NOVATEUR PUBLICATIONS
INTERNATIONAL JOURNAL OF INNOVATIONS IN ENGINEERING RESEARCH AND TECHNOLOGY [IJIERT]
ISSN: 2394-3696
VOLUME 4, ISSUE 5, May-2017

SENTIMENT ANALYSIS USING FEATURE SELECTION AND CLASSIFICATION ALGORITHMS

SHRUTI PANT
Department of computer science and engineering, MPUAT University/College of technology and engineering, Udaipur, India, pant.shruti5@gmail.com

KALPANA JAIN
Assistant professor of CSE, College of technology and engineering, Udaipur, India,kalpana_jain2@rediffmail.com

ABSTRACT
Here we present a technique to compute the sentiments of movie review dataset so that the overall performance of the model is optimised. This model is certain to train and test the model and find the performance constraints. We first pre-process the dataset followed by feature selection and then we will classify the features to investigate the performance. A textual movie review is important as it reveals strong and weak points of the movie plot and by doing the deeper analysis of a movie review one can tell if movie will meet the expectations of the reviewer.

KEYWORDS: Sentiment Analysis, Feature Selection, Classification, Naïve Bayes, Chi Square.

INTRODUCTION
The present Techno savvy people has brought upon the social media era which has become a huge database for gigantic amount of data. It makes data mining important because of the availability of structured and unstructured data which is created and consumed by the users. Users express their judgement and belief on various social media channels such as Facebook, Twitter, Rotten Tomatoes, Blogs and many others.
Text classification (TC) is basically an instance of text mining which categorises the given object into a predefined category. In other words, given a set of categories, and a collection of text documents, text categorization or TC is the process of finding the correct topic for each document [1]. Text classification is generally divided in single label where exactly one category is assigned to each document and multi-label where a document may belong to more than one category or class.
Sentiment analysis is the field of study which analyse the text, opinions, sentiments, evaluations, attitudes, and emotions using combination natural language processing (NLP) and machine learning. As sentiments are key influencers of our behaviour, this technology will be very important in next few years.
Sentiment analysis has three different levels: document level deals with classifying entire document in positive or negative [2] [3], sentence level is very similar to document level but instead of document each statement is classified [4] [5] and finally aspect level in which object is classified with respect to a particular facet or aspect [6].
In this paper we will look at various feature selection and classification techniques. The rest of the paper is organized as follows. First is thebackground study of sentiment analysis, next is the proposed methodology, then experimental results and finally we will cover the conclusion and future score.

SENTIMENT ANALYSIS
(a) FEATURE SELECTION
A "feature" (attribute or variable) refers to the characteristic of the data. Features that may be discrete, continuous, or nominal are usually collected before the features are specified or chosen. Features can be:

- Irrelevant: Irrelevant features are those which do not influence the output.
- Relevant: These are features which have an influence on the output as they have an inherent meaning which cannot be assumed by the rest.
- Redundant: A redundancy exists whenever a feature can take the role of another.

There are four basic steps in a typical feature selection method (figure 1):

- A generation procedure generates next candidate subset which retains enough information for better performance of the model ;
- An evaluation function evaluates the candidate subset;
- A stopping criterion decides when to terminate;
- A validation procedure validates the subset.

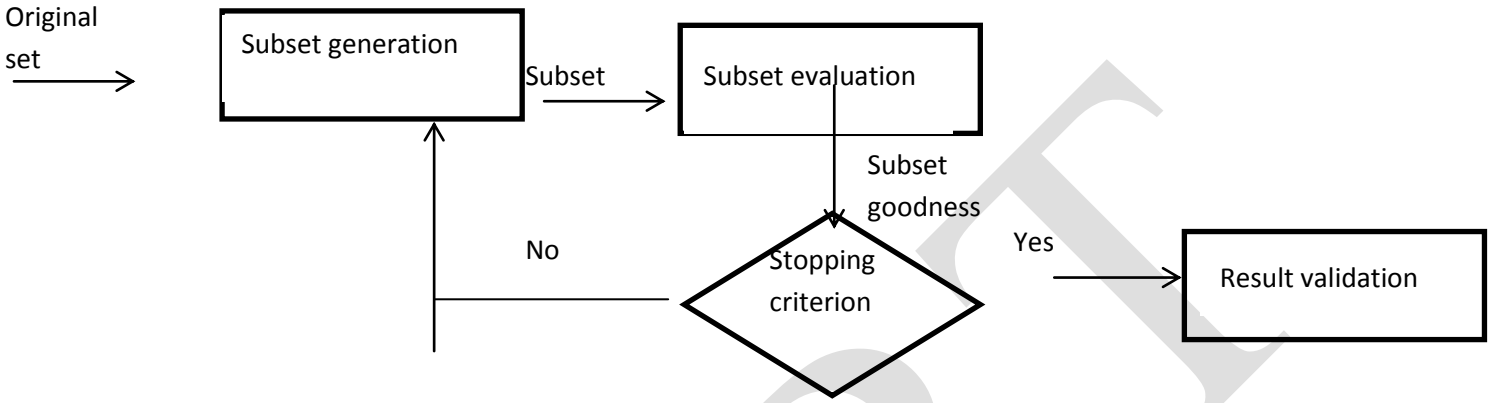

Figure 1 steps in Feature Selection

The generation procedure is essentially a search procedure [7] [8] whichbasically aim to generate subsets of features for evaluation process. The generation function begins:

- With no features,
- With all features,
- With a random subset of features.

Features are iteratively removed or added respectively in the first and second case whereas features are either added or removed iteratively or produced randomly in third case [8].

New goodness of subset produced by generation function is compared with the old one and is replaced with if needed. This process is done by evaluation function.

Stopping criteria can be based on either generation function or evaluation function. It helps in deciding when to stop.

CLASSIFICATION

It is a supervised function whose work is to classify text into a given class or label. There are various types of classifiers in machine learning like Fisher's linear discriminant, Support vector machines, k-nearest neighbour, Boosting (meta-algorithm), Decision trees, Random forests, Neural networks, Bayesian classifiers. Some examples of fields in which text classification is used are:

- News filtering: Automated methods re now a days used in categorization of news in a variety of web portals [9].
- Email Classification and Spam Filtering: Also referred to as spam filtering or email filtering, it classifies email in order to determine whether the email is a spam or not in an automated way[10][11][12][13].
- Document Organization and Retrieval: These are used for digital libraries of documents, scientific literature, web collections, or even social feeds. [14].

PROPOSED METHODOLOGY FOR SENTIMENT ANALYSIS

In the presented work we train and test an automated model for sentiment analysis which not only classify the feature as positive or negative but also finds the system's performance. To design the model we have used various technologies in each step of construction showed in figure 2.

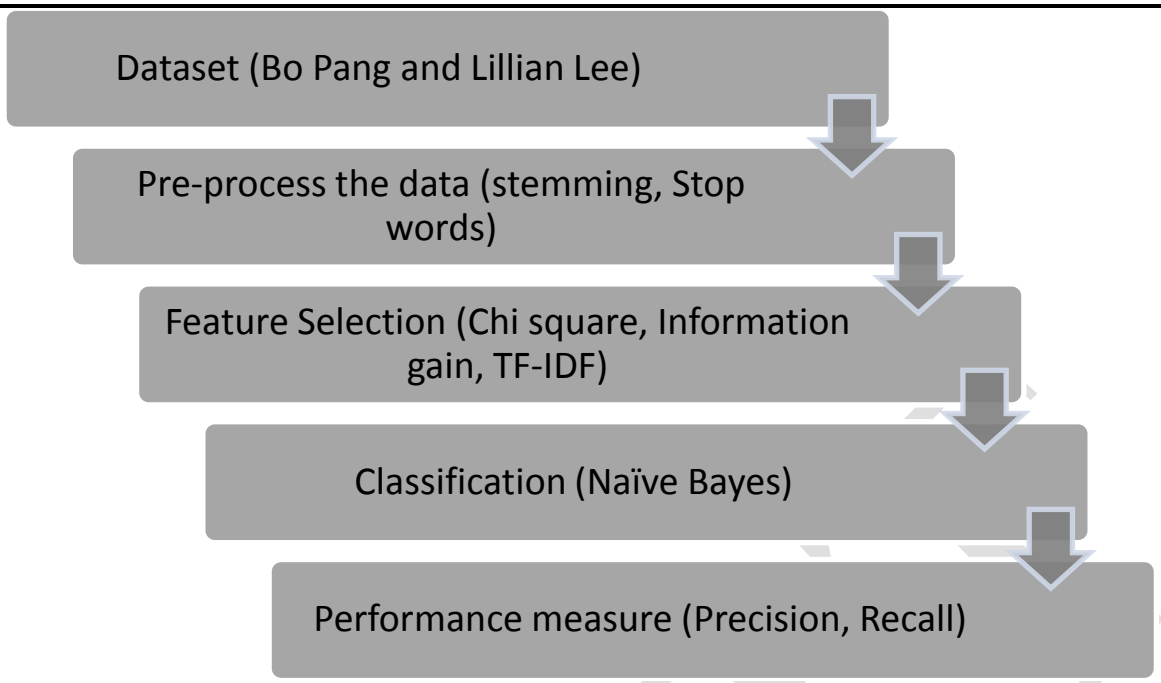

Figure 2 Steps in Proposed methodology of sentiment analysis

STEMMING

Stemming techniques tries to find out the root of a word. Stemming convert words to their stems which take into account language-dependent linguistic knowledge. As the words with the same root mostly describe same or relatively close meaning, these words can be conflated. For example, the words, work, workers, worked, using all can be stemmed to the word 'WORK'.

CHI SQUARE

Chi Square Test is used in the field of statistics to test the independence between two events. For the calculation of chi square, we take the square of the difference between the observed (o) and expected (e) values and then divide it by the expected value. Chi Square measures the deviation between expected counts (e) and observed Count (o).

$$\chi 2 = \Sigma \frac{(o - e)^2}{e}$$

INFORMATION GAIN

Information gain measures the relevance of a feature for prognosis of a class by knowing the presence or absence (frequency) of a particular term in a document. In short, after the value of feature is obtained the information gain measures the reduction in entropy of the class variable or we can also say that it measures how frequent a feature is in one class when compared to other classes. Information gain is formulated as:

$$I(w) = -\sum_{i=1}^{k} P_i \cdot \log(P_i) + F(w) \cdot \sum_{i=1}^{k} p_i(w) \cdot \log(p_i(w)) +$$

$$+(1 - F(w)) \cdot \sum_{i=1}^{k} (1 - p_i(w)) \cdot \log(1 - p_i(w))$$

TF-IDF

Tf-idf (term frequency-inverse document frequency) weight is a statistical measure often used in text mining for information retrieval which evaluates the importance of a term in the document or a corpus. The variations of this technique are often used by search engines for scoring and ranking the relevance of the document. When the no of times the word appearance in a document increases, the importance of a word also increases proportionally but is offset by the frequency of the word in the entire database.

TF (t) = (Number of times term t appears in a document) / (Total number of terms in the document).

IDF (t) = $\log_e$ (Total number of documents / Number of documents with term t in it).

NAÏVE BAYES

Naïve Bayes' is based on Bayes' Theorem which relates conditional probabilitiesand finds its roots in probability theory that shows the effect of the occurrence of one event on another. The important terms in bayes theorem are the prior probability which is the probability obtained in the beginning before any additional information is obtained and posterior probability is the revised probability after some evidence is obtained.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where,

- P(A) is the prior probability of A
- P(B) is the prior probability of B
- P(A|B) is the posterior probability of A given B
- P(B|A) is the posterior probability of B given A

EXPERIMENTAL RESULTS

We used Bo Pang and Lillian Lee dataset to evaluate the performance of our trained and tested model not only by accuracy but also by other measures which help in the better understanding of model computations.

ACCURACY

Accuracy is the most common performance measure and it is a ratio of correctly predicted observation to the total observations. It may seem that if the model has high accuracy then the model is best but only when you have symmetric datasets where values of false positive and false negatives are similar. Therefore, it's advisable to look at other parameters to evaluate the performance of the model.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

PRECISION (COMPLETENESS)

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. High precision corresponds to the low false positive rate.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall (Sensitivity)

Recall is the ratio of correctly predicted positive observations to the all observations in actual class.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F-MEASURE

The F_1 score is the harmonic mean of precision and recall where an F_1 score reaches its best value at 1 and worst at 0. F-score is an 'average' of both precision and recall. Using harmonic mean provides an appropriate way to average ratios.

$$F1 = 2\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

The dataset of Bo Pang and Lillian Lee contains a total of 2000 reviews (1000 positive, 1000 negative). The table 1 shows the no of initial features are 35000, after the application stop words no of features remaining are 11666 and finally after stemming the features that remain are 8100.

Table 1 Features after pre-processing

Pre-processing	No of Features
Total	35000
Stop Words	11666
Stemming	8100

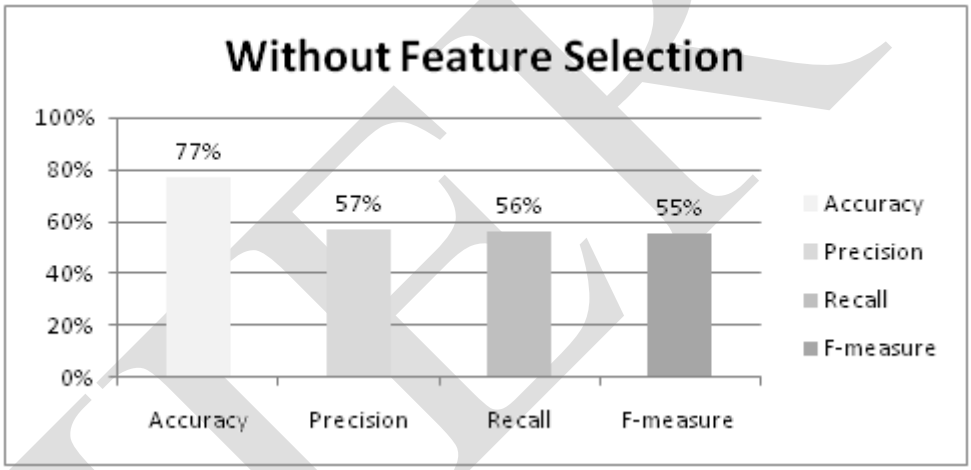

Figure 3 Results before application of feature selection

Figure 3 shows the result after the model is tested on the dataset which uses no feature selection techniques. Accuracy obtained is 77%, precision is 57%, recall is 56% and f-measure is 55%.

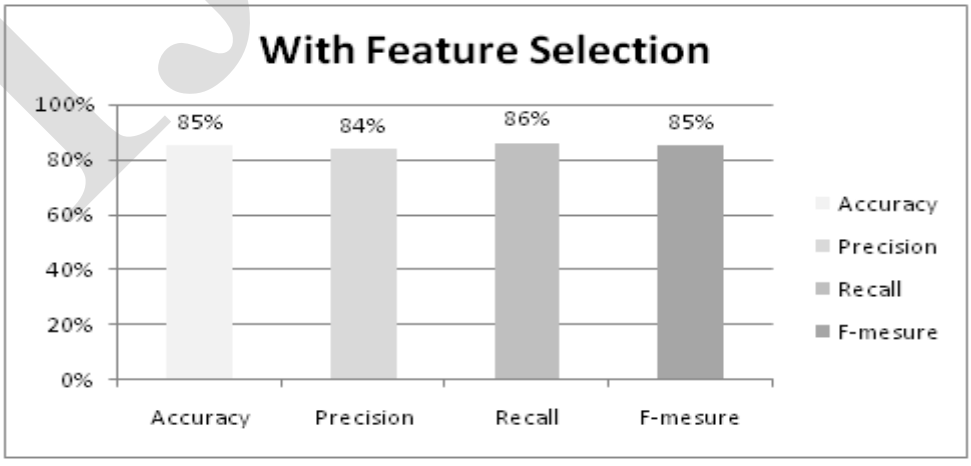

Figure 4Results after application of feature selection

Figure 4 portrays that after applying feature selection our model's accuracy has increased by 8%, precision which shows approximately 28% improvement, recall shows 30% improvement and F-measure also shows 30% improvement. Thus we have constructed a model which classifies the sentiment of users with less false positives and less false negatives.

CONCLUSION

In this paper, we consider the task of opinion and sentiment classification on the Document Level. We have done sentiment analysis on supervised classifier like Naïve Bayes using Bo Pang and Lillian Lee dataset for the two classification categories (positive/negative, opinionated/factual).

The model proposed is just an initial step of improvement in the techniques for sentiment analysis. There is substantial scope for improvement in the corpus formation and effective pre-processing and feature selection. In future, we would like to extend this technique on other domains of opinion mining likes newspaper articles, product reviews, political discussion forums etc. We would like to apply in-depth concepts of NLP for improved prediction of the polarity of the document.

REFERENCES

1) Doquire, G., &Verleysen, M. (2013). *Mutual information-based feature selection for multilabel classification. Neurocomputing, 122, 148-155.*

2) Pang, B., Lee, L., &Vaithyanathan, S. (2002, July). *Thumbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.*

3) Turney, P. D. (2002, July). *Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.*

4) Riloff, E., & Wiebe, J. (2003, July). *Learning extraction patterns for subjective expressions. In Proceedings of the 2003 conference on Empirical methods in natural language processing (pp. 105-112). Association for Computational Linguistics.*

5) Terveen, L., Hill, W., Amento, B., McDonald, D., &Creter, J. (1997). PHOAKS*: A system for sharing recommendations. Communications of the ACM, 40(3), 59-62.*

6) Hu, M., & Liu, B. (2004, August). *Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.*

7) Kira, K., & Rendell, L. A. (1992, July). *The feature selection problem: Traditional methods and a new algorithm. In AAAI (Vol. 2, pp. 129-134).*

8) Narendra, P.M. and Fukunaga, K., A *branch and bound algorithm for feature selection. IEEE Transactions on Computers, C-26(9):917–922, September 1977.*

9) Koller, D., &Sahami, M. (1996). *Toward optimal feature selection. Stanford InfoLab.*

10) Siedlecki, W. and Sklansky, J., *On automatic feature selection. International Journal of Pattern Recognition and Artificial Intelligence, 2:197–220, 1988.*

11) Langley, P. (1994, November). *Selection of relevant features in machine learning. In Proceedings of the AAAI Fall symposium on relevance (Vol. 184, pp. 245-271).*

12) Lang, K. (1995, July). Newsweeder: Learning to filter netnews. *In Proceedings of the 12th international conference on machine learning (pp. 331-339).*

13) Carvalho, V. R., & Cohen, W. W. (2005, August). *On the collective classification of email speech acts. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 345-352). ACM.*

14) Cohen, W. W. (1996, March). *Learning rules that classify e-mail. In AAAI spring symposium on machine learning in information access (Vol. 18, p. 25).*

15) Lewis, D. D., & Knowles, K. A. (1997). *Threading electronic mail: A preliminary study. Information processing & management, 33(2), 209-217.*

16) Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998, July). *A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).*

17) Chakrabarti, S., Dom, B., Agrawal, R., &Raghavan, P. (1997, August). *Using taxonomy, discriminants, and signatures for navigating in text databases. In VLDB (Vol. 97, pp. 446-455).*