International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol. 2, Issue. 5, Sep.-Oct. 2012 pp-3265-3270 ISSN: 2249-6645

Mining Two Class Opinions Using Optimized Recurrent Neural Network

J. Isabella, R. M. Suresh,
Research Scholar,Sathyabama U/niversity,Chennai,India
Principal,Jerusalem College of Engineering,Chennai,India

Abstract: *Opinion mining automatically classifies the sentiment of the reviews of the customers into positive or negative. Sentiment classification helps analyzing the customer's opinions from the information gathered online. Opinions of products and services are available at various sources such as feedback at websites, blogs in the internet. Automatic opinion mining is required as it is not possible to manually classify the amount of review available online. In this paper, it is proposed to classify movie reviews using a novel Recurrent Neural Network with genetic algorithm optimization.*

Keywords: *Opinion Mining, Sentiment Classification, Neural Network, Genetic algorithm*

I. Introduction

With the emergence of Internet, number of people participating in posting opinions, feedbacks has drastically increased. These reviews are very useful for users to take decisions. Consumers tend to decide whether to buy a product or not based on the opinion. Reviews of products, movie, books, events, and political issues and so on, are available in abundance but it is not possible to go through all the content. So, opinion mining is used for creation and automatically updated review websites, which summarizes opinions on the whole of a particular product or event. Sentiment classification classifies the reviews into positive or negative opinions [1]. An individual finds it useful to see summary of opinions of existing users before making an informed decision. User can also compare with that of competing products. Reviews of movies, books, music are other oft-looked up issues.

Opinion mining uses information retrieval techniques for the process of searching, retrieving relevant documents for the information required. For efficient information retrieval, documents are transformed into a standard format and indexed. Stopword list is prepared which lists the words that are not relevant for retrieving documents. Common words are reduced to its stem or root word. Term frequency is computed, which expresses the number of times a term occurs in a document. Relative term frequency is obtained by term frequency to the total number of occurrence of all terms in the document. Similar documents have similar relative frequencies. Thus, using a similarity measure like cosine measure, documents similar to the query is retrieved. With the advance in machine learning, opinions and reviews of various products and services are semi-automatically or automatically retrieved and further classified. Several works in literature deal with mining reviews of automobiles, banks, movies, travel destinations, electronics and mobile devices [2].

Movie review mining has some special challenges when compared with product review mining. As movie review mining is very domain specific and word semantics in review could contradict with overall sentiment polarity (good or bad) of that review [3]. For example, an "unpredictable" service gives negative meaning to product review, whereas for a movie with "unpredictable" storyline gives positive opinion to moviegoers. Machine learning methods and semantic orientation methods are the main approaches used for sentiment classification. Bo Pang et al., [4] investigated the effectiveness of classification of documents by overall sentiment using machine learning techniques. Experiments showed that the machine learning techniques give better result than human produced baseline for sentiment analysis on movie review data. In this paper, it is proposed to classify movie reviews using Recurrent neural network with genetic algorithm.

II. Previous Research

Goldberg, et al., [5] presented a graph-based semi-supervised learning algorithm for rating the movie reviews. In the proposed method, the problem of the scarcity of labeled movie reviews to infer rating is addressed. A graph-based semi-supervised learning is proposed for learning from a set of movie reviews with ratings. The unlabeled documents are rated on the basis of the sentiment expressed in the review. This was done by creating a graph on both labeled and unlabeled data to encode certain assumptions for rating the reviews. The optimization problem is then solved to obtain a smooth rating function over the whole graph. Experimental results show that the proposed method achieved better predictive accuracy when compared with other methods with limited labeled data.

Dave, et al., [6] presented a classifier for sentiment classification of product reviews. Information retrieval techniques were used for feature extraction and scoring of words. Product reviews are obtained from the sites or clipping services. Structured reviews are used for training and testing to determine sentiment. The classifier is then used to identify and classify review sentences. Noise and ambiguity limits the performance, so sentences are grouped into attributes for better results. Experiments show that the method works as well as or better than traditional machine learning.

Ye, et al., [7] proposed a semantic approach for sentiment classification of movie reviews in Chinese. Word segmentation was introduced for classification process and also an optimal reference-word-pair selecting process for determining extremely positive and negative opinions. A semantic orientation (SO) value is calculated for the phrases. The average SO value of the review is then calculated to classify it as positive if the value exceeds the threshold and negative otherwise. Results show that the proposed method is comparable with the figures of other movie classification studies.

Zhuang, et al., [8] proposed an approach for movie review mining and summarization by integrating WordNet, and statistical analysis. In the proposed approach, a key word list is generated based on WordNet, movie and labeled training data to find features and opinions. Polarity of the opinion word is determined. Suitable feature-opinion pairs are identified by applying grammatical rules. Summaries of the reviews are generated using the extracted feature-opinion pairs. IMDB dataset was used for experimentation and the results showed the proposed method performed better than the other review mining algorithms.

Inferences made by Pang et al., [4] are that machine learning techniques are better than human baselines for sentiment classification. Whereas the accuracy achieved in sentiment classification is much lower when compared to topic based categorization. The experimental setup consists of movie-review corpus with randomly selected 700 positive sentiment and 700 negative sentiment reviews. Features based on unigrams and bigrams are used for classification. Learning methods Naïve Bayes, maximum entropy classification and support vector machines were employed. Machine learning algorithms using unigrams got better accuracy than human selected unigram. Whereas in topic based classification using bag of unigram features achieve 90% accuracy suggesting that sentiment categorization is more difficult than topic categorization.

Peter Turney [9] proposes an unsupervised learning algorithm, using semantic orientation of the phrases containing adjectives and adverbs, to classify reviews. The approach initially extracts phrases containing adjectives and adverbs; the semantic orientation of the phrase is estimated using PMI-IR; based on the average semantic orientation the phrases the review is classified as recommended (Thumbs up) or not recommended (Thumbs down). Experiment was conducted using 410 reviews on various topics; an average accuracy of 74% was achieved. Movie reviews had accuracy of 66% whereas banks and automobiles attained 80% to 84% accuracy. The advantage of this algorithm is its simplicity of using PMI-IR. The time required for queries is the main disadvantage of this method.

III. Research Method

Bo Pang and Lillian Lee have created a collection of movie-review documents from the Internet Movie Database (IMDb) archives. It is an online database of information related to movies, television shows, actors. In this paper, it is proposed to use online movie reviews as data. The dataset is labeled with respect to their overall sentiment polarity (positive or negative) or subjective rating (e.g., two stars). List of stop words for commonly occurring words and stemming words with similar context is prepared. A total of two hundred reviews with 100 positive and 100 negative are chosen in this work.

3.1 Recurrent Neural Network (RNN)

Neural networks are an artificial representation of human brain, artificial neurons use mathematical or computational mode for information processing. Neurons exhibit complex global behavior due to the connections between the neurons and its parameters. Neural networks are adaptive in nature, changing according to the information flow through the network. Neural networks learn by examples, during learning the connections between the neurons are adjusted.

The neural network is made up of layers; input layer, hidden layer and output layer. The connections between the neurons between layers are referred to as weights. The inputs are multiplied by the weights and then fed through a transfer activation function to generate an output. During training, the weights of each neuron are so adjusted to reduce the error between the desired output and actual output. Back propagation algorithm [10] is the most popular learning algorithm used to train the neural networks.

Recurrent neural networks (RNNs) are dynamical systems used for both classification and prediction using the temporal information of the inputs. The interrelations between the inputs and the internal state are exploited to produce output during training of RNNs. Thus, the outputs represent the relevant internal states of the past [11]. During training process, a second source of information helps constitute the target values which specify the relevant interrelations in the input sequence. Inputs for RNNs are generally in a time series and the target is a time series or a sequence of constant value. RNNs when used for classification especially in language learning, the output are given in form of a constant class label whereas in prediction tasks the output consists of time series.

The neurons activations in the RNN depend on the input series and also the internal dynamics of the RNN. Thus, on training of the weights, the output neurons can classify an input sequence. Generally, a time series prediction task is included in the training as it improves the classification ability. This also helps to extract the best features from the input series. Topology of an extended Elman-network [11] is illustrated in the figure 1. The network consists of two parts; a feed-forward part and a memory part. The memory part stores the activation of the neurons of the previous cycle and forwards it as additional inputs.

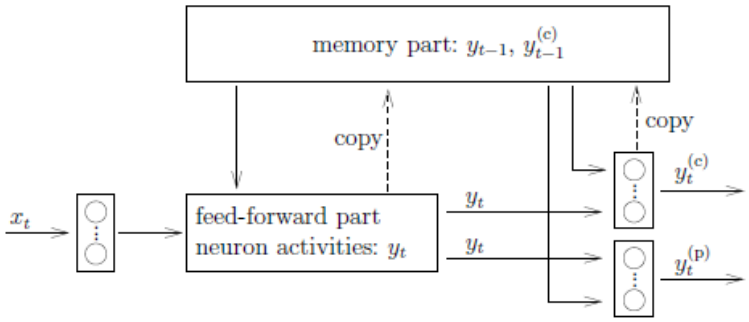

Figure 1: Topology of an extended Elman-network

The induced local field y_k of the neuron is given by equations as follows:

$$y_k = b_k + \sum_i \sum_j w_{kij} x_i u_j$$

$$x_k(n+1) = \varphi(v_k(n))$$

$$= \frac{1}{1+\exp(-v_k(n))}$$

where b_k is the associated weight

$x_k(n)$ is the state of neuron k,

$u_j(n)$ is the input applied to source j

w_{kij} is the weight of neuron k

For learning in simple recurrent networks, error at state layer $\delta_j(t)$, is used to modify weights and errors are backpropagated according to equation 5.13

$$\delta_{ij}(t-1) = \sum_h^m \delta_{ih}(t) u_{hj} f'(y_{ij}(t-1))$$

Where *h* is the index for the activation receiving node and *j* for the sending node (one time step back). This allows calculation of the error as assessed at time *t*, for node outputs (at the state or input layer).

Though backpropagation is well suited for simple training problems, with increase in complexity the performance is reduced significantly. This is due to the fact that gradient search techniques get trapped at local minima. To overcome this problem, it is proposed to use genetic algorithm optimization for learning of neural network.

3.2 Genetic Algorithm

Genetic algorithms are nature-inspired optimization methods that could be advantageously used for optimization problems. GAs imitates basic principles of life and applies genetic operators like mutation, crossover, or selection to a sequence of alleles which is the equivalent of a chromosome in nature and is built by a representation which provides a string of symbols to every possible solution of the optimization problem. Earlier work revealed that behavior and performance of GEAs was strongly influenced by the representation used. Due to this many recommendations for a proper design of representations was made over the years. But most design rules are of a qualitative nature and are not very helpful for estimating how different representation types influence problem difficulty.

Genetic algorithms (GA) are algorithms used for optimization and learning. In GA, the solutions are encoded on chromosomes. Fitness function is used to evaluate the chromosomes. A population of chromosomes or solutions is initiated, and GA operators such as reproduction, mutation, and crossover are applied to generate the next generation of chromosomes. The population reproduces until stopping criteria is met or for a specified number of iterations. During each iteration, one or more parents are chosen to reproduce. The selection of parents depends on the fitness values, the chromosomes with high fitness value are chosen often to reproduce. Children are produced and inserted into the population. Thus, GA produces population of better and better solutions, converging towards global optimum.

The pseudo-algorithm of GA is as follows:
BEGIN
1. Generation t=0;
2. Initialize the population P(t);
3. While not Termination criterion do
a. Evaluate the population P(t);
b. Next P'(t) (by crossover/mutation) over P(t);
c. Evaluate P'(t);
d. P(t+1) = select from P'(t);
e. Generation t=t+1;
4. End while.

IV. Experimental Setup

It is proposed to investigate the effectiveness of GA to evolve the number of neurons in the hidden layer of the Neural Network architecture. The GA architecture used to optimize the number of hidden neurons is shown in Table 1.

Table 1: Parameters used in the GA optimization

Number of epoch	200
population size	10
Maximum generations	5
Neuron optimization lower bound	10
Neuron optimization lower bound	30
Encoder mechanism	Roulette
Cross over type cross over probability	Two point 0.9
Mutation Mutation probability	Uniform 0.01

The parameters used in the experiment are shown in Table 2. The table gives the details of the number of input neurons, number of neurons in the hidden layer and the momentum.

Table 2: The parameters used in proposed neural network design

Number of neurons in input layer	48
Number of neurons in output layer	2
Number of hidden layer	1
Context unit time constant	0.8 second
Transfer function of context unit	Integrator
Number of neurons in hidden layer	10, 15, 20, 25
Learning rule	Back propogation
Number of epochs for termination	500

V. Results

The classification accuracy obtained is compared with existing feed forward network and recurrent neural network. Figure 2 shows the classification accuracy of the three neural network techniques.

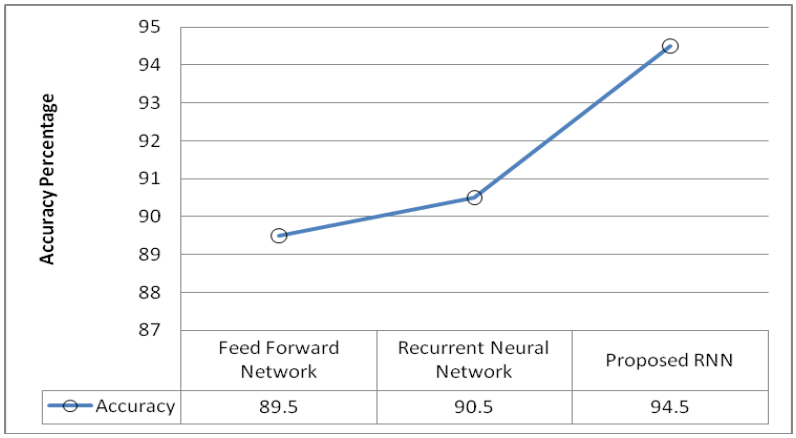

Figure 2: The classification accuracy compared with different classifiers

The plot of Avg MSE vs Epoch (number of iterations) for various number of hidden neurons is shown in Figure 3.

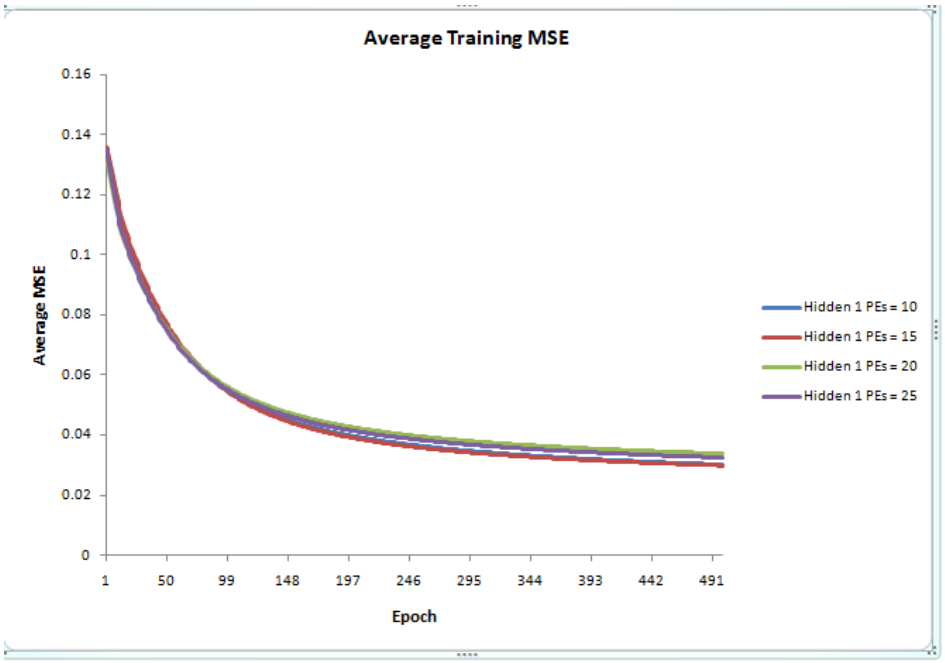

Figure 3: Avg MSE versus No.of Epoch

The best fitness obtained and average fitness obtained is shown in Table 3. Figure 3 shows the plot of MSE with the generation.

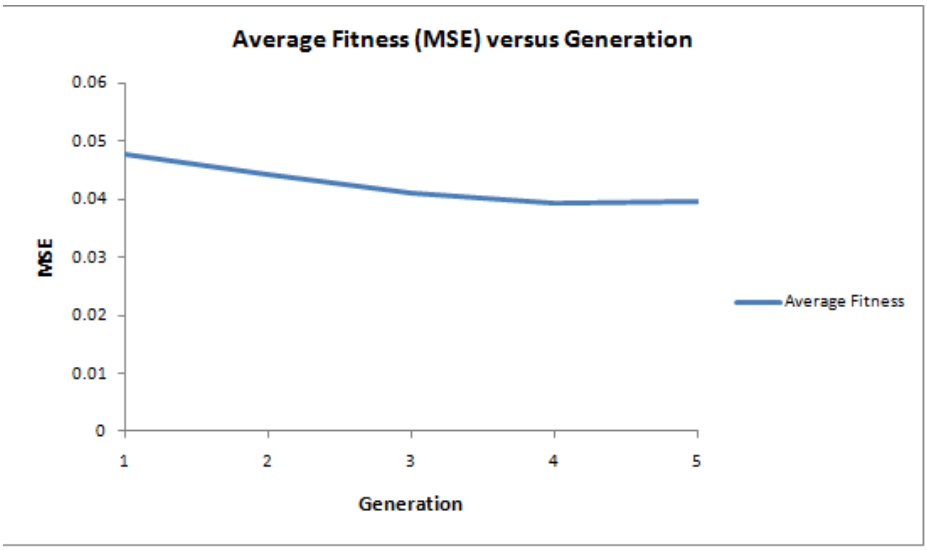

Figure 4: The average fitness value obtained

Table 3: Fitness values obtained

Best Fitness	**Average Fitness**
0.04174498	0.04770372
0.03654528	0.0443747
0.03654528	0.04110919
0.03654528	0.03932239
0.03654528	0.03977443

VI. Summary and Concluding Remarks

In this work an improved Recurrent Neural Network with Genetic Optimization was proposed. Features were extracted from the unstructured movie review data and feature reduction was done based on correlation between features. The proposed classifier algorithm shows an improvement of 4.42 % over conventional Recurrent Neural Network algorithm. Using genetic optimization, the ideal number of neurons in the hidden layer was optimized.

References

[1] Cody W, Kreulen J. T, Krishna V, Spangler S W, (2002) "The Integration of Business Intelligence and Knowledge Management", *IBM Systems Journal*, Vol. 41, No. 4, pp. 697-713, 2002.

[2] Hong Yu and Vasileios Hatzivassiloglou. "Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences". In Michael Collins and Mark Steedman, editors, Proceedings of EMNLP-03, 8th Conference on Empirical Methods in Natural Language Processing, pages 129–136, 2003.

[3] D. Houser and J. Wooders. Reputation in Auctions: Theory and Evidence from eBay. *Journal of Economics and Management Strategy*, 15:353-369, 2006.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.

[5] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in TextGraphs: HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing, 2006.

[6] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in Proceedings of WWW, pp. 519–528, 2003.

[7] Ye, W. Shi and Y. Li. 2006. Sentiment classification for movie reviews in Chinese by improved semantic oriented approach. In Proceedings of 39th Hawaii International Conference on System Sciences, 2006.

[8] L. Zhuang, F. Jing, X.-Y. Zhu, and L. Zhang, "Movie review mining and summarization," in Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 2006.

[9] Peter Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417–424, 2002.

[10] D.E. Rumelhart, G.E. Hinton and R.J. Willams, "Learning Representations by BackPropagating Errors,"' Nature 323, pp. 533-536 (1986).

[11] J. L. Elman. Finding structure in time. Cognitive Science, 14:179{211, 1990.