Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

CONVINCED ENHANCEMENTS IN FEATURE DIVERSITY FOR DATA MINING AND ITS APPLICATION IN OPINION MINING

Dr. A. SURESH
Department of Computer Science, Sona College of Arts and Science, Salem, India

A. KALEEMULLAH
Department of Computer Science, Mazharul Uloom College, Ambur, India

Abstract - Opinion Mining (OM), which is also known as Sentiment classification or a Polarity classification, is that binary classification of task labelling of an opinionated text/document that expresses either a positive or an overall opinion that is negative. Several technique of analysis of a subjective information in the classification of sentiment is available. The extraction of feature chooses sufficient features for characterizing the opinion. The feature selection (FS) that is also known as the variable selection, subset selection or attribute reduction available in machine learning. It does select an input of dataset which closely defines its specific outcome. Here in this research a feature selection mechanism that is wrapper based is used based on a heuristic algorithm. The experiments demonstrate that the proposed method improves the efficiency of the classifiers and achieve higher accuracy.

Keywords: *Opinion Mining (OM), Sentiment classification, Feature extraction, Feature Selection (FS), wrapper based feature selection*

I. INTRODUCTION

An Opinion mining or a sentiment analysis will come into play in case where the data is huge and makes it difficult for evaluating them personally. The opinion mining will learn the views of the individuals, the tests, the behaviour, the issues, the activities, the subjects and their features. The opinion is considerable if they are the primary influences of these behaviours. The subjects in the movies will be of significant interest from among the social networking of the individual community that has been recognised by a vast number of individuals that talk about movies and with a significant different in sentiments. The opinion mining of the movie reviews will be measure being more challenging than that of opinion mining of the other categories which are product reviews [1]. A Sentiment analysis is that emerging research area of text mining with its computational linguistics, that has attracted plenty of considerable research attention within the last few years.

The Sentiment analysis is a text classification which classifies text that is based on sentimental orientation of the opinions contained in them [2]. Several methods of opinion mining have been released either for determining if the sentences of feedback in that of a natural language will be objective or subjective of if they are negative or positive. Such methods have been released on the basis of real life needs like that of movie reviews that can lead to a success commercially for this research and the main components of these will be a produce fabrication that is based on the methods of opinion mining. This is a relatively new as well as a challenging field that is dedicated to the detecting of subjective content in the documents that are opinionated. Opinion mining will determine polarity of the comments that can be classified into either positive or negative or neutral by means of the attributes of the objects as commented in their reviews. These methods are required for sentiment classification, feature extraction, and opinion summarization which are the areas of

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

classification. It is evident from that of a prior example that making of a linguistic distinction among the objective words which will express the subjective words expressing opinions which will be important. There are several important segments in the sentiment analysis for the purpose of providing feature extractions. This is perhaps the most challenging task in the analysis and opinion mining. For automatic extraction of features it will need Natural Language Processing (NLP) techniques.

The Opinion or Sentiment words are those phrases or the words that imply both positive or negative feelings. Even though the opinion words that are typically either adverbs or adjectives, verbs or nouns that can be used for conveying emotions. The quantity of such high-dimensional features have been mentioned for enhancing the classification's precision, the selection methods are used. The feature selection is that subset of features from the documents. Feature Selection will be carried out by means of retaining the words with that of the greatest score according to the predefined metrics of prominence. A high dimensionality of the features space will be an issue in the classification of text. A feature selection which will be the process of selecting a small subset of M features from a set of N features that are decreased based on certain criteria [3]. The technique of data mining dimensionality reduction is that technique of selection which is an original feature set for specific criteria.

This will bring down the number of features that will remove the redundant or noisy data that provides the effects that will include the data mining speeding up that improves the performance of mining. This is an active research field that has developed in machine learning and the data mining for many years that is applied to those fields like that of the text mining, the intrusion detection, the image retrieval of the genomic analysis. If the new applications emerge there are several challenges that arise that require new methods or theories for addressing complex and high dimensional data. Removal of optimal redundancy, selection of stable features and exploitation of knowledge along with auxiliary data are the challenges in feature selection. Large volumes of literature have been published in feature selection [4].

II. LITERATURE REVIEW

Kaur and Saini [6] made a presentation and an analysis of various approaches for Sentiment Analysis and Opinion Mining. The formal and the informal text pieces are made in 8 different international languages. There are formal text pieces as poetry, documents, essay and informal text in the form of SMS, emails, micro blogs and chats. There are four parameters of feature selection that have been analysed for identifying the emotional states that are associated with such written texts. The parameter, the Information Gain (IG) and the Term Document Frequency with the Inverse Document Frequency (TDF-IDF), have been experimented by the Researchers several times and the IG outperformed all the other feature selectors. Prusa et al.,

[7] further made a study on the impact of ten of the filter- based techniques of feature selection by using ten feature subsets and the results demonstrated that the feature selection was able to improve significantly compared to not using a feature selection. In addition to this the choice of the ranker and the feature subset size had a significant impact on performance.

Karegowda et al., [8] made a proposal of a wrapper approach for the feature selection. Here the authors made use of a wrapper approach along with a genetic algorithm that was wrapped with four different classifiers and their induction algorithms which are the decision tree C4.5, the Naïve Bayes, the Bayes networks and the Radial basis function for the purpose of evaluating based on four other standards which are the Pima Indians Diabetes Dataset, the Breast Cancer, the Heart Stat log and the Wisconsin Breast Cancer. Furthermore, their attributes have been proposed by wrapper which are validate with the

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

classifiers. The results of experiment that employs the feature subset by using the proposed wrapper approach for enhancing the accuracy of classification. Claypo and Jaiyen [9] made a proposal for opinion mining and classifying the reviews and to analyse the consumer attitude. Here the artificial neural network has been applied for classification of both positive and negative reviews. Additionally, the mRMR feature selection is used for selecting the features and to reduce the feature number in the dataset. As a consequence of this the time of computation of the learning algorithms for the neural networks that have been reduced. The results of the experiment showed that this will be an effective model to classify the reviews on Thai restaurants.

Uribe [10] made proposals for a method of hybrid feature selection for distinguishing salient features which permit the identification of the viewpoint that is underlying that of a text review which is to determine the polarity of the sentiment. The method has used the fundamental tasks of pre-processing called filter and wrapper techniques. This approach has been demonstrated on a dataset in which each document will be represented by tow feature vectors that are based on two rule sets. Shahid et al., [11] made a proposal of a Biogeography based optimization algorithm for the selection of optimal features from that of a given data and further by means of using the Naïve Bayes and the Support Vector Machine techniques, authors further performed reviews of sentiment classification of the products. This technique may be applied to certain classification problems in which the feature set is large.

Kwon et al., [12] made a proposal for a novel method using a sentiment trend analysis that makes use of the Ant Colony Optimization (ACO) algorithm and also the SentiWordNet. Initially the authors collect some social data as that of the Resource Description Framework (RDF) and triples, after which the ACO algorithm is used for digitizing the RDF triples that are amassed. By using the ACO algorithm the authors further compute the values of pheromone for extracting the trends of the sentiments for the users and for this the authors make a comparison of the sentiments that have been analysed with their real and daily lives. Liang et al., [13] made a presentation of an optimized framework by means of incorporating the sentiment and contextual information and the authors exploited two of the phenomena of sentiment which are sentiment matching: for the polarities of documents and their sentiment words that are similar and sentiment consistency: the polarities of two of the frequently occurring words that are similar. The results demonstrated that these models had outperformed significantly the current approaches. Kumar et al., [14] further studied the state-of-art swarm intelligence algorithms that have been used for the subset selection inside the framework of sentiment analysis.

This study proved that techniques of swarm intelligence will bring some significant gains in terms of accuracy. There are also swarm algorithms that have been applied and more than can be explored providing an insight into those expounded for improved analysis. Sumathi et al., [15] made a proposal for an Artificial Bee Colony (ABC) algorithm for the purpose of optimization of the feature subset. The Naïve Bayes, the Fuzzy Unordered Rule Induction Algorithm (FURIA) and the Ripple Down Rule Learner (RIDOR) and their classifiers have been used for classification. This method that has been proposed with the features that are extracted on the basis of the Inverse Document Frequency (IDF). This is used for reduction of the size of feature subset and the complexity of increasing the accuracy of classification. The facets in the sentiment of medical sphere as well as the potential use cases have been identified by Saraswathi and Tamilarasi [16]. The study further proposed the Ant Colony Optimization-2OPT algorithm which is a classification framework that extracts the feature sets from the reviews using the Term Frequency- Inverse Document Frequency. The features that are chosen have been classified using the Naïve Bayes, and the Support Vector Machine

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

(SVM) and the experimental results showed this method improved the efficiency of classifiers to classify opinions.

III. METHODOLOGY

The feature selection Maximum Relevance Minimum Redundancy (mRMR), the Particle Swarm Optimization(PSO), the naïve bayes and the k nearest neighbor have been detailed.

Maximum Relevance Minimum Redundancy (mRMR)

The approaches to Feature selection may be roughly grouped into the filter based methods, the wrapper-based methods [17] and finally the embedded methods. Another special group of such approaches to filter-based feature selection will choose a highly predictive and also uncorrelated feature. An ideal example will be the Maximum Relevance Minimum Redundancy (mRMR) algorithm that has been developed for the purpose of feature selection. This will select the features having maximum correlation with that of a class relevance along with least correlation among themselves and here in this algorithm the be the Maximum Relevance Minimum Redundancy (mRMR) algorithm that has been developed for the purpose of feature selection. This will select the features having maximum correlation with that of a class relevance along with least correlation among themselves and here in this algorithm the features have been ranked in accordance to the criteria of minimal- redundancy-maximal-relevance criteria.

This mRMR is that feature selection approach which will choose the features that have a high correlation with that of a low correlation and for the continuous features for calculating calculation with that of a class relevance the F-statistic will be used and the coefficient of correlation of Karl Pearson will be used for calculating the correlation among the features called redundancy. After this the features will be chosen by means of applying another greedy search for maximizing this objective function and the two types of objective function are the MID (the Mutual Information Difference criterion) and the MIQ (the Mutual Information Quotient criterion) that represent the quotient of the redundancy or the relevance. For the temporal data the mRMR approach to feature selection needs some techniques of preprocessing which will flatten the temporal data within a single matrix. This can also lead to loss of information from among the temporal data.

Particle Swarm Optimization (PSO)

The Particle swarm optimization or PSO, was initially proposed by Kennedy and Eberhart [18] that had been inspired by the fish schooling or a bird flocking social behavior. The swarm has certain particles each with a component that represents a certain solution with a component velocity that represents the direction of the movement of the particle in case of a solution space. The PSO is also an iterative algorithm that has three main steps. The first being the initialization of the population by means of generating the velocity component of the particle and the component with its position component in a random manner. The next step is the evaluation of the solutions that are represented by the positions of the particles and the final step is updating the velocity of the particles using equation. Both the second and the third positions are repeated till such time the stop criterion is duly met.

A detailed algorithm has been shown below. So while evaluating the fitness value of this particle, such a training set can be divided into 10 folds and will run on the training set with a scheme of supervised machine learning that the particle represents. After this a fitness value will be that of the average accuracy of the 10 runs. An algorithm for this feature selection [19] is::

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

```
Initialize Parameters of PSO
Randomly Initialize Swarm
While stopping criterion not met do
    For i=1 to swarm size do
        Calculate P_i's fitness value
        Update the pbest of P_i
        Update the gbest of P_i
    End
    For i=1 to swarm size do
        For j=1 to dimension do
            Update the velocity of P_i
            Update the position of P_i
        End
    End
End
Return the best feature subset found by the swarm
```

Classifiers

The Two Supervised Machine Learning algorithms that have been used are the Naïve Bayes 'and the K-Nearest Neighbor for the calculation of accuracies, the precisions and the recall values. Sentiment Analysis is an opinion word treated as the way a positive side is considered as that of a negative inanother situation. The degree of negativity or positivity will have a great impact on their opinions and the latest text mining will give room for the advanced analysis that measures the word's intensity. In this point btoh accuracy and efficiencyof algorithms may be scaled.

Naive Bayes (NB)

The Naive Bayes classifier is that simple model that is used for classification [20] being simple and working well with text classification. It is probably the simplest form of the Bayesian Network wherein all the attributes are independent with a given value of that of the class variable. This is known as conditional independence that assumes each of the feature which is conditionally independent to that of the other problem classes. From the numerical approach group this Naïve Bayes has many other advantages of them being fast, simple and with high accuracy.

K-Nearest Neighbour (KNN) Classifier

The K-NN is that type of instance that is based on either learning or lazy learning in which the function is locally approximated. This is a non-parametric method that is used for either regression or classification. In the output classification will be the class membership (in which the most common clustercan be returned) where the object has been classified by means of a majority of the vote of the neighbors which is being assigned to the class that is most common in the k nearest neighbors. The rule hasretained the whole set of training at the time of learning and will assign a class that represents a majority label of k nearest neighbors in that of the training set.

The Nearest Neighbor rule (NN) which is one simple form of K- NN when the K = 1. With an

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

unknown example as well as a training set the distance that exists between the unknown sample and all of the samples in the training set will be computed. The actual distance between that of the smallest value will correspond with the sample in the training set that will be the closest to a sample that is unknown. So an unknown sample can be classified on the basis of classification to the nearest neighbor. This K-NN will be an easy to understand algorithm and also for implementation purposes. This is a powerful tool for sentiment analysis as it does not assume anything about its data except for the distance measure that is consistently calculated among two instances [21]. This is also called nonparametric or even non-linear not assuming a functional form.

IV. RESULTS AND DISCUSSION

Table 1 and figure 1 to 3 shows the results of recall, precision and F Measure respectively.

	NB-MRMR	NB-PSO	NB-KNN	PSO-KNN
Recall	0.846867	0.871533	0.832867	0.852
Precision	0.8637	0.8881	0.849967	0.868067
F Measure	0.853533	0.8782	0.8395	0.858367

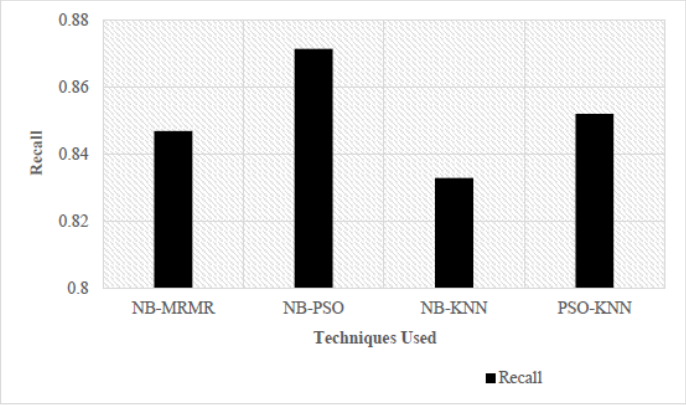
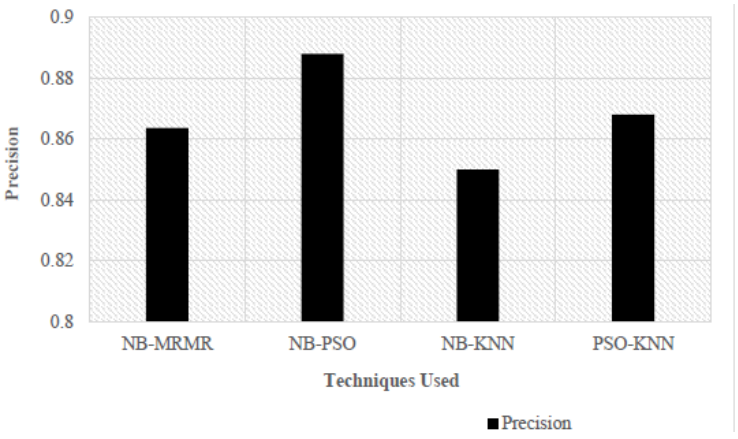


Figure 1 – Recall

From table 1 and figure 1 it is observed that the recall of NB-PSO performs higher by 2.87%, by 4.54% and by 2.27% than NB-MRMR, NB- KNN and PSO-KNN respectively.



Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

Figure 2 Precision

From table 1 and figure 2 it is observed that the precision of NB-PSO performs higher by 2.79%, by4.39% and by 2.28% than NB-MRMR, NB-KNN and PSO-KNN respectively .

From table 1 and figure 3 it is observed that the F Measure of NB- PSO performs higher by 2.85%, by4.51% and by 2.28% than NB- MRMR, NB-KNN
and PSO-KNN respectively.

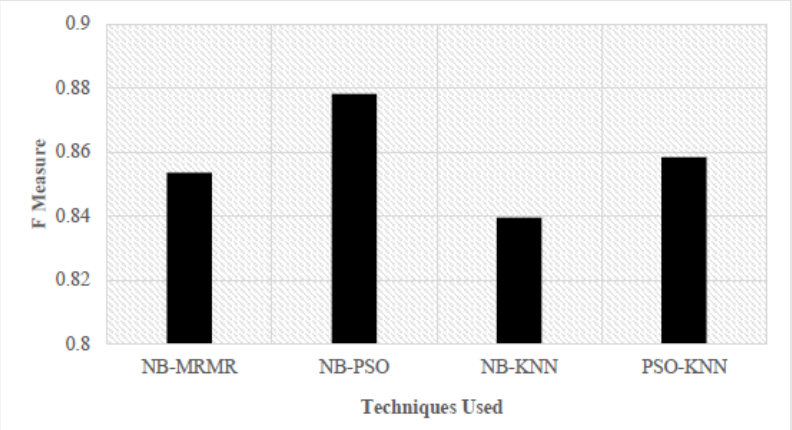


Figure 2 F-Measure

V. CONCLUSION

An approach to sentiment classification is through making use of a machine learning algorithms. The tasks of sentiment analysis will also include classification of a text polarity at the level of a document, a feature, a sentence which may be either positive or negative or even neutral. Feature selection is a procedure of learning which saves the time of operation by eliminating the redundant and irrelevant features. The learning algorithms without intervention with the irrelevant, the redundant and the noisy features and also build a simple and a precise data model. This Feature selection will build a simpler and a common model to get a better insight inside the fundamental perception of a task. This work proposed and evaluated PSO based feature selection. The results have demonstrated that precision of the NB-PSO will perform much higher by about 2.79%, 4.39% and 2.28% than that of the NB-MRMR, the NB-KNN and the PSO-KNN respectively. This recall of that of the NB-PSO will perform much higher by about 2.87%, 4.54% and 2.27% than that of the NB- MRMR, the NB-KNN and the PSO- KNN respectively. This F Measure of the NB-PSO has performed higher by about 2.85%, 4.51% and 2.28% than that of the NB-MRMR, the NB-KNN and the PSO-KNN respectively.

REFERENCES

1. Basari, A. S. H., Hussin, B., Ananta, I. G. P., & Zeniarja, J. (2013). Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization. Procedia Engineering, 53, 453-462.
2. Govindarajan, M. (2014). Bagged Ensemble Classifiers for Sentiment Classification of Movie Reviews. International Journal of Engineering and Computer Science, 3(2), 3951-3961.

Tianjin DaxueXuebao (ZiranKexueyuGongchengJishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online): 0493-2137
E-Publication: Online Open Access
Vol:54 Issue:04:2021
DOI 10.17605/OSF.IO/K9NE8

3. Madhusudhanan, & Srivatsa. (2016). Feature Selection Using Improved Shuffled Frog Algorithm For Sentiment Analysis of Book Reviews. IIOAB JOURNAL, 7(9), 526-534.
4. Jotheeswaran, J., & Kumaraswamy, Y. S. (2013). Opinion Mining Using Decision Tree Based Feature Selection Through Manhattan Hierarchical Cluster Measure. Journal of Theoretical & Applied Information Technology, 58(1).
5. Brindha, G. R., Prakash, S., Santhi, B., & Swaminathan, P. (2015). Enhanced Facet Ranking and Text Classifier for Opinion Mining. Applied Mathematics & Information Sciences, 9(3), 1147.
6. Kaur, J., & Saini, J. R. (2013). An analysis of opinion mining research works based on language, writing style and feature selection parameters. Int. J. Adv. Netw. Appl.
7. Prusa, J. D., Khoshgoftaar, T. M., & Dittman, D. J. (2015, May). Impact of Feature Selection Techniques for Tweet Sentiment Classification. In FLAIRS Conference (pp. 299-304).
8. Karegowda, A. G., Jayaram, M. A., & Manjunath, A. S. (2010). Feature subset selection problem using wrapper approach in supervised learning. International journal of Computer applications, 1(7), 13-17.
9. Claypo, N., & Jaiyen, S. (2014, July). Opinion mining for Thai restaurant reviews using neural networks and mRMR feature selection. In Computer Science and Engineering Conference (ICSEC), 2014 International (pp. 394-397). IEEE.
10. Uribe, D. (2011, November). Optimizing feature selection techniques for sentiment classification. In Electronics, Robotics and Automotive Mechanics Conference (CERMA), 2011 IEEE (pp. 103-107). IEEE.

11. Sumathi, T., Karthik, S., & Marikkannan, M. (2014). Artificial Bee Colony Optimization for Feature Selection in Opinion Mining. Journal of Theoretical & Applied Information Technology, 66(1).

12. Quadri, K.A., Imafidon, C.E., Akomolafe, R.O.Kolaviron mitigates proteinuria and potentiates loop diuresis in Wistar rats: Relevance to normal renal function(2019) Journal of Complementary Medicine Research, 10 (1), pp. 58-67.

13. Radovic, M., Ghalwash, M., Filipovic, N., & Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. BMC bioinformatics, 18(1), 9.

14. Kennedy J, Eberhart R (1995) Particle swarm optimization. Proc IEEE Int Conf Neural Netw 4:1942–1948.

15. Shang, L., Zhou, Z., & Liu, X. (2016). Particle swarm optimization- based feature selection in sentiment classification. Soft Computing, 20(10), 3821-3834.