association Rule Mining, Section VI contain the related work for the sentiment analysis of the twitter data set, section VII explains the methodology of proposed work with flow chart, Section VIII describes results and discussions of proposed work, and Section IX concludes the entire work.

II. SENTIMENT ANALYSIS

Sentiment Analysis, also called opinion mining is defined as the process to identify and categorize opinions which are expressed in piece of text especially to determine whether the writer's attitude is positive, negative and neutral. Nowadays sentiments and reviews are becoming much evident due to growing interest in the ecommerce [1]. Basically. it aims to

determine people's sentiment about a topic by analyzing their posts and different actions on social media. Then it classifies the polarity of posts. With the growth of data day by day, importance of sentiment analysis is increasing. Earlier meetings, interviews, question answering sessions were used to gather opinions about various things but now Semantic Analysis has become an effective way of understanding user's opinion about a particular thing, product, object etc [2]. Machines should be much efficient and reliable for interpreting and understanding the emotions and feelings of a human.

The flowchart of sentiment analysis is depicted in figure 1 which gives the general flow of the process of sentiment analysis. As shown in the figure, data for segmentation is taken from the database where the data is stored.
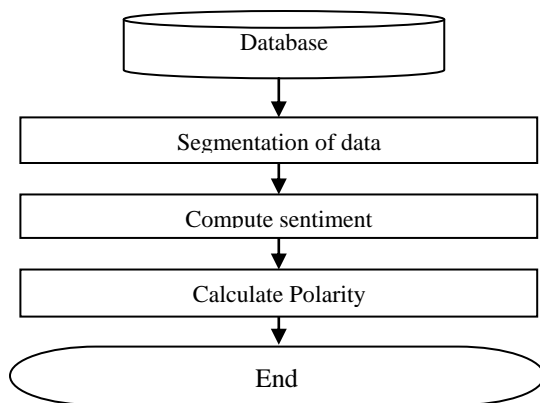

Figure1. Sentiment Analysis

Segmentation can be done according to behaviour and life cycle etc. Then, in sentimental analysis like natural language processing, text analysis has been done. It is basically applied to the voice of the customer like online and social media. In the last step, the data polarity is measured. If the polarity of data is positive, then it is the positive sentence and if polarity is negative it is the negative sentence [3].

III. GENETIC ALGORITHM (GA)

Genetic algorithm (GA) is the set of computational prototypes which are inspired through the evolutionary theory of Darwin. According to the theory of Darwin, the fittest species and the species which can adapt to changing surroundings can outlive; and the rest tends to die. The basic working mechanism of GA is as stated as: Firstly the algorithm is started with a solution set (represented by chromosomes) called as population. From one population, solutions are taken and are used for forming a new population (Reproduction). Basically, this is driven by optimism that the newly generated population will be superior to the old one i.e. old population will be inferior to the generated population. Because of this reason, they are

named as optimistic search algorithm. The reproductive possibilities are distributed in such a way that the chromosomes representing a better solution for the target problem are provided more chances for reproduction than the ones that represent inferior solutions. Search is done through a wide combination of parameters for finding the best match. Basic mechanism of this algorithm is natural selection which has been successfully used in machine learning and optimization problems.

The advantage of GA is its adaptive nature; that is, this algorithm can change/fit itself with respect to the changing environment. The main steps involved in Genetic Algorithm are named as the selection, crossover and mutation GA begins with population. Each individual in population is called as chromosome. The chromosomes progress through iterations, labeled as generations. In every generation, with the help of fitness function, chromosomes are evaluated. Then in next generation, formation of offspring (new chromosomes)are done either by fusion of two chromosomes with the help of crossover operation or by modifying the chromosomes through mutation operation. Then creation of new generation occurs by selection with the help of the fitness value and rejecting the values which do not meet the condition of fitness function. Hence, in this manner the population size remains persistent. In this way, this algorithm provide user with the best chromosomes which represents the optimal solution for the problem [4].
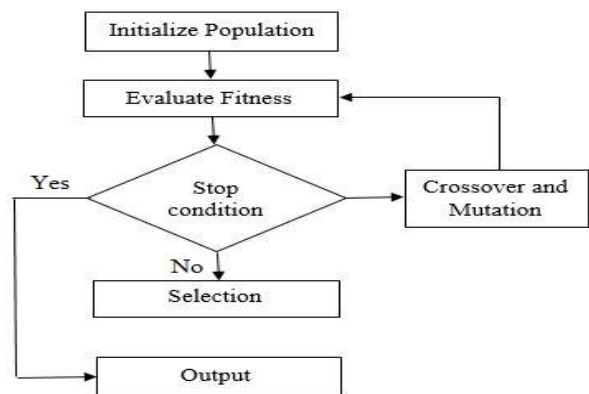

Figure2. Flow Chart of Genetic Algorithm

IV. ARTIFICIAL NEURAL NETWORK (ANN)

Whenever we talk about a neural network, we should more popularly say Artificial Neural Network (ANN), ANN are computers whose architecture is modelled after the brain. They typically consist of hundreds of simple processing units which are wired together in a complex communication network. Each unit or node is a simplified model of real

neuron which sends off a new signal or fires if it receives a sufficiently strong Input signal from the other nodes to which it is connected. ANN is mathematical model or computational model, an information processing paradigm i.e. inspired by the way biological nervous system, such as brain information system. ANN is made up of interconnecting artificial neurons which are programmed like to mimic the properties of m biological neurons. These neurons work in unison to solve specific problems. ANN is configured for solving artificial intelligence problems without creating a model of real biological system. These applications are done through a learning process, like learning in biological system, which involves the adjustment between neurons through synaptic connection. Same happen in the ANN. A simple structure of ANN is shown below:
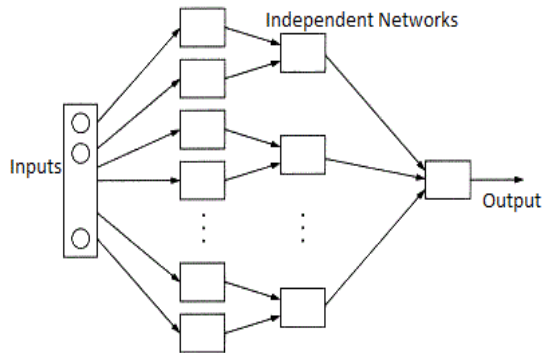

Figure 3. ANN Structure

The input layer consists of input neurons that transmit information to the hidden layer and the data from the hidden layer are transmitted to the output layer. All the neurons used in the process comprises of (i) weighted input, (ii) Activation function (iii) and an output value [5].

V. ASSOCIATION RULE MINING

Association rules are expressions that will help to indicate the likelihood of relationships between data in a wide variety of datasets. As an associate, culture involves the use of machine learning models to analyze data for samples or events in a single database, on a basic level. The combination rules often identify compounds called rules. The merging rule consists of two parts: one (if) and other one is (then). It is a substance in the preset information. Then, it is a predetermined substance.

Association guidelines are often created by using criteria and support for identifying the most important relationships and searching for extreme patterns. Support shows how often products appear in the data. Security shows how many times IF statements are true. The third metric is called elevator,

which can be used to compare confidence with the expected confidence. Association rules are calculated from item sets of two or more items. The rules may contain many rules that contain very little meaning if the rules are to analyze all possible item sets. However, merging rules usually arise from rules that are well-documented.

VI. RELATED WORK

Sharma, et al. [6] presented a web based application that helps to visualize present sentiments that are related to the keywords namely hash-tag, phrase or words respectively. The sentiments as well as the geography intensity have been measured. The code has been developed in "Python" software. The sentiments have been plotted on the real tweet map of the world as well as on USA. Hence the tweets sentiments has been analyzed and plotted on the map as per the location .

Jimenez-Zafra, et al. [7] proposed a technique to detect negation in "Spanish sentiment analysis". An un-supervised polarity classification model has been created which is created on the basis of external knowledge. The performance has been compared in the presence and absence of negation using the tweets in Spanish language. From the experiment it has been analyzed that the accuracy has been enhanced in the presence of negation.

El Alaoui, et al. [8] proposed an adaptable sentiment analysis approach that analyzes social media posts and extracts user's opinion in real-time. The proposed approach consists of first constructing a dynamic dictionary of words' polarity based on a selected set of hash tags related to a given topic, then, classifying the tweets under several classes by introducing new features that strongly fine-tune the polarity degree of a post. The results of prototype tests have performed a good accuracy in detecting positive and negative classes and their sub-classes.

Dos, et al. [9] proposed a novel deep convolution neural network that utilized character to sentence level data to execute short text sentiment analysis. The approach has been applied for two corpora's of two different areas named as Stanford Sentiment Treebank and the Stanford Twitter Sentiment corpus (STS) that comprises of sentences taken from the movies reviews and Twitter messages respectively. The accuracy up to 85.7% has been achieved by using STS corpus.

Neri, et al. [10] studied a sentiment analysis which has been performed over 1000 Facebook posts (news and makes comparison for the sentiments obtained from "Rai" (Italian broadcasting service) and the La7 which is the private company)). The proposed model comprises of different

components (i) crawler used to collect data from internet (ii)Semantic Engine utilized to recognize relevant information in the texts (iii) machine translate engine that is capable to translate search results automatically (iv) classification phase which differentiate search results in the form of clusters and sub clusters.

VII.   PROPOSED WORK

The expanding social networking, web has increased and people have started sharing data through various means. They share their data in the form of reviews, forums, discussions, blogs, and micro-blogs. Emotions are associated with these data which can be positive, negative as well as neutral. Identifying these emotions through face to face communication is easy as compared to textual communication. Sentiment analysis or opinion mining is a text mining technique which is used to determine the opinions which are expressed by the author of the text. We can gain understanding of the attitudes, opinions and emotions expressed with an online mention. So proposed technique of association rule mining with ANN and GA is a better option with the help of which classification accuracy will be much better.

In the work, main concern is on the optimization of feature set. As when feature set will be extracted after the pre-processing step, it also fetches an irrelevant feature set. To get the relevant feature set, the use of optimization technique (Genetic Algorithm) is must. System will not be trained properly if ANN will be provided with the irrelevant feature set. Improper training of the system will further lead to inaccurate classification of the categories of test data. The proposed research work is to classify the sentiments using optimization technique i.e. GA algorithm and to train the system using ANN as a classifier on the basis of the features extracted.

In the proposed work, classification system will be designed to classify the sentiments from text data using ANN as a classifier on the basis of optimized feature set extracted using genetic algorithm as an optimization technique. It can be designed using MATLAB simulator. Firstly, a GUI (Graphical user interface) can be created in MATLAB that will contain two panels namely testing panel and training panel. In training panel, the designed model can be trained using ANN as a classifier. In testing panel, text data can be tested for three different kinds of sentiments i.e. positive, negative and neutral by firstly applying the pre-processing techniques like data cleaning, data transformation, data reduction, tokenization and feature selection. These features are supportive to sentiment analysis for selecting a feature of a particular sentiment.

The features of the signal can be extracted as weights. Then the features of the extracted text can be optimized using genetic algorithm. After that the polarity of the sentiment can be classified using ANN classifier and efficiency can be determined on the basis of performance parameters
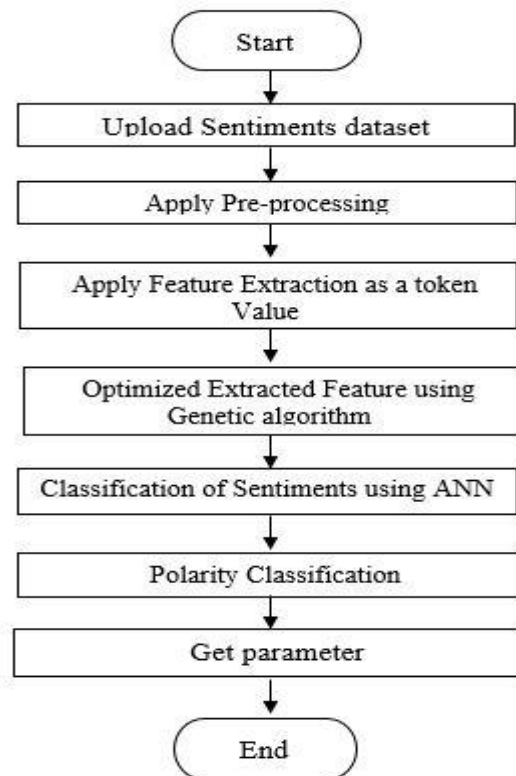

Figure 4. Flowchart of Proposed Work

VIII.   RESULTS AND DISCUSSION

The result which will be obtained from the system, sentiment analysis using ANN as a classifier will contain the sentiments of the given sentences in the form of reviews. Input by the system will be accepted in the form of twitter tweets. Then steps of pre-processing will make the data according to the requirement. Pre-Processing will show impressive positive effects on the quality of feature extraction and results of sentiment analysis. Then with the application of feature extraction technique called as tokenization technique on pre-processed data, feature sets will be extracted. Feature optimization using GA will be done that can find better feature sets which will be given to ANN, hence providing more accurate prediction by its efficient feature selection. Training of system with the ANN classifier will be done. Test data will be classified according to trained structure. Then performance parameters like precision, recall, F-measure and accuracy will be calculated to validate the system.

From the studied literature, the average accuracy determined by the author is nearly 90.21 % [8]. The accuracy which will be obtained using optimized ANN along with the GA will be more than existing work due to the optimization of the extracted features.

IX.   CONCLUSION

In present research works, main focus is on the classification of the text using some classification techniques like ANN, CNN, and SVM etc. The optimization of the feature set is barely done. But in this proposed research work, along with the classification technique named as ANN, optimization Technique (GA) will be used   for the sentiments analysis of the twitter data. A lot of work done earlier analyzes the polarity from text but the accuracy of existing technique is not acceptable because of the lack of feature optimization selection. The concept of feature optimization will be used to find out the relevant data according to the sentiment classes. By using the concept of feature optimization technique, the chances of removal of irrelevant data is more and we will able to achieve better accuracy.

REFERENCES

[1] K. Sarvakar, U. K. Kuchara, "*Sentiment Analysis of movie reviews: A new feature-based sentiment classification*" International Journal of Scientific Research in Computer Science and Engineering, Vol.**6**, Issue.**3**, pp. **8-12** , **2018.**

[2] C. Nanda, M. Dua, "*A Survey on Sentiment Analysis*" International Journal of Scientific Research in Computer Science and Engineering, Vol.**5**, Issue.**2**, pp.**67-70**, **2017.**

[3] A. Palve, R. D. Sonawane, Amol D. Potgantwar, "*Sentiment Analysis of Twitter Streaming Data for Recommendation using Apache Spark*" International Journal of Scientific Research in Network Security and Communication, Vol.**5**, Issue.**3**, pp. **99-103**, **2017.**

[4] S. Jain, "*Mining Big Data using Genetic Algorithm*" International Research Journal of Engineering and Technology (IRJET 2017), Vol.**04**, Issue.**07**, pp.**743-747,2017.**

[5] S. B. Maind, P Wankar, "*Research Paper on Basic of Artificial Neural Network*" International Journal on Recent and Innovation Trends in Computing and Communication, Vol.**2**, Issue.**1**, pp.**96-100, 2014.**

[6] N. Sharma, R. Pabreja, U. Yaqub, V. Atluri, S. A. Chun, and J. Vaidya, "*Web-based application for sentiment analysis of live tweets*," In the Proceeding of 19th International Conference on Digital Government Research Governance in Data Age, , Delft, **Netherlands**, pp. **1–2, 2018**.

[7] S. M. Jimenez-Zafra, M. T. Martin Valdivia, E. Martinez Camara, and L. A. Urena-Lopez, "*Studying the Scope of Negation for Spanish Sentiment Analysis on Twitter*," IEEE Transactions on Affective Computing. , Vol. **14**, Issue. **8**, pp. **1–14, 2017**.

[8] El Alaoui, I., Gahi, Y., Messoussi, R., Chaabi, Y., Todoskoff, A., & Kobi, A, "*A novel adaptable approach for sentiment analysis on big social data*." Journal of Big Data, Vol.**5(1)**, pp.**1-18, 2018**.

[9] C´ and cero N. dos Santos, "*Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts*," 25th International Conference on Computing Linguistic, pp. **69–78, 2014**.

[10] F. Neri, C. Aliprandi, F. Capeci, M. Cuadros, and T. By,

"*Sentiment analysis on social media*," In the proceeding of IEEE/ACM International Conference on Advanced Social Networks Analysis Mining, (ASONAM 2012), Birmingham, **UK** pp. **919–926, 2012**.

Authors Profile

Ms. Gitanjali studied Bachelor of Technology in Computer Science Engineering from RIMT-Institute of Engineering and Technology Mandi Gobindgarh (Punjab) in 2016. She is currently pursuing Master of Technology in Computer Science & engineering from Guru Nanak Dev Engineering College, Ludhiana (Punjab). Her main research focuses on development of sentiment analysis system for twitter data set using genetic algorithm and ANN.

Ms. Shailja received a Bachelor degree in Computer Science Engineering in 2005 and a Master degree in Computer Science & Engineering in 2011. She is with Guru Nanak Dev Engineering College, Ludhiana (Punjab) India as Assistant Professor, in CSE department since 2013- 2019. Her research area is image processing.