Performance Analysis Of Ensemble Feature Selection Method Under SVM And BMNB Classifiers For Sentiment Analysis

M.Gunasekar, Dr.S.Thilagamani

Abstract: The process of sentiment analysis identifies whether a given piece of text is positive, negative or neutral. It helps the business organizations to make use of the data in an effective way and to take informed decisions. It also saves human time and effort since this is an automated process. To automate the sentiment prediction the comment or review of a user has to be synthesized accurately. Since the feature selection is an important factor in sentiment prediction, this paper uses composite n-gram model with two feature selection methods mRMR (Minimum redundancy and Maximum Relevance) and improved Gini index. Then the sentiment prediction is carried out using SVM (Support Vector Machines) and BMNB (Binary multinomial Naïve Bayes) classifier. Experiment results shows that BMNB classifier performs better under both feature selection methods.

Key words: Text classification, feature selection, Gini index, Naïve Bayes, mRMR, SVM

INTRODUCTION:

Now-a-days social networks has become a common platform for the public to share their opinions, express their views and give their comments on a product, brand name or service, etc.. These opinions are of great use for business organizations to know about their customer's feedback. But it is not possible for organizations to read such huge data. Here comes the need for proper monitoring and organization of such huge data into easily understandable form. The process of identifying reader's attitude towards a brand, or a product or a service is called Sentiment Analysis. It is also called Opinion mining because it acts on the opinions of the public. Sentiment Analysis is of great importance in today's world because; it gives an idea about the opinion of the general public. The opinion of the users can be computed using different methods; supervised learning is one such method which shows better performance comparatively. Feature selection is an important task in sentiment analysis, which rightly picks the most contributing feature for the sentiment prediction. The research in sentiment analysis is progressing in a perspective to capture the opinion intuitively from user. Varela et.al. [7] Compared the performance of feature selection methods for classifying the polarity of a document. In his work SVM and Multinomial Naive Bayes classifiers are applied at two situations. first without using feature selection method and using feature selection in that the latter method shows 10% more accuracy that the former. Then among the two classifiers Multinomial naive bayes shows better result. Harun et.al. [8] Created two step feature selection and extraction mechanism. In first step the important feature in the document are identified using its Information gain value. The step two involves Genetic algorithm to further optimize the feature selection method and using Principal component analysis the promising features for the classification is extracted. KNN and Decision tree are applied to classify the polarity.

RELATED WORK:

Sufficient research works are being carried out on sentiment prediction, still a space is available to improve the accuracy of the sentiment prediction models. Understanding the data is a crucial task and data has to be transformed into a format which allows the classifier models to capture the sentiment more accurately. Preprocessing enhances the performance of the model. Angiani et al. [4] have converted the emoticons into simple positive and negative opinion for better classification. Since stemming process converts a word to its root word capturing the essence of words like "couldn't go good", "wouldn't sufficient" are difficult. Negation in words influences the meaning of the word which surrounds it. Omitting these words would lead to misclassification. These words are synthesized by appending 'not' (negation) before it. Michelle Annett [5] compared the Lexical and ML approaches to find the accuracy of the classifier. They have proved that Machine Learning approach provides high accuracy than Lexical methods. Among SVM, NB and ADtree Machine Learning algorithms, SVM shows better result. Wenqian Shang, et al. [6] modified the gini index equation to improve the feature selection result. Here the probability function $P(W)$ is replaced with posterior probability $P(W|C_i)^2$ which finds out the presence and absence of the text feature in the document. Narayanan, et al. [9] developed a fast sentiment classifier model using Enhanced Naïve Bayes algorithm. Vivek et.al uses laplacian smoothing to classify the word which is not found in the training data. Negation words are handled by appending 'not' before the word. Mutual information is used to calculate the dependence of the random feature and its classes. [10] used two feature selection method based on unigram, bigram and composite feature. Information Gain and minimum redundancy and maximum relevance are applied on an IMDB dataset for feature selection. SVM and Multinomial Naïve Bayes are used to classify the text polarity.

PREPROCESSING:

Text categorization research area has gained its importance since the inception of social media. Before analyzing the data, the data preprocessing is an essential task, as not all data in the dataset are going to contribute for knowledge discovery. We consider movie review dataset prepared by Pang & Lee, which is one of the Dataset much explored by the researchers.

Preprocessing involves the following steps:

1) Removing the stop words from the document to make the dataset more clear for data processing.
2) Converting the documents into tokens and transforming the upper case alphabets to lower case and removing the high frequency words.
3) Stemming process transforms the words in the document to its original or root word, which helps in identifying the similar words in the document. The words with similar meaning are removed from the document.

Categorization of text:
In this module the data in the document are modeled based on the n-grams model. In n-gram model n represents the sequence of words to be grouped from a given dataset. In this experiment n are assigned with values 1 and 2. Example: In unigram model (1-gram) text data "every people has unique dream" is modeled as "every", "people", "has", "unique", "dream". In bigram model (2-gram) "every people", "people has", "has unique", "unique dream".

Feature selection:
The efficiency of the classifier can be improved by removing the words which do not contribute for the text classification. Feature selection is an important task in sentiment analysis. In this paper two feature selection algorithms, novel gini index and minimum redundancy maximum relevance algorithm are used. These two methods calculate the dependency of words towards the class and words with low dependency are removed from the document.

Improved gini index:
Gini index method was widely used to calculate the influence of a word in a document. This algorithm was first considering only the mutually exclusive features in the above dataset. This can be achieved with the help of following,

min L(S), L = $\frac{1}{|S|^2}\sum_{x_i,x_j \in S} I(x_i, x_j)$

max ϕ(R,L), ϕ = R – L

The condition merging the maxR and min L is called mRMR. The above max ϕ optimizes the R and L, it also finds the near optimal value when incremental search is applied.

Classifiers:

BMNB:

Naive Bayes is one of the classification model used widely for text classification. The classical NB is a simple probabilistic model which utilizes the bayes theorem for its computation. In Multinomial NB model [9] frequency of a term in the document is considered, whereas in the binary MNB it considers only the occurrence of a feature not its frequency.

At first the documents are labeled with its class using $P(c_i)$ and to restrict the more than one occurrence of the feature $P(W_k/c_j)$ is calculated.

$P(c_i) = \frac{|docs_i|}{|total\ docs|}$

$P(W_k/c_j) = \frac{n_{k+\propto}}{n+\propto|Vocabulary|}$

Where the docs represents the documents in the corpus and $P(c_i)$ calculates the probability of each documents belonging to class i. n_k refers to the frequency of W_k in $text_j$. all the duplicate words in the dataset is removed and NB is computed using the following equation,

$C_{NB} = argmax\ P(c_i)\ \prod_{i \in positions} P(\frac{W_i}{c_i})$

SVM:

The SVM classifier has been adopted by many researchers for sentiment classification. This [5] works on both regression and classification problems. The non-linear inseparable problems are well addressed. The non-linear inseparable data's are transformed into linear data's using the kernel of SVM. This kernel transforms the data into linear separable by building hyperplanes in a multidimensional space that distinguishes data's belonging to different classes. The following minimizes the error function,

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N} \epsilon_i$$

Follows the constraints,

$y_i(w^T \phi(x_i) + b) >= 1-\epsilon_i$ and $\epsilon_i \geq 0, i = 1, \ldots, N$

where b, c are constant and capacity constant respectively. ϵ_i handles non separable inputs. N represents the training cases and i labels the training cases and w is the vector of coefficients. ϕ transforms the input non-linear data to an independent feature space.

coined in 1984 by Breiman. In general, Gini index computes the impurity of the words in text categorization. The general form of Gini index is,

Gini(S) = 1- $\sum_{i=0}^{m} P_i^2$

Where S is the sample set and there are m classes, each s_i belongs to a class c_i. P is the probability of each sample s_i/s. This suffers from the default maximum value problem i.e. when all the samples are equally distributed to classes. [6] The following formula depicts the modified gini index algorithm,

GiniT(W) = ∑ $P(W/C_i)^2$ $P(C_i/W)^2$

The above formula replaces P_i^2 with $P(W/C_i)^2$, which considers the unbalanced class distribution. This new formula considers posterior and condition probability of all the features to reduce the unbalanced class problem.

mRMR:
This feature selection method reduces the redundancy in the dataset. mRMR is a heuristic approach which matches the maximum relevancy among the data's in the dataset and removes the redundant data's from the dataset.

mRMR algorithm [10] finds the discriminant attributes of a class. It works by grouping the features having high relevance. The maximal relevance among data's in a dataset is identified using the following

max R(S,c), R = $\frac{1}{|S|}\sum_{x_i \in S} I(x_i, c)$

The features with maximum relevance would have high redundant feature set. This redundancy can be minimized by

EXPERIMENTS AND RESULTS:

The experiment is on the IMDB movie review dataset prepared by pang et.al. This study is based on sentiment classification on movie review dataset. Dataset consists of 2000 reviews which are equally segregated as 1000 positive and 1000 negative reviews. Experiment follows the process flow shown in figure 1.

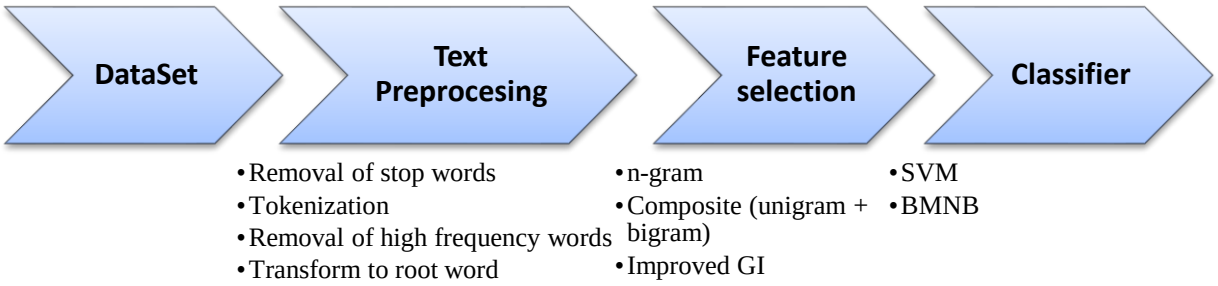

Figure 1: *Process flow diagram*

The preprocessing methods such as tokenization, stop word removal, transformation to root word are applied on the dataset. Feature vector is generated using the n-gram model. Feature selection methods mRMR and improved Gini index are used identify the influencing features in the This study clearly shows that the feature vector generation based on composite model (unigram + bigram) shows better performance. Feature selection methods both mRMR dataset. Then two classifiers BMNB and SVM are applied to classify the sentiment in the document. The performance of the classifiers is captured using the confusion matrix constructed from the results of the classifiers. Precision, Recall and F-score are the parameters considered for analyzing the performance of the model.

and improved gini index shows significant improvement than classical n-gram methods. The performance of the classifiers SVM and BMNB are below.

Table 1*: Results of the proposed model*

Method	SVM			BMNB		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Unigram	83.7	83.4	83.55	87.2	83.2	85.22
Bigram	82.1	81.7	81.79	86.4	84.8	85.62
Composite (unigram+Bigram)	85.9	82.4	84.14	85.9	84.2	85.04
Composite modified GI	88.5	89.2	88.84	92.1	89.6	90.84
Composite mRMR	89.7	88.4	89.04	90.2	89.3	89.74

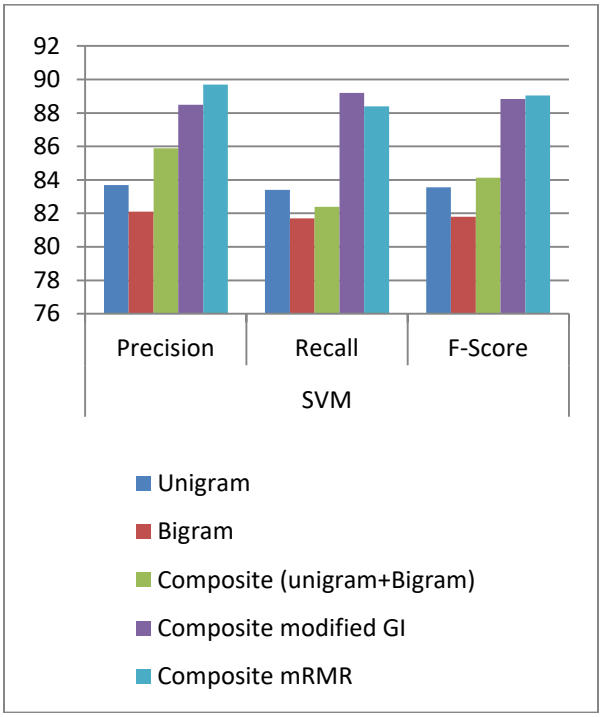

Figure 2: *comparison of Feature selection methods on SVM*

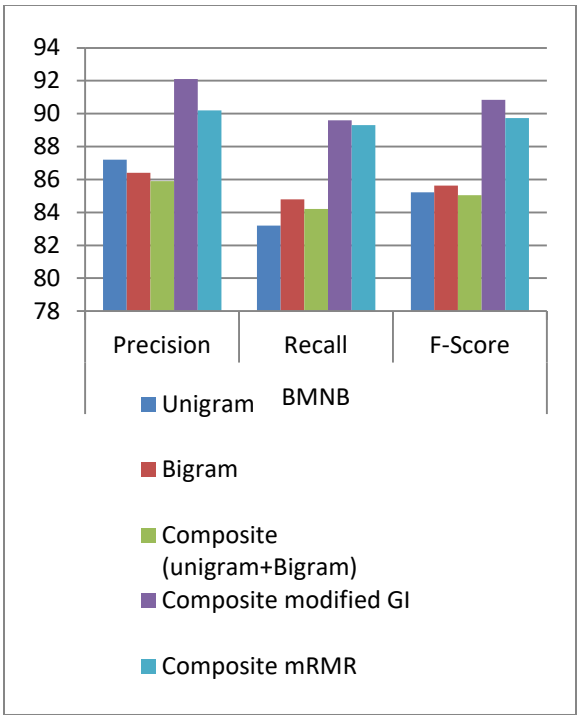

Figure 3: *comparison of Feature selection methods on BMNB*

optimization method with the classical feature selection method to improve the overall performance of the classifier model.

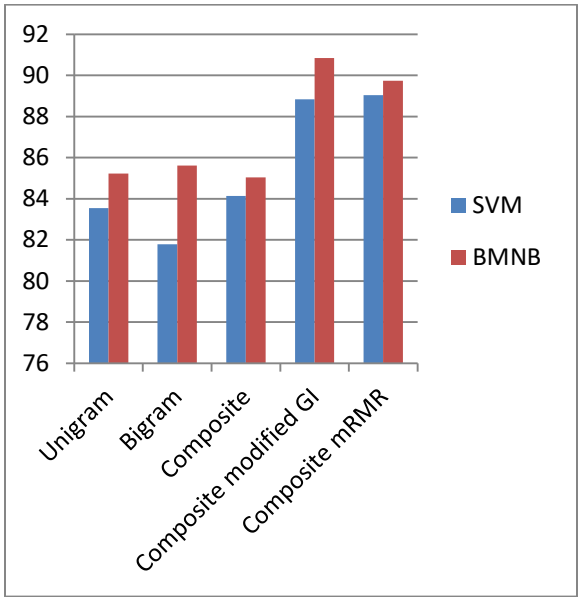

Figure 4: *Performance comparison of SVM and BMNB*

The results shows that the BMNB classifier performs better than SVM under both improved and mRMR feature selection methods.

CONCLUSION:

Sentiment analysis becomes an important research area because of the increase in the use of social media by people. Opinion of the customers has great impact in the profit of an organization. So capturing customer's feedback and understanding their opinion is an essential task for organizations. In this study we proposed a model which uses two feature selection methods improved Gini index and mRMR. The results of feature selection methods are fed into machine learning classifiers SVM and BMNB for opinion mining. BMNB performs better than other classifier model. Since the feature selection method has a great influence in the performance of classifier models, in future the study can be enhanced by including the

REFERENCES:

[1] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

[2] Angiani, Giulio, et al. "A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter." KDWeb. 2016.

[3] Annett, Michelle, and Grzegorz Kondrak. "A comparison of sentiment analysis techniques: Polarizing movie blogs." Conference of the

Canadian Society for Computational Studies of Intelligence. Springer, Berlin, Heidelberg, 2008.
[4] Wenqian Shang, et al. "A novel feature selection algorithm for text categorization." Expert Systems with Applications 33.1 (2007): 1-5.
[5] Varela, Pedro & Martins, André & Aguiar, Pedro & Figueiredo, Mário. An Empirical Study of Feature Selection for Sentiment Analysis. (2013)
[6] Uğuz, Harun. "A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm." Knowledge-Based Systems 24.7 (2011): 1024-1032.
[7] Narayanan, Vivek, Ishan Arora, and Arjun Bhatia. "Fast and accurate sentiment classification using an enhanced Naive Bayes model." International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, 2013.
[8] Agarwal, Basant, and Namita Mittal. "Optimal feature selection for sentiment analysis." International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, 2013.
[9] Ghosh, Monalisa, and Goutam Sanyal. "Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis." Applied Computational Intelligence and Soft Computing 2018 (2018).
[10] Mitra, Vikramjit, Chia-Jiu Wang, and Satarupa Banerjee. "Text classification: A least square support vector machine approach." Applied Soft Computing 7.3 (2007): 908-914.
[11] Zhang, Wen, Taketoshi Yoshida, and Xijin Tang. "Text classification based on multi-word with support vector machine." Knowledge-Based Systems 21.8 (2008): 879-886.
[12] W. Zhao, Y. Wang, D. Li, A dynamic feature selection method based on combination of GA with K-means, in: 2nd International Conference on Industrial Mechatronics and Automation 2010, pp. 271–274.