Sentiment classification of social media reviews using an ensemble classifier

Savita Sangam, Subhash Shinde
University of Mumbai, India

ABSTRACT
These days it has become a common practice for business organizations and individuals to make use of social media for sharing the opinions about the products or the services. Consumers are also ready to share their views on certain products or commodities. Thus huge amount of unstructured social media data gets generated day by day. Gradually heap of text data will be formed in many areas like automated business, education, health care, and show business and so on. Opinion mining also referred as sentiment analysis or sentiment classification, deals with mining of the review text and classifying the opinions or the sentiments of that text as positive or negative. In this paper we propose an ensemble classifier model consisting of Support Vector Machine and Artificial Neural Network. It combines the knowledge from two feature sets for sentiment classification. The proposed model shows the acceptable performance in terms of accuracy when compared with the baseline model.

Keywords: ANN, Ensemble model, Feature selection, Opinion mining, Reviews, Sentiment classification, SVM

1. INTRODUCTION
The recent trend shows that entrepreneurs, artists, companies, agencies and individuals are interested in knowing the public opinions about their brands, products, services, commodities etc. The review data produced by the social media applications may not be in a proper structure and may require lot of processing in order to make it usable. In order to process these reviews, data models need to be constructed. The focus of this research work is to process and analyze the opinions or the sentiments of the social media reviews by applying some data mining techniques. According to Liu [1], the study of analyzing opinions in written language, is termed as opinion mining. According to Bo Pang and Lillian Lee [2, 3] “the task of analyzing the opinion, sentiment, and subjectivity computationally is known as Opinion Mining” and it is also called as Sentiment Analysis (SA). Decision makers rely on SA for making their decisions. For example various shopping sites like Amazon, Flipkart etc. take feedback from the customers that will help them to take proper decisions for improving the quality of their services and marketing strategies. SA techniques have been applied widely in many areas like business, entertainment, medicine, politics etc. Sentiment Classification (SC) process classifies the sentiments of the text reviews into negative or positive or sometimes neutral. The two main approaches for sentiment classification are: Lexicon Based approach and Machine Learning approach. In Lexicon based approach, a sentiment score is calculated using a dictionary of positive and negative words with a positive or negative sentiment value assigned to each of the words. The overall sentiment of the entire text passage is sum or average (or any other function) of all the words. This approach is domain specific and gives low recall. The Machine Learning algorithm uses labeled data sets in order to perform the classification task. The classifier gets trained on training data in the form of features which are the words or phrases in the text. It then classifies the unseen test data based on its training. There

are three types of machine learning techniques namely supervised, unsupervised and semi-supervised. Naïve Bayes (NB), Decision Tree, Support Vector Machine (SVM), and Maximum Entropy (MAXENT) are some of the machine learning systems mentioned by the researchers for the sentiment classification work. Different feature selection mechanisms to select the features in the text, and deeper analysis of the sentences as a whole are the main points that are to be considered for the accurate sentiment classification. Machine learning methods basically rely on the features with which they are trained to do the classification task. Along with the traditional methods, deep learning methods are now catching attraction for many tasks including sentiment analysis [4, 5].

In this paper we have used Most Persistent Feature Selection (MPFS) method (a novel feature selection method proposed in our previous work [6]) and developed a baseline classifier models using NB, MAXENT and SVM. The classifiers are trained on bigrams as well as trigrams features. The feature set is further optimized using a technique based on Genetic Algorithm (GA) to produce an optimized feature set. An ensemble classifier model is proposed which includes SVM and Artificial Neural Network (ANN) classifiers. The performance of the proposed model is tested using 10-fold cross validation technique and its accuracy is compared with our baseline classifier models. The results produced with the proposed method are found to be satisfactory and are discussed in detail in the Result section.

The rest of the paper is arranged as follows: Section 2 includes related work; proposed framework for the opinion mining is explained in detail in Section 3; experimental results are discussed in Section 4; and conclusion and future scope of the work is presented in Section 5.

2. RELATED WORK
Lot of research work is being undertaken in opinion mining in recent times. Researchers are working on classifying the sentiments of the reviewers for different domains like restaurant reviews, product reviews, and movie reviews etc. Sentiment classification task has been done using machine learning approach, lexicon approach or the combination of both the approaches which will produce a hybrid approach. It can be performed at three levels; document level, sentence level and feature level [7]. NB and SVM models are normally used as baselines for other systems in text labeling and sentiment analysis research. Pang and Lee [8] first used these classification methods in their experiments to classify movie reviews. A lexicon-based approach is based on the overall sentiment score of the sentiment words in the passage of text [9]. This approach is mentioned by Hu and Liu for the first time for aspect level and sentence level sentiment classification. Sentiment classification at the sentence level is analogous to document sentiment classification as sentences are part of the documents. But this task is difficult as sentences are less informative when compared to the entire document. There are different types of sentences like direct sentence (e.g. the movie is superb) and indirect sentences (e.g. Race 3 is almost like its previous version) which require more understanding of the problem. Feature level classification tries to determine the sentiment on certain aspects in the text reviews. The words, terms or the phrases present in the text passage which contribute in finding the polarity of the sentiment of the text passage are called as features. The machine learning systems first get trained on these features and then classify the unseen text. Selection of the best features ensures better accuracy of the classifier by reducing the dimensionality of the training data set. There are several approaches mentioned in the literature for finding out the finest features [10-12]. Opinions can be expressed in any language. Many researchers have worked on multilingual data. The work usually translates data from one language to another and then finds the sentiments of the original data. Cross-language sentiment classifiers are built for various languages like Chinese, Spanish, Arabic and Indonesian language etc. by many researchers achieving comparable results with the monolingual ones [13-17]. AbinashTripathy et al. [18] and Yuhui Cao et al. [19] mentioned that the combination of two different machine learning algorithms like SVM and ANN for sentiment classification yield better results when compared with other hybrid models. Yassine Al Amrani et al. [20] chose to used SVM and Random Forest for sentiment classification and introduced a novel hybrid approach to identify product reviews obtained by Amazon. They showed their hybrid approach increased the accuracy of the classifier model when compared with the individual algorithms. Back Propagation Neural network and Probabilistic Neural Network are employed by G.Vinodhini, R.M. Chandrasekaran because of their superior classification ability [21]. The authors of the paper “A Hierarchical Neural-Network Based Document Representation Approach for Text Classification” [22] integrate hierarchical neural architecture into traditional neural network methods and showed that their proposals outperform the corresponding neural network models for document classification. Nurulhuda and Ali [23] have mentioned three different weighting schemes to generate the word vectors which are Term Frequency-Inverse Document Frequency Binary Occurrence and Term Occurrence. Daniel Jurafsky and James H. Martin [24] showed that Naïve Bayes with binarized features seems to work better for several text classification tasks. Asha S Manek et al. [25] proposed a statistical method using weight by Gini Index method for selecting the features. Ouyang et al. [26] introduced word embedding features based on deep
learning technology for optimizing the accuracy of their proposed model to carry out attribute-level sentiment analysis. Lohann C. et al. [27] proposed a Genetic Algorithm approach to balance the corpus of texts for sentiment classification using a SVM classifier. They showed that balancing the corpus would increase the performance of the classifier up to 86.14% from 76.58% with the imbalanced corpus. The authors of the paper “Genetic Algorithm based Feature Selection in High Dimensional Text Dataset Classification” [28] used genetic algorithm based meta-heuristic optimization algorithm to improve the F1 score of classifier hypothesis and chose best features for SVM, MAXENT and stochastic gradient descent classification algorithms to find the classification models of public available datasets. With the selected features they achieved 97% accuracy as the best case.

Many researchers have developed feature selection algorithms which lack in finding the most informative features that are needed for the machine learning algorithms to produce accurate results. Single words or unigrams are considered as best features but they require more space and time for processing. Hence in this study, most persistent bigrams and trigrams are selected as informative features and further due to optimization of these informative features using genetic algorithm, a better performance is achieved in classifying the sentiments of the text reviews. Since ensemble methods combine a set of base classifiers in order to obtain more accurate and reliable classifier model, we propose an ensemble model in which the information provided by two kinds of feature sets is collected. Combination of feature sets is quite effective in the task of sentiment classification.

3. PROPOSED FRAMEWORK FOR OPINION MINING

The machine learning approaches require a set of useful features for sentiment classification. The feature selection approaches intend to select a small subset of features in order to minimize redundancy and maximize relevance to the target such as the class labels in classification. Different feature selection techniques include Information Gain, Relief, Fisher Score, Lasso etc [29]. A novel feature selection method called MPFS that make use of feature score and information gain of the features in the text is applied on bigram and trigram features in the documents. The feature set is further optimized using a genetic algorithm based technique to generate Optimized Feature Set (OFS). The feature set produced by MPFS is used to train ANN to produce ANN Feature Set (ANNSFS). The proposed ensemble classifier model SVM\text{A2N2} (SVM and ANN) uses both OFS and ANNSFS for classification task. The performance of this model is compared with the base classifier models. The Opinion Mining System architecture is shown in Figure 1. The brief description of the proposed framework is given in Algorithm1.

3.1. Review data collection

The review data can be collected from web which contains the social media data like Facebook, Twitter and blogs etc. Several review datasets of movies, products, restaurants etc. are available for sentiment classification of social media reviews using an ensemble classifier (Savita Sangam)
classification task. The dataset used in this work is movie reviews dataset developed by Pang and Lee. It contains 2000 processed positive and negative text files. The reviews on the movies are considered because they contain range of emotions or sentiments.

3.2. Data preprocessing
The activities involved here are:
\begin{itemize}
  \item a. Removal of punctuations marks (".", ",", "?" etc.)
  \item b. Filtering out natural language specific stop words (in, on, an etc.)
  \item c. Elimination of special characters ("@", ",", "/" etc.).
  \item d. Discarding repetitive characters like in okkkk, goo, noooo etc.
\end{itemize}

3.3. Feature selection
Feature selection techniques can be used to identify and remove irrelevant and redundant features that do not contribute in increasing the accuracy of the model. Several combinations can be made for selecting features which involves lots of effort. Therefore, sophisticated methods are required to perform feature selection in practice. Feature Ranking methods are generally used as they are simple and give good results. A suitable ranking criterion is used to score the variables and a threshold is used to remove variables below the threshold. The basic property of feature ranking is to identify the relevance of the features. The features that are not relevant to the class labels can be discarded. The MPFS method is applied here to find the most relevant features in the review documents.

3.3.1. MPFS method
The MPFS method tries to find out the most persistent features in the documents. Initially the feature set consists of all the bigrams like “movie is”, “is very”, “very beautiful” etc. Instead of considering all the bigrams, only useful bigrams like “very beautiful” which contributes mainly in finding the sentiment can be considered. The feature score of the features is used to find out the most persistent features. This score is calculated using the chi square statistic. Similar to bigrams, trigrams (e.g. “not so good”) are also considered here for the experimentation purpose. The experiments are conducted using top 5000; top 10000 and so on up to top 30000 features to test the performance of the classifier models. The models performed well as we increase the number of features but the time taken to train them also increased. The results were better for top 10000 features with moderate training time.

3.3.2. Optimization
One of the most sophisticated algorithms for feature selection is the genetic algorithm. This heuristic optimization technique is population-based and is very adaptive method for feature selection. It operates on chromosomes which are in bits. These operations include selection, crossover and mutation operations. Based on certain fitness value, these operations need to be applied iteratively to get the fittest chromosome in the population. The initial population is randomly selected from the available feature set. The iterative operations operate on one population of chromosomes to produce a new population. Crossover operator recombines two single chromosomes which are called as parents to produce a child. This child is further mated at some position in order to produce new population. In this work, OR operator is used to carry out crossover operation and single bit of the chromosome is muted or flipped to perform mutation operation. The features which are absent in the chromosomes are discarded and with the new feature set thus produced, the fitness value is found out. Here classification accuracy is taken as the fitness value. Thus at the end of the whole GA process the Optimized Feature Set will be generated. This OFS is then fed to the ensemble classifier model. A sample chromosome of 10 bits with 1 indicating the presence of a feature and 0 indicating absence is shown in Table 1. The crossover and mutation operations on this chromosome are presented in Table 2 and Table 3 respectively. The initial population here is the feature set produced by MPFS method. The detail explanation of the method is presented in Algorithm 2.
3.4. **Ensemble classifier model**

By combining different machine learning techniques, one can expect a better performance by the combined model than the individual one. The ensemble classifier SVMA2N2 model is an ANN-SVM combined model for sentiment classification. The ANN model is applied to learn feature vector representations produced from MPFS method for the labeled training data. The learned feature vectors produced by ANN are fed to train the SVMA2N2 classifier along with the OFS produced by GA method. Such a combined model is expected to combine the advantages of both ANN and SVM on efficient classification. The ANN model is trained on MPFS and generates ANNS. The SVMA2N2 classifier treats ANN model as the feature learner and SVM as the sentiment classifier. Thus the proposed model combines the advantages of ANN on feature learning and SVM on efficient classification. The model is shown in Figure 2. The entire process of supervised learning is illustrated in Figure 3. The detailed feature learning process of ANN to generate ANNFS is discussed in Algorithm 3.

**Algorithm 1: Procedure for the proposed approach**

Input: Reviews document set  
Output: Review documents classified as positive or negative  
Threshold = 0.5  
Step 1: Preprocess the review documents to filter out punctuations, stop words and special characters  
Step 2: Tokenize the preprocessed documents into bigrams (trigrams) features  
Step 3: Select the features with minimum occurrence of three or more times  
Step 4: Calculate Feature Score for all the features using chi-square score  
Step 5: Select feature with Feature Score greater than Threshold as most persistent feature  
Step 6: Create MPFS set with the most persistent features generated from Step 4-5  
Step 7: Generate OFS  
Step 8: Generate ANNS  
Step 9: Train SVMA2N2 using the OFS and ANNS  
Step 10: Test SVMA2N2 using cross validation method  
Step 11: Evaluate the performance of SVMA2N2  

**Algorithm 2: Generate OFS**

Input: MPFS  
Output: OFS  
Step 1: Generate Initial population of members with features set randomly selected from MPFS  
Step 2: Create chromosome of n bits indicating presence (1) or absence (0) of n features in the member  
Step 3: Assign number of members i.e. feature sets from the initial population to Iterations  
Step 4: Set Fitness value as accuracy of the model  
Step 5: Perform Crossover operation on parent chromosomes using OR operator to produce a child chromosome C  
Step 6: Carry out Mutation operation on C by mutating a single bit with each individual having a probability Pm to mutate where Pm=1/m, m being the number of features  
Step 7: Calculate accuracy of the model with feature set generated after crossover and mutation operations  
Step 8: Select the model with maximum Fitness value and assign its feature set to OFS  
Step 9: Repeat Steps 4-7 Iteration times  
Step 10: Return OFS  

**Algorithm 3: Generate ANNS**

Input: MPFS  
Output: ANNS  
Step 1: Initialize Initial Set (IST) with MPFS and create an empty set ANNS  
Step 2: Sort IST based on FS and Create Candidate Feature Set for training with set of ‘n’ Temporary Subsets (TSS, to TSS_n)  
Step 3: Initialize ANN. The number of input layer nodes is the size of the TSS.  
Step 4: for i=1 to n
Train ANN partially with TSS$_i$
Test ANN and find current classification accuracy (CACC)
if $\text{CACC} > 0.5$
Update ANNFS with TSS$_i$
end if
end for
Step 5: Return ANNFS

Figure 2. Ensemble classifier model: SVMA2N2

Figure 3. Supervised learning process of SVMA2N2

4. RESULTS

The experiments are conducted on Intel core i3-3220 CPU @ 3.30 Ghz processor with 32 bit operating system running Windows 7 Professional. Python 3.5.2 with NLTK 3.2.1version is used for programming purpose. The experiments are performed on movie reviews dataset. The movie reviews polarity
dataset used in this work contains a folder named movie_reviews. In this folder there are subdirectories called ‘pos’ and ‘neg’ which contain 1000 positive and 1000 negative processed text files respectively. The document feature vectors are generated from the aggregation of the feature sets OFS and ANNFS. The performance of the model is evaluated using 10-fold cross validation method. The evaluation parameters are calculated as given in the Equations (1) to (4).

\[
\text{Accuracy} = \frac{\text{NCCD}}{\text{TND}} \quad (1)
\]

\[
\text{Precision} = \frac{\text{NCCDX}}{\text{TNDX}} \quad (2)
\]

\[
\text{Recall} = \frac{\text{NCCDX}}{\text{TNDC}} \quad (3)
\]

\[
F1 = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (4)
\]

Where,

- TND = Total Number of Documents
- NCCD = Number of correctly classified documents
- X = Positive or Negative category
- NCCDX = Number of correctly classified documents belonging to X
- TNDX = Total Number of Documents in X
- TNDC= Total number of documents actually classified

The performance comparison of ensemble classifier model SVMA2N2 with baseline classifier models is shown in Table 4. The result shows that the performance of SVMA2N2 improved little bit over SVM and it outperforms NB and MaxEnt. The performance comparison of SVMA2N2 in terms of classification accuracy, with classifier models proposed by other researchers is tabulated in Table 5.

Table 4. Performance comparison of SVMA2N2 with a baseline classifier models

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>NB</td>
<td>0.814</td>
<td>0.836</td>
<td>0.817</td>
<td>0.826391</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>0.79</td>
<td>0.821</td>
<td>0.794</td>
<td>0.807274</td>
</tr>
<tr>
<td>SVM</td>
<td>0.963</td>
<td>0.979</td>
<td>0.973</td>
<td>0.975991</td>
</tr>
<tr>
<td>SVMA2N2</td>
<td>0.974</td>
<td>0.964</td>
<td>0.963</td>
<td>0.9635</td>
</tr>
</tbody>
</table>

Table 5. Performance comparison of SVMA2N2 with other classifier models

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Classifier</th>
<th>Reference</th>
<th>Dataset</th>
<th>Accuracy (in%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NB, MAXENT, SVM</td>
<td>[2]</td>
<td>IMDb</td>
<td>81, 80.4, 77.1</td>
</tr>
<tr>
<td>2</td>
<td>Hybrid of SVM and ANN</td>
<td>[16]</td>
<td>IMDb</td>
<td>95</td>
</tr>
<tr>
<td>4</td>
<td>Convolutional Neural Network</td>
<td>[31]</td>
<td>StockTwits</td>
<td>90.9</td>
</tr>
<tr>
<td>5</td>
<td>Model B (Tf-idf + Linear SVM)</td>
<td>[32]</td>
<td>newspaper headlines</td>
<td>91.52</td>
</tr>
<tr>
<td>6</td>
<td>SVMA2N2</td>
<td>Proposed</td>
<td>Movie-Review</td>
<td>97.4</td>
</tr>
</tbody>
</table>
5. CONCLUSION

Ensemble learning proves to be better than machine learning by combining various models. It is proved by researchers that by combining several models the accuracy of the new model can be improved and by combining more models the result can be better. Since the combination of most informative feature set processed by one model and optimized feature set generated by other, is used for ensemble learning the accuracy is better compared to the individual models. Feature optimization is one reason for the improved accuracy and the other reason is parallel processing of feature sets by SVM and ANN. The model is tested only on one domain ie movie reviews. The future work can include different domains and also the deep analysis of the input data.

REFERENCES


Sentiment classification of social media reviews using an ensemble classifier (Savita Sangam)