International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2020 IJSRCSEIT | Volume 6 | Issue 3 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT206314

Word Embedding and Feature Reduction for Sentiment Analysis Using GA

Prof. Prajakta P. Shelke, Ankita N. Korde
Department of Computer Science and Engineering, Government College of Engineering, Amravati, India

ABSTRACT

Sentiment analysis (SA), also called as opinion mining is the technique for the removal of opinions of a specific entity or feature from reviews dataset. The opinions of other users help in decision making process of people. This paper studies different methods that are aimed at SA. These approaches vary from semantic based methods, machine learning, neural networks, syntactical methods with each having its own strength. Although hybrid approach also exists where the idea is to combine strengths of two or more methods to increase the accuracy. A framework in which sentiment analysis is done by using word embedding and feature reduction techniques is also proposed. Word embedding is a technique in which low-dimensional vector representation of words is provided. Feature reduction method is used with Support Vector Machine (SVM) classifier. The framework will perform sentiment analysis of user opinions by using a machine learning approach and provides a recommendation system for the ease of decision making for users. The proposed system in this paper has solved the scalability problem and improved the accuracy.

Keywords : Sentiment analysis, opinion mining, genetic algorithm, word embedding, feature reduction.

I. INTRODUCTION

The Internet and associated web technologies have dramatically changed the way our society works. Social Networks such as Facebook and Twitter have become commonplace for exchanging ideas, sharing information, promoting business and trade, and promoting products and services. There are various ways to analyse the social media such as collecting business intelligence for products and services, monitoring malicious activities for detecting and mitigating cyber threats, and sentiment analysis for analysing people's feedback and reviews. Sentient analysis, also called as opinion mining, is the process of extracting, identifying, or characterizing the sentiments from text using Natural Language Processing, or Machine Learning (ML) technique. The mostly used approach for Sentiment Analysis (SA) is ML in which the significant dataset is required for training and learning the association between different sentiments. In ML technique, various learning algorithms and datasets are used. Sentiment analysis is mostly discussed in the context of product reviews like whether the product review is positive or negative.

This paper proposes a sentiment analysis which employs SVM algorithm and UCI ML Repository dataset to analyse and classify the reviews. Feature reduction is also done by using by using genetic algorithm (GA) technique which works by developing a fitness function.

Sentiment analysis is a significant research part as people are being more expressive over social networks like Facebook as well as Twitter. Numerous approaches intended at sentiment analysis has been used among which the most common approach is

machine learning (ML) which learns the different sentiments by using a significant dataset. ML methods are used to train the classifiers and determine the sentiments by using different learning algorithms and datasets. Sentiments are generally found in comments, reviews or feedbacks. These sentiments are either positive or negative. In respect to this, sentiment analysis works as a task of classification in which every classified set signifies the sentiment. SA shows the customer satisfaction for a product or an entity.

Word embedding is used in sentiment analysis task which provides low dimensional vector representation of words. Sentiment embedding is used for acquiring both semantic and syntactic similarity between the words which avoids generating similar vector representation for semantically similar words. There are various opinions of users on different products. For example, '*The phone is good but the voice quality is poor'*. In such reviews where both negative and positive sentiments are included in a single sentence, the word embedding technique makes it easy to analyse the sentiments. Feature reduction is an important technique in sentiment analysis which makes the classifiers more accurate and efficient. A there are plenty of features in online reviews, it makes classifier infeasible. So to eliminate the unnecessary features, feature reduction technique is used.

II. RELATED WORK

Here we discuss the protruding related research being carried out in the area of sentiment analysis.

In [5] the author describes the different ways to enable sentiment analysis systems. They focus on methods to fulfil challenges produced by sentiment analysis applications in comparison with the available applications in more traditional systems. They included the problems related to confidentiality, management, and financial influence of opinion leaning facilities.

In [6], they provided a system used for SA of small, social network grades. The system is demonstrated with twitter reviews to show glad and down sentiments as well as demonstrate that the system gives better performance compared to Naïve Bayes. They analysed the twitter reviews and provided a scheme containing data acquisition as well as calculation to analyse the sentiments, which is a scalable system.

In [7], author has attempted a complete summary of last update in field of SA. Many projected algorithms' improvements and several SA applications have been examined and presented concisely. The correlated fields to SA that are concerned by researchers recently are deliberated. They offered almost full image of SA techniques and the correlated fields with brief details and included the refined categorizations of a outsized number of recent articles and the design of the recent development of research in the sentiment analysis and its related zones.

Sentiment analysis of data on Twitter is examined in [8]. They presented POS-specific preceding divergence features. The practice of a tree kernel on the way to prevent requirement of deadly feature manufacturing is discovered. The novel features and tree kernel accomplished nearly with a similar way, together outstripping an advanced reference point.

In [9], usefulness of etymological structures for identifying sentiment in Twitter communications is studied. It estimated worth of current lexical assets and also features which seizure data around creative as well as casual language applied with microblogging. They acquired the supervised method for problem, nonetheless for construction of training data which influenced current hashtags in the Twitter data.
In [10], word embedding aims to the vector illustration of arguments through leveraging the appropriate data with huge reviews datasets is stated.

A pioneering work is proposed and castoff the neural network language system towards study of word embedding built with a prior context of every word [11]. For eluding the production of analogous vector representations of sentimentally differing words, current research ought to propose embedding [12] to capture both semantic and syntactic data so that sentimentally related words possess comparable symbols [13].

III. PROPOSED FRAMEWORK

In this system, we present a sentiment analysis framework which includes data cleansing, data pre-processing and sentiment analysis. For better recommendation system, a summarized review is necessary to the customer for decision making ability. The proposed framework illustrates the review summarization and allows users to make a quick decision.

The proposed system consists of different stages to carry out the internal working of system. In this framework, different processes take place such as data cleansing, data pre-processing, population initialization, feature generation, feature selection using GA, word embedding and sentiment analysis. Figure 1 describes the whole sentiment analysis framework.

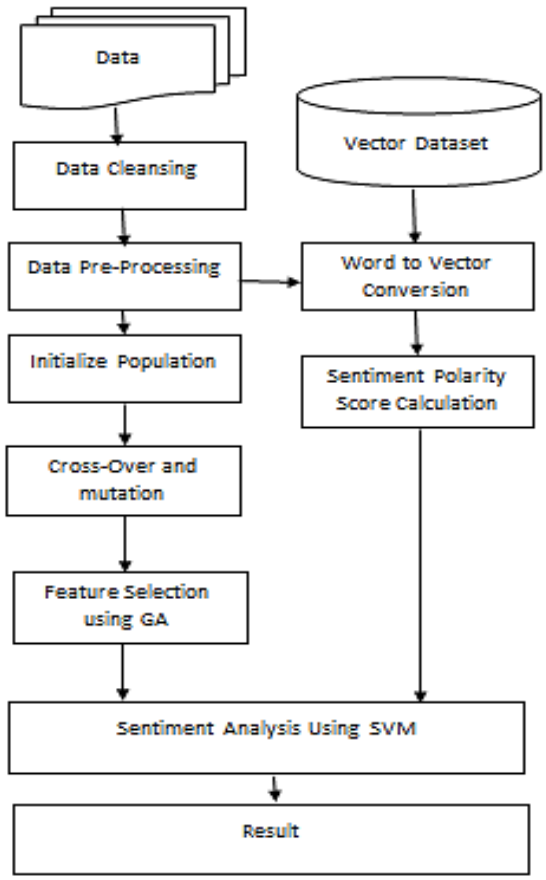

Figure 1: Sentiment Analysis Framework

3.1 DATA CLEANSING

Data cleansing is the first module in the proposed system. In this step, extracted data is streamed from the files and saved for cleansing. This stage includes three sub-stages.

1) GARBAGE REMOVAL

In this step, unwanted characters such as URLs, web addresses, and online links are removed from the text.

2) SLANG CORRECTION

In this step, any slang and abbreviated word that is used in online conversations are corrected. Predefined dictionaries and maps are used to translate slangs or abbreviation to their original and abbreviated form. e.g., "*ttyl*" to "talk to you later". This is cooperative for far ahead stages as, throughout sentiment analysis, the abbreviated words make no sense for analysis engine.

3) STOPWORD REMOVAL

This step is used to remove common words in the text such as “a”, “an”, “the”, “about”, etc. Such words are of no use in NLP.

3.2. DATA PREPROCESSING

Pre-processing includes various NLP tasks such as tokenization, word stemming, and part-of-speech tagging.

1) TOKENIZATION

Tokenization is the method of breaking a stream of text into words, phrases, symbols, or other meaningful features called tokens. In order to tokenize the text, LingPipeTokenizer from Apache Lucene package is used which preserves punctuations. An important point to remark is that custom data structures are planned to grip tokens (Keyword) and sentences (list of Keywords) of each file.

2) STEMMING

In stemming, the infected words are reduced to its root word. The porter-2 algorithm is used to stem the infected words.

3) POS-TAGGING

In POS tagging, the word in a text is tagged as a particular part of speech, based on its definition and context. After pre-processing, the data is sent to the next module in the framework.

3.3 SENTIMENT ANALYSIS

This is the most vital module of the framework. In this step, each sentence is given to the sentiment analysis engine and it calculates the sentiment polarity score. ML algorithms are used to classify sentiment values of the given text. In order to solve scalability problem, this system provides an efficient technique to reduce the feature set size. To find the sentiment polarity score, the score of all the keywords is aggregated on a document level to find the global score and sentiment value of either positive or negative is assigned.

3.4 WORD EMBEDDING

In this step, the words from the text are converted to its vector form. This word to vector conversion is done by using the Word2Vec dataset. The words that are useful for sentiment analysis are extracted from the text and converted to its vector form by extracting the vector value of words from Word2Vec database. The aim of word embedding is to have words with similar context occupy close spatial positions.

3.5 GENETIC ALGORITHM

A genetic algorithm is a search and an optimized feature selection algorithm which integrates with ensemble methods to improve the performance and overcome the limitations of traditional method. An optimization is a process of finding the best or an optimal solution for a sentiment classification. Genetic Algorithm is basically used as a problem solving strategy in order to provide with an optimal solution. GA is carried out in three steps.

1. Population Initialization

In GA, the initial populations of n strings are randomly generated and collection of such strings is called initial population. The information gain feature weights are used as the final strings in the initial population. The information gain solution features are used as the solution string in the initial population. The solution features are represented using binary string character. Specifically 1 represents a selected attribute or feature and 0 represents the discarded one. Generate random population of individual. Each attribute is switched on with the probability Pi.

2. Feature Selection

As the average fitness of the population increases, the strength of the selective pressure also increases and the fitness function becomes more discriminating. This method can be helpful in making the best selection

later on when all individuals have relatively high fitness and only small differences in fitness distinguish one from another.

3. Crossover

Once the individuals have been selected the next thing is to produce the offspring. Crossover is the process of exchange of information between two parents to produce a new offspring. Choose two individuals from the population and perform crossover based on a crossover probability. We use uniform crossover by selecting two individuals and swapping substring at a randomly determined crossover point x. If the mixing ratio is 0.5, then half of the genes in the offspring will come from parent 1 and half will come from parent 2. Mutation is randomly mutated individual feature characters in a solution string based on a fixed probability

3.6 SVM CLASSIFICATION

A support vector machine (SVM) is a supervised machine learning algorithm that uses classification algorithms for two-group classification problems. After giving SVM algorithm sets of labelled training data for each category, we are able to categorize new text. A support vector machine takes the data points and outputs the hyperplane that best separates the features. This line is the decision boundary, anything falls to one side of it we will classify as positive, and anything that falls to the other as negative sentiments.

IV. RESULTS

This section includes our results and discussion. The result shows the overall performance of the system for classification of sentiments. Here we use the parameters Quality, Service, Price and Miscellaneous. The different performance metrics are considered such as precision, recall, F-measure, and execution time.

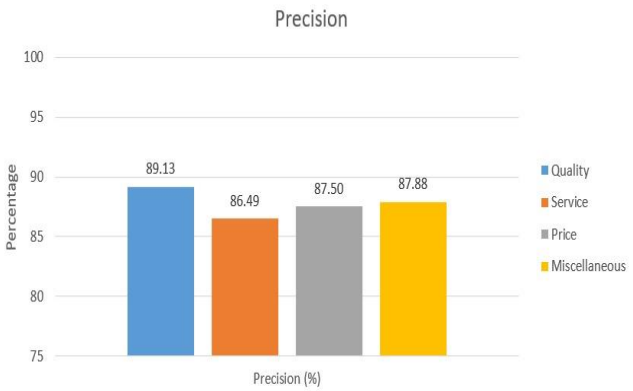

Figure 2: Precision

The precision for different parameters such as quality, service, price, and miscellaneous is calculated. Figure 2 shows that the quality has the highest precision.

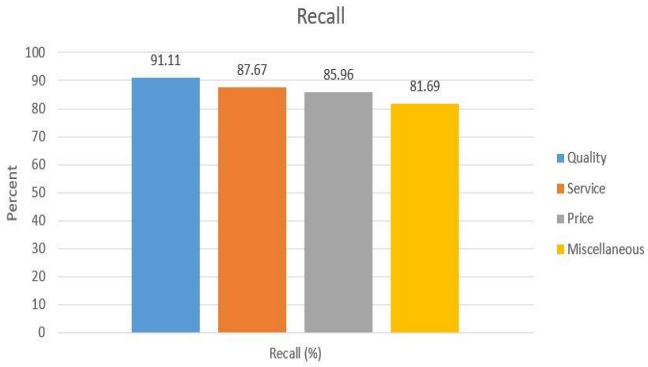

Figure 3: Recall

The recall factor is calculated for all the parameters. Figure 3 shows that the Quality has highest recall.

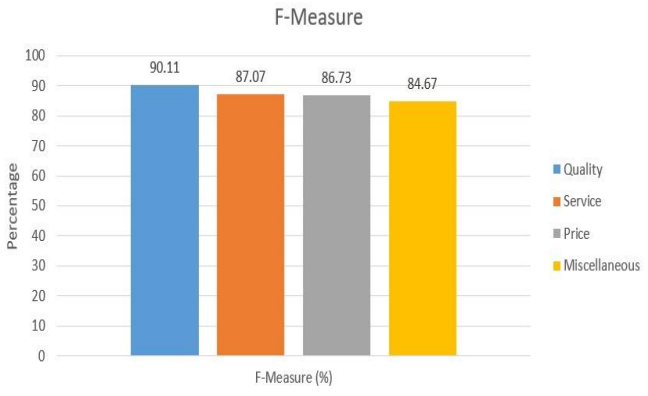

Figure 4: F-Measure

F-Measure is calculated for all the parameters. Figure 4 shows that the highest F-measure is found for Quality parameter, which is 90.11.

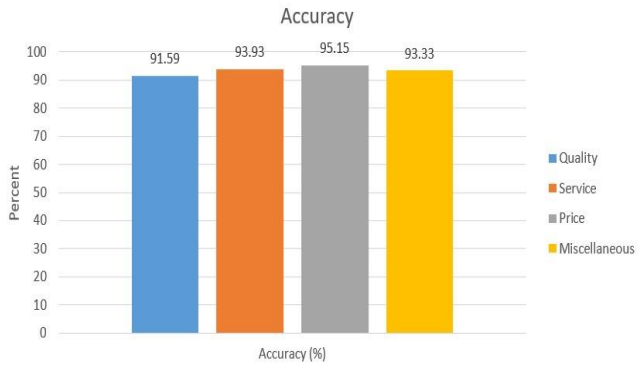

Figure 5: Accuracy

From Figure 5, we can see that the highest accuracy metric is of price. The accuracy for price is 95.15.

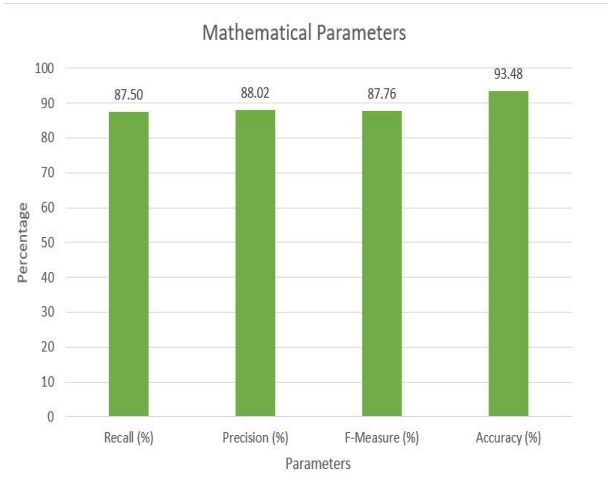

Figure 6: Mathematical Parameters

Figure 6 shows the overall performance of all the mathematical parameters including Recall, Precision, F-Measure, and Accuracy.

V. CONCLUSION

In this paper, we have presented the design, development, and evaluation of our sentiment analysis framework in detail. We proposed and developed a model for feature selection using GA's evolutionary model. This model resulted in reduced feature size and increased efficiency without compromising in accuracy.

We conclude that the proposed framework has proved to be an excessive addition in sentiment analysis techniques. By means of additional aids of GA based optimization, it reduces feature size and improves efficiency while maintaining the accuracy. This framework provides a better recommendation system. In the future, we intend to outspread this framework for hotels reviews system.

VI. REFERENCES

[1]. Farkhund Iqbal,Jahanzeb Maqbool Hashmi and Benjamin C. Fung, "A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm based Feature Reduction",2017.

[2]. M. Pontiki et al., "SemEval-2016 task 5: Aspect based sentiment analysis," in Proc. 8th Int. Workshop Semantic Eval. (SemEval), 2014.

[3]. P. C. S. Njølstad, L. S. Høysæter, W. Wei, and J. A. Gulla, "Evaluating feature sets and classifiers for sentiment analysis of financial news," in Proc. IEEE/WIC/ACM Int. Joint Conf. Web Intell. (WI) Intell. Agent Technol. (IAT), vol. 2, Aug. 2014.

[4]. M. Govindarajan, "Sentiment analysis of movie reviews using hybrid method of naive Bayes and genetic algorithm," Int. J. Adv. Comput. Res., vol. 3, no. 4, 2013.

[5]. B. Pang and L. Lee,"Opinion mining and sentiment analysis',' Found. Trends Inf. Retr., vol. 2, 2008.

[6]. A. Davies and Z. Ghahramani, "Language-independent Bayesian sentiment mining of Twitter," in Proc. Workshop Social Netw. Mining Anal., 2011.

[7]. W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, 2014.

[8]. A. Agarwal, B. Xie, I.Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," in Proc. Workshop Lang. Social Media, 2011.

[9]. E. Kouloumpis, T.Wilson, and J. Moore, “Twitter sentiment analysis: The good the bad and the OMG!'' in Proc. ICWSM, vol. 11. 2011.

[10]. Ling-Chih Yu, J. Wang, K. Robert and X. Zhang,"Refining Word Embeddings using intensity scores for Sentiment Analysis", Vol. 26, March 2018

[11]. Y. Beingo, R. Ducharme,"A Neural Probabilistic Language Model", J. Mach Learn Res., Vol. 3.

[12]. A. L. Maas, R.E.Daly,"Learning Word Vectors for Sentiment Analysis". In Proc. ACL, 2011.

[13]. Y. Ren, Y. Zhang and D. Ji,"Improving Twitter Sentiment Classification using Topic-enriched Multi prototype Word Embedding", in Proc. AAAI, 2016.

[14]. A. Collomb, C. Costea, D. Joyeux, O. Hasan, and L. Brunie, "A study and comparison of sentiment analysis methods for reputation evaluation,'' Tech. Rep. RR-LIRIS-2014-002, 2014.

[15]. L. M. Schmitt, "Theory of genetic algorithms,'' Theor. Comput. Sci., vol. 259, May 2001.

[16]. A. Pak and P. Paroubek, ``Twitter as a corpus for sentiment analysis and opinion mining,'' in Proc. LREC, 2010.

Cite this article as :

Prof. Prajakta P. Shelke, Ankita N. Korde, "Word Embedding and Feature Reduction for Sentiment Analysis Using GA", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 6, Issue 3, pp.111-117, May-June-2020.
Available at
doi : https://doi.org/10.32628/CSEIT206314
Journal URL : http://ijsrcseit.com/CSEIT206314