Efficient Sentiment Analysis using Optimal Feature and Bayesian Classifier

Neha Gupta
M.tech Student,
Dept. of C.S.E, JMIT
Radaur, India

Shabnam Parveen
Assistant Professor,
Dept. of C.S.E, JMIT Radaur, India

ABSTRACT
Sentiment analysis refers to a broad range of fields of the natural language processing, computational linguistics, and text mining. Mining is used to extract previously unknown information from the different written resources. This extracted information is helping in decision making process. Sentiment analysis has gained much attention in recent years. It determines the opinion and attitude of the people towards a particular topic. This paper focuses to improve the accuracy by using the optimal feature and reduces the complexity by Naïve Bayes classifier. In proposed work, comparing the results with the existing model regarding the accuracy, precision, recall and f-measure which shows that performance are improved in each and every case.

Keywords
Sentiment analysis, Natural language processing, Optimal feature, Naïve Bayes classifier.

1. INTRODUCTION
Sentiment analysis is defined as finding the attitude/emotion/opinion manifest by a person about a particular topic. Internet has become an important part of in each and everyone life. It is mainly useful for user to share their opinion and views in a very short time. Sentiment analysis is useful for extracting the important information from the user's view. Sentiment analysis uses natural language to identify the subjective information from the source material.

Sentiment analysis and subjectivity are the automatic identification of private states of the human mind (i.e., opinions, emotions, sentiments, behaviors and beliefs). Further, subjectivity detection focuses on identifying whether data is subjective or objective. Where in, sentiment analysis classifies data into positive, negative and neutral categories.

There is a large explosion today of sentiment available from social media including face book, twitter, blogs, message boards and user forums. These snippets of text are the gold mine for companies and individuals that want to monitor their reputations and get timely feedback about their products and actions. Sentiment analysis offers these organizations the ability to detect the contrary social media sites in real time and act accordingly.

Sentiment analysis occurs at different level i.e. document level, sentence level, aspect/feature level. In document level classification sentiment is extricated from the whole review, and a total opinion is classified based on the all-inclusive sentiment of the opinion holder. The objective is to classify a review as positive, negative, or neutral. In sentence level the process involves two steps .The sentence level sentiment analysis is used to identify whether the sentence is subjective or objective and then only subjective sentences are determined to be positive, negative or neutral. The objective types of sentences are ignored, as there is no sentiment bearing words in objective type of sentences. In feature level the aim is to recognize and extract object features that have been commented on by the opinion bearer and decide whether the opinion is positive, negative, or neutral.

With the development of micro blogging websites, such as Twitter and Weibo, have achieved the huge popularity and stimulated hundreds of millions of users. People post huge short messages day-to-day to release latest news and share their opinions on the miscellaneous topics. Exploring the sentiments in these large-scale micro blog messages can help finding the public's opinions on products, political events, companies and so on, which has many applications [11].

The two main approaches of sentiment analysis are machine learning based and lexicon based. The first method is a lexicon-based approach which is used for sentiment dictionary with opinion words and finds the polarity by matching them with words. The second method is a machine learning approach that is used to classify the text.

2. RELATED WORK
[1] N. D. Valakunde et al. computes the document level sentiment analysis by calculating the aspect level sentiment score based upon underlying entity. It shows that SVM has better accuracy over NB. [2] Pavitra.R et al. proposes a document level sentiment classification in co-occurrence with topic sentiment analysis of bigrams and topic detection at the same time. The usage of bigrams for classification rather of unigrams enhances the efficiency of sentiment classification. [3] Ms.K.Mouthami et al. prefers a document level sentiment classification with bag of words in actual system to determine the opinion document whether positive or negative and proposed the new algorithm sentiment fuzzy classification with parts of speech tag to improve accuracy on the benchmark datasets of the movie reviews. [4] Siti Rohaidah Ahmad et al. prefers the metaheuristic algorithm which is used as feature selection in sentiment analysis. This algorithm is used for selecting the optimum features from the customer reviews. [5] Abdullah Dar et al. deals with the study of different methodology which is used for opinion mining and sentiment analysis and it is used to calculate about the movie review, product review etc. [6] Deepali Virmani et al. proposed algorithm which integrate aspect factor with sentiment value. Preciseness of aspect is evaluated by the aspect factor. The review about the aspect is more specific when aspect factor is high. [7] Mohammad Sadegh Hajmohammadi et al. uses novel learning model which is based on composition of uncertainity based active learning approaches and semi-supervised self-training in cross lingual sentiment classification to increase the performance. [8] A. khan et al. uses the rule based domain independent sentiment analysis method that classifies the subjective and objective

sentences with reviews and comments where subjective sentence is taken out from SentiWordNet to calculate the polarity and achieves the accuracy. [9] Soujanya Poria et al. proposed novel technique for multimodal sentiment analysis that contains the features of text and visual data and uses these features and decision level for fusing the feature extracted from different modalities. [10] Mostafa Karamibekr et al. focuses on sentiment analysis of social issues where they conduct statistical investigation on difference between the sentiment analysis of social issues and products. Use of BOW approach to classify the sentiment orientation of comments about social issues. [14] Kim Schouten et al. focus on aspect level sentiment analysis that is used to find aggregate sentiment on entities mention in document or aspect. [15] Leonardo Rocha et.al proposed a sentiment analysis by collective inspection that uses the lexicon-based unsupervised method that extracts the collective sentiment without concerning the individual classification.

3. PROPOSED WORK

The methodology used in this research is shown in Fig. 1.

3.1 Dataset

Use of dataset of the movie review data (*www.cs.cornell.edu/people/pabo/movie-review-data/*) collected by the reviewers which has more than 500 comments in the terms of positive and negative comments. For the experiments, we read the dataset which have two steps firstly, extract the file list from the dataset which gives the folder name and the file name after that read the document file.

3.2 Preprocessing

The first step is preprocessing which is used to remove inconsistent data or noisy data from dataset. It includes several tasks.

3.2.1 Tokenizer

Textual data consist of block of characters called tokens. Documents is divided into tokens and used for further processing.

3.2.2 *Stop word removal*

After tokenizing, some undesired words are discarded form sentence by using removal of stop words algorithm. Some of the frequently used stop words are “a”, “me”, “of”, “the”, “he”, “she”, “you”.

3.2.3 Stemmer

Stemming is the process of reducing the words to the root words. For example "argue", "argued", "argues", "arguing", and "argus" reduces to the stem "argu".

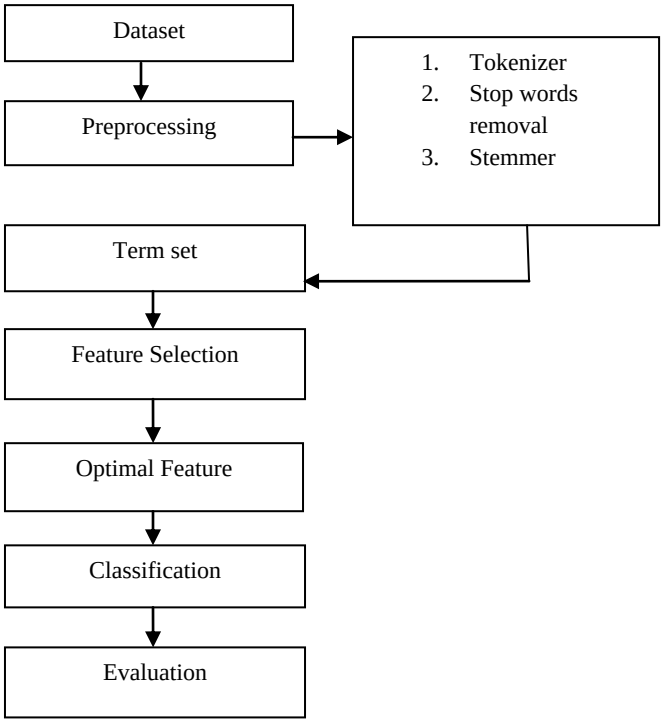

Fig1. Sentiment analysis steps and techniques

3.3 Term Set

Having two methods to find the positive and negative words:-

1. In existing work, the positive and negative words are manually mentioned then do the feature selection on the basis of these words. After feature selection, optimal words are left which always improved the results.
2. In proposed work, automatically calculate the term set using total set of positive and negative words then apply the filtering on the set and get the words which are lie on the dataset and then use the optimal feature with the concept of gain ratio on the filter words. By doing that, get the optimal words which gives the better results always.

3.4 Feature Selection

TF-IDF is one of the widely used representations. TF-IDF method assesses term frequency in document or the relevance of the term in the collection of document. The TF and IDF are defined as

TF (t) = Number of times the adjective term occurs in document (d)/Total Number of adjective in document (d) (1)

IDF (t) =log {ND/DF (t)} (2)

Here ND is total number of document in the document collection and DF (t) is number of documents in which adjective term (t) occurs in the document collection.

The output that comes from feature selection is VSM (vector space model). VSM is an algebraic model for representing text document as vectors of identifiers such as index terms. It is used in information filtering, information retrieval indexing and relevancy rankings.

3.5 Optimal Feature

Feature subset selection is of significant importance in the field of data mining. The high dimension data makes

training and testing of general classification methods difficult. So, uses the gain ratio which is filter feature subset approach that has been used to rank the attributes of datasets.

Gain ratio (GR) is a modification of the information gain that decreases its bias. When choosing an attribute gain ratio takes size of branches and number into account. It corrects the information gain by taking the intrinsic information of a divide into an account. Intrinsic information is entropy of distribution of instances into branches (i.e. how much info do we need to tell which branch an instance belongs to). Value of attribute decreases as intrinsic information gets larger. [12].

$$Gain\ Ratio(attribute) = \frac{\text{Gain (attribute)}}{\text{intrinsi c_info (attribute)}} \qquad (3)$$

3.6 Classification

Classification is performed on features extracted for every entity set. For each entity there is a classifier which is evaluated for a set of movie reviews. In this paper, Naïve Bayes classifier is used to decrease the complexity.

3.6.1 Naïve Bayes

Naïve Bayes classification is based on Bayesian theorem of statistics. A Naive Bayes classifier is a probabilistic model which is based on the Bayes rule with assumption of independence. This makes the algorithm faster and does not affect its accuracy.

This classifier merely calculates the conditional probabilities of the distinct classes given the values of attributes and then selects the class with the highest conditional probability. If an instance is represented with n attributes a_i(i=1…n),then the class that instance is classified to a class v from set of possible classes V according to a Maximum a Posteriori (MAP) Naive Bayes classifier is,

$$V = \arg\max P(v_j)^n \ \pi_{i-1}\mathrm{p}\,(a_i|v_j) \qquad (4)$$

Eq. (4) gives conditional probability acquired from the approximate of the probability mass function using training data. This Bayes classifier minimizes the probability of classification error under the assumption that the sequence of points is independent [13].

3.7 Evaluation

In the context of classification, having four parameters such as True Positives (TP), True Negatives (TN), False Negatives (FN) and False Positives (FP) are used to compare the class labels assigned to documents by a classifier with the classes the items actually belongs to. True positives (TP) are examples that the classifier correctly labeled as belonging to the positive class.

False positive (FP) are examples which were not labeled by the classifier as belonging to the positive class but should have been. True Negative (TN) are the examples that the classifier correctly labeled as belonging to the negative class. At last there is False Negative (FN), is example which was not labeled by the classifier as belonging to the negative class but should have been.

3.7.1 Accuracy

Accuracy is one of the common measure of the classification performance. Accuracy is used to describe the closeness of the measurement to the true value. When the term is applied to sets of measurements of the same measurand, it involves a component of systematic error and a component of random error. In this case trueness is the closeness of the mean of a set of measurement results to the actual value. Accuracy is the proportion of correctly classified labels to the total number of labels.

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \qquad (5)$$

Table1. Contegency table

		Correct labels	
		Positive	Negative
Classified labels	Positive	TP(True positive)	FP(False positive)
	Negative	FN(False negative)	TN(True negative)

3.7.2 Precision and recall

Precision and recall are two most generally used metrics for calculating the performance in text mining and in other text analysis field. Precision measures the exactness whereas recall measures the completeness. Precision is defined as the proportion of true positives against all positive results while recall is defined as proportion of true positive against positive results that should have been returned.

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \qquad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad (7)$$

3.7.3 F-Measure

F-measure is the harmonic mean of precision and recall. It considers the both precision and recall of the test to compute the score. F-measure is more useful than accuracy. If false positive and false negative have same cost then it works best or if they have different cost it's better to look precision and recall.

$$\text{F} = \frac{2*\text{Precision}\ *\text{Recall}}{\text{Precision}\ +\text{Recall}} \qquad (8)$$

4. EXPERIMENTAL SET UP

The goal is to study the sentiment analysis of movie reviews datasets. Dataset containing a more than 500 comments that expressed the public opinions in terms of positive and negative reviews. This paper focus only on the positive and negative reviews.

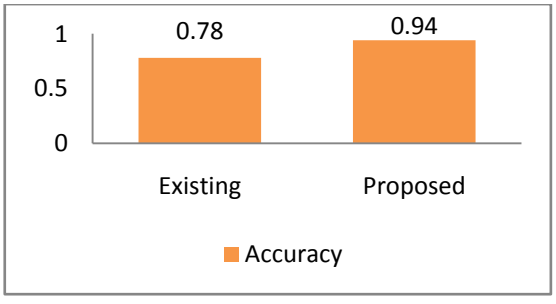

Fig2. Graphical analysis of accuracy

Fig 2 shows the graphical analysis of accuracy is increased in proposed model as comparing with the existing model.

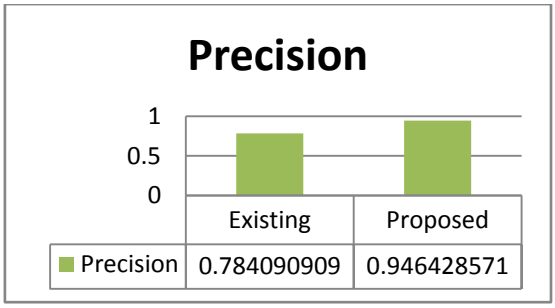

Fig3. Graphical analysis of precision

Fig 3 shows the graphical analysis of precision is improved in proposed model as comparing with the existing model. It also called the positive predictive value.

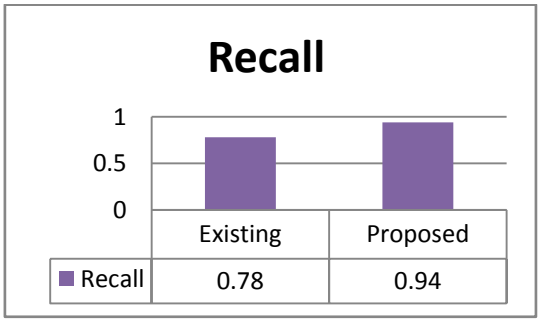

Fig4. Graphical analysis of recall

Fig 4 shows the graphical analysis of recall which is also called sensitivity is improved in proposed model as comparing with existing model.

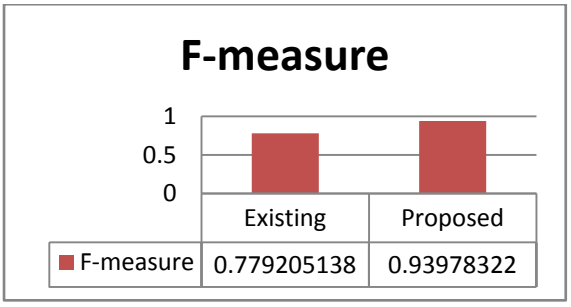

Fig5. Graphical analysis of f-measure

Fig 5 shows the graphical analysis of f-measure is improved in proposed model as comparing with existing model.

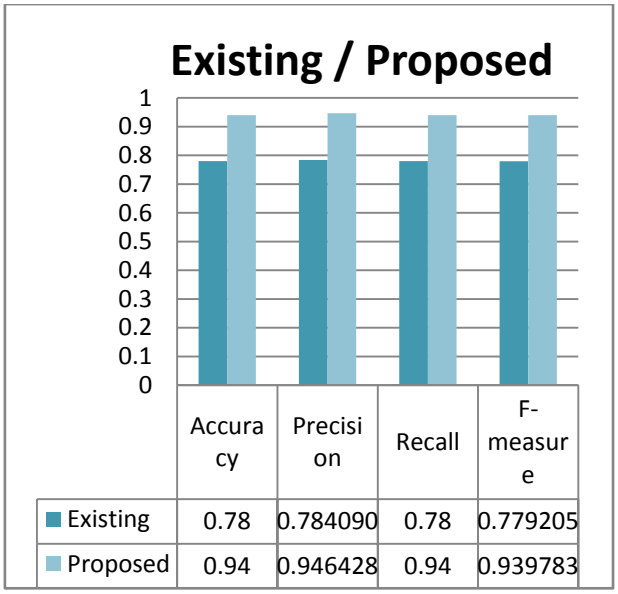

Fig6. Graphical analysis of overall

Fig 6 shows the graphical analysis of overall in which accuracy, precision, recall and f-measure are improved in proposed model as comparing with the existing model by using the optimal feature with the concept of gain ratio and naïve bayes classifier.

Comparative Analysis:
The comparison of accuracy, precision, recall, f-measure used by proposed model with existing model, which enhanced the sentiment analysis by using optimal feature and naïve bayes classifier can be shown in table 2 below:

Table2. Comparison table

	Existing	**Proposed**
Accuracy	0.78	0.94
Precision	0.784090909	0.946428571
Recall	0.78	0.94
F-measure	0.779205138	0.93978322

5. CONCLUSION

In this paper, examine the sentiment analysis of movie review in terms of positive and negative. The experiment shows that the accuracy, precision, recall, f-measure of proposed work results are improved than the existing work by using the optimal feature which uses the concept of gain ratio and naïve bayes classifier which reduces the complexity.

In future, plan to continue the research in the following directions (i) validation on large dataset as proposed model work on small dataset (ii) Ensemble classifiers such as decision tree or random forest can be used to improve the performance of the proposed work.

6. REFERENCES

[1] Valakunde, N.D and Patwardhan, M.S. 2013. Multi-Aspect and Multi-Class Based Document Sentiment Analysis of Educational Data Catering Accreditation Process. International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, pp.188-192.

[2] Pavitra, R. and PCD, Kalaivaani. 2015. Weakly supervised sentiment analysis using joint sentiment topic detection with bigrams. IEEE sponsored 2nd International Conference on electronics and communication system (ICECS '2015).

[3] Mouthami, K. and Nirmala, K.D. 2013.Sentiment Analysis and Classification Based On Textual Reviews. International Conference on Information Communication and Embedded Systems (ICICES), pp.271-276.

[4] Ahmad, S.R., Bakar, A. and Yaakub, M.R. 2015. Metaheuristic Algorithms for Feature Selection in Sentiment Analysis. Science and Information Conference 2015 July 28-30 | London, UK..

[5] Dar, A. and Jain, A. 2014. Survey paper on Sentiment Analysis: In General Terms International Journal of Emerging Research in Management & Technology ISSN: 2278-9359, Vol-3, Issue-11.

[6] Virmani, D., Taneja, S. and Bhatia, P. 2015. Aspect Level Sentiment Analysis to Distil Scrupulous Opinionated Result. International Conference on Computing, Communication and Automation, pp. 59-65.

[7] Hajmohammadi, M.S., Ibrahim, R., Selamat, A. and Fujita, H. 2015. Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. Information Sciences 317, pp. 67–77.

[8] Khan, A. and Baharudin, B. 2011. Sentiment classification using sentence-level semantic orientation of opinion terms from blogs. IEEE National Postgraduate Conference.

[9] Poria, S., Cambria, E., Howard, N. and Hussain, A. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. Neurocomputing 174, pp. 50–59.

[10] Karamibekr, M. and Ghorbani, A.A. 2012.sentiment analysis of social issues. International Conference on social informatics, pp.215-221.

[11] Ren, F., Wu, Y. 2013. Predicting user-topic opinions in twitter with social and topical context, IEEE Trans. Affect. Comput. 4(4) 412–424.

[12] Han, J. and Kamber, M. 2001.Data Mining Concepts and Technique.

[13] Ye, J., Povinelli, R.J and Johnson, M.T. 2002 Phoneme classification using naive bayes classifier in reconstructed phase space. Proc. of IEEE Signal Processing Society 10th Digital Signal Processing Workshop, pp. 37- 40.

[14] Schouten, K. and Frasincar, F. (2015). Survey on Aspect-Level Sentiment Analysis. IEEE Transactions on Knowledge and Data engineering.

[15] Rocha, L., Mourao, F., Ferreira, R. (2015). SACI: Sentiment analysis by collective inspection on social media content. Web Semantics: Science, Services and Agents on the World Wide Web 34, 2015, pp. 27–39.