Performance Evaluation for Sentiment Classification of Movie Review

Shinija S[1] Mrs P. Kalaivani[2]
[1]P.G. Student [2]Associate Professor
[1,2]Department of Information Technology
[1,2]St. Joseph's College of Engineering, Chennai

Abstract— Sentiment Analysis focuses on analyzing the movie review extracted from online media which helps in understanding the public opinion about the product and utilizing it for future enhancement of business. Information Gain (IG) and principal component analysis (PCA) are used to reduce the dimensionality of different combination of feature vector and combines with learning algorithm to learn the sentiment classifier to improve the performance based on accuracy, precision and recall. In this paper, we consider three classification algorithms such as Support Vector, Naïve Bayes and Neural Network are used to provide accuracy for different combination of IG and PCA for different feature vector and find out which classifier provides good performance based on accuracy, precision, recall and F-Measure.

Key words: Sentiment Analysis, Neural Network Learning, Feature Selection, Feature Reduction, Support Vector Machine, Naïve Bayes

I. INTRODUCTION

Sentiment Analysis which is also called as opinion mining. Sentiment analysis is the study of sentiments, attitudes, reactions, evaluation of the content of the text. It is used to determine the attitude of the user based on polarity reviews. Many times while analyzing people's opinion, sentiment, attitude, reaction towards entities such as services, products, organizations, individuals, events, topics, issues and their attributes. Sentiment Analysis is used to classify the polarity of a given text in a document level, sentence level or feature/aspect level. Opinion mining focus on polarity classification and emotion recognition. Polarity classification also identify positive and negative expressions in online reviews and help make the product evaluation more credible. The basic task of sentiment analysis are removing URL, data pre- processing, feature extraction, feature reduction and learning the sentiment classifier using machine learning algorithm. The most common challenges in Sentiment analysis are word sense disambiguation. There is problem how effectively use sentiment words to improve performance of sentiment classification of sentence or documents. Several social networks allow us to collect data with Application Programming Language.

In addition to the main functionalities, the API allows developers to collect every kind of YouTube data. Sentiment analysis is used to determine the subjective attitude of the text. It is one of the special task in text mining. In bag of word model, a review is represented as vector of independent words. The statistical machine learning algorithm such as Naïve Bayes, Support Vector Machine and maximum entropy classifier is used to train the sentiment classifier. Polarity shift is one of the linguistic phenomenon to reverse the original polarity of text. Sentiment analysis gives less accuracy when compared to different combination of feature vector dataset.

In this paper, we propose a simple efficient model to analyze the efficiency of accuracy for different combination for feature vector such as unigram, bigram and trigram for three different classifier (SVM, Naïve Bayes, and Neural Network) and two feature selection algorithm. Sentiment classification has two review data set positive and negative review dataset. We first propose a data pre- processing techniques for each review dataset by creating bag of word model for unigram, bigram and trigram. Feature selection by Information Gain and Feature reduction by Principal Component Analysis are used to reduce the dimensionality of feature vector for a given review dataset and analyze which one gives best outcomes. Neural network learning, support vector and Naïve Bayes are used to learn the sentiment classifier and analyze the performance level for different combination of n gram Bag of word model to provide the best outcome.

The organization of this paper is as follows. Section II reviews the related work. In Section III, we present the proposed work. In Section IV, we analyze the performance level for different combination of unigram, bigram and trigram. In Section V, we draw a conclusion for the summary of the work.

II. RELATED WORK

We first summarize the work of sentiment analysis and then review the techniques to analyze the level of accuracy for different n gram feature vector. According to the level of granularity, Sentiment analysis is classified into different levels: document level, sentence level and aspect/ feature level.

While analyzing twitter sentiment analysis, kishori et al [1] discussed about the level of sentiment analysis. Review dataset are collected from the twitter, dataset are pre-processed and machine learning algorithm is used to improve the sentiment classifier at various level of sentiment. It cannot capture implicit sentiments. Overall result shows the less accuracy in sentiment classifier. Budhika and kasun [2] proposed the enhanced bag of word model is to ensure feature vector contains opinionated words in a textual data and improve the efficiency of the algorithm. Its efficiency is better than baseline algorithm but it provide insufficient local information to determine the polarity. Hyeoncheol et al [3], describes about sentiment analysis on social network using probability model. It does not support polarity shifting problem. It still needs to improve performance based on accuracy, precision and recall. Pre-processing of text done by human coders.

Erik cambria et al [6] describes about the new avenues in opinion mining and sentiment analysis to improve customer relationship management through positive and negative feedback. Many new areas might be useful in sentiment analysis such as facial expression, body moment or a video blogger choice of music or color filter. It still needs to understand the better understanding of natural language processing. While analyzing the different feature selection algorithm on different sentiment lexicons in sentiment analysis, anuj et al [7] investigates the performance based on accuracy, precision and recall. It clearly represents which feature selection provides the better performance in sentiment analysis while learning the sentiment classifier.

Wiltrud et al [8] proposed the classification of inconsistent sentiment words using syntactic construction. It improves the performance for inconsistency classifier by syntactic construction. It still needs to improve quality and coverage of the automatic annotation which are available in large quantities of data. Chandrakala and sindhu [9] proposed the survey of opinion mining and sentiment classification. It gives the best outcome of analyzing the various machines learning algorithm to improve the performance of the sentiment analysis. It does not support polarity shifting problem. Nidhi Mishra [10] describes about the classification of opinion mining techniques. Opinion mining is useful for detecting and removal of fake opinions. It does not support polarity shifting problems. It does not support hidden product features. It does not deal with negative expression.

In this paper, we improve the performance for different combination of feature vector based on precision, accuracy and recall by using feature selection algorithm and neural network learning to learn the sentiment classifier.

III. PROPOSED WORK

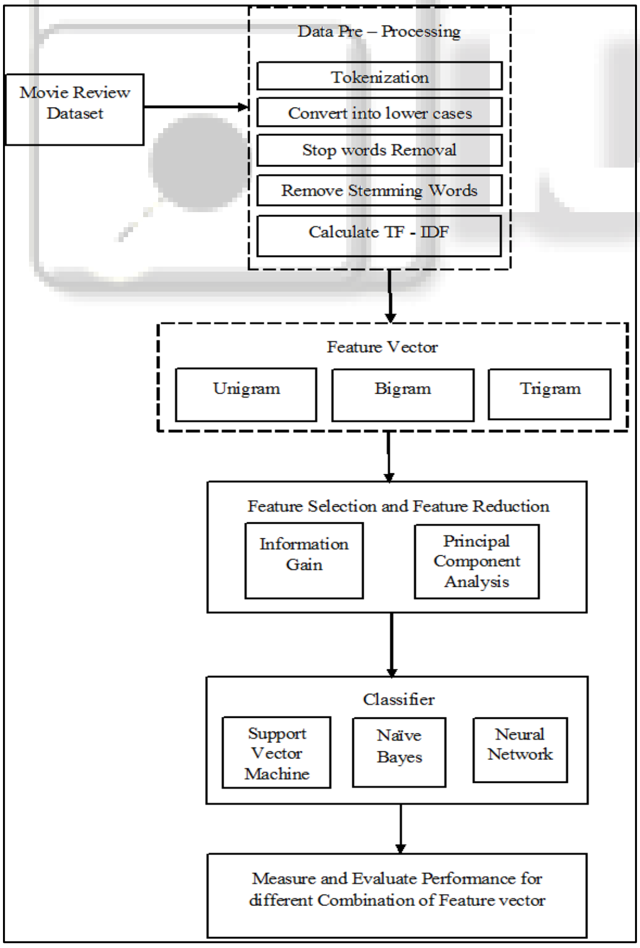

Fig. 1: Overall Architecture Diagram

The overall design is implemented to identify the possible work flows and process execution of the system. Movie review dataset are collected from the amazon product reviews. The original dataset are pre- processed to remove irrelevant data from the review dataset. Feature selection and Feature reduction are used to reduce the dimensionality for different combination of unigram, bigram and trigram of feature vector. Neural network learning, Support Vector and Naïve Bayes are used to learn the sentiment classifier by tenfold cross validation. Then, Measure and validate the effectiveness of different outcomes.

A. Data Pre- Processing

Review Dataset such as movie, DVD are collected from amazon product review. The process of data pre- processing are tokenization, convert word into lower cases, stop word removal, remove stemming words and calculate term frequency and inverse document frequency and construct n gram feature vector by unigram, bigram and trigram. First, the given sentence is tokenized into words by removing URLs, Tweets, Punctuation and special character. Second, stop words are removed such as a, an, the, of, as, before, aside etc. in given tokenized dataset. Third, snowball stemmer is used to remove the stemming algorithm such as –ing, -ful, -ed etc. and finally, Term Frequency can be can be calculated as follows

$$Term\ frequency = 1 + \log(1 + \log(\mathrm{freq}(w, D_w)))\text{---- (1)}$$

Inverse Document Frequency can be calculated as follows.

$$IDF = \log(\frac{1+D}{D_w}) \quad \text{--------- (2)}$$

Where D – Total Number of Documents and D_t – Total Number of Terms t in the Document. After calculating Tf – Idf, construct unigram, bigram and trigram for a given featured dataset.

B. Feature selection and Feature Reduction

Feature Selection and Feature Reduction are used to reduce the dimensionality of feature vector. Feature selection can improve the result by reducing the interference brought by some useless feature. Feature reduction is finding a subspace which has less dimension that of original feature space.

1) Information Gain

Information Gain is a type of feature selection. The feature dataset can be given as input. The expected information need to classify review for review dataset R is known as entropy and is given by

$$Info(R) = -\sum_{i=1}^{m}(p_i \log_2(P_i))\text{----------------- (3)}$$

Where m represents the number of features in a review dataset and p_i represents the probability for classifying the review by positive or negative label. The amount of information that the class in bits require for each attributes measured in class label is given as follows

$$Info_w(R) = \sum_{i=1}^{m}(\frac{R_j}{R}) * Info(R) \quad \text{-------------- (4)}$$

Where R_j/R is the weight of the jth partition for each word for a given review dataset by classifying the positive and negative reviews and Info (R_j) is calculated using entropy formula. Information Gain for each word A is given as follows

$$Information\ Gain(w) = Info(R) - Info_w(R) \text{ -- (6)}$$

Select the features ranked as per the highest information gain score. The features below the threshold value are removed.

2) Principal Component Analysis

Principal component analysis is the widely used statistical method to reduce the dimension of feature set. Given X (n. m) matrix as the standardized word vector data with n reviews and m product attributes as input from data pre-processing. Calculate covariance matrix cov(X,Y) to represent how words in dataset are interrelated to each otherand is calculated as follows

$$Cov(X,Y) = {}^{1}/_{n-1} \sum (X - X')(Y - Y') \text{ -------- } (7)$$

Where X' and Y' are the mean vectors and n be the dimensional word vector. Calculate Eigen values and Eigen vectors for covariance matrix and is calculated as follows

$$|Cov(X,Y) - \lambda I| = 0 \text{ ---------- } (8)$$

Eqn (8) shows how to find Eigen values and calculate Eigen vector from Eigen values. The final result is an n × p matrix of domain features. The feature can be reduced from the original pre- processed dataset based on setting up of variance covered.

C. *Classification Algorithm*

Classification process includes two steps: Training Step and Testing Step. The classifier is built from training review dataset and their associated class label positive and negative. Classification algorithm is used to estimate the performance level based on accuracy, precision, recall and F- Measure.

1) Neural Network Learning

Back propagation learns by iteratively processing a dataset of training data, comparing the networks prediction for each data with the actual known target value. For each training dataset, the weights are modified so as to minimize the mean squared error between the network prediction and the actual target value.

Feature vector X= {X_1, X_2.............. Xn}, weight W_{ij} calculated by information gain and principal components and their associated class label positive and negative as input. Setting up the value for learning rate l, Momentum **θ** and hidden layer h. calculate the input and output unit for hidden layer and output layer is calculated as follows

$$Input_i = \sum W_{ij} * Output_j \text{ ---------- } (8)$$

$$Output_i = {}^{1}/_{1 + e^{-Input_j}} \text{ ---------- } (9)$$

The error is propagated backward from hidden layers and output layer by updating the weights and biases to reflect the error of the networks prediction. The error for output layer is computed as follows

$$Err_j = Output_j(1 - output_j)(T_j - Output_j) \text{ -- } (10)$$

The Error for hidden layer is computed by summing the error and weight from the output layer as follows

$$Err_j = Output_j(1 - Output_j) \sum Err_k W_{jk} \text{----- } (11)$$

The weights and biases are updated to reflect the propagated errors. Weights are updated as follows

$$\Delta w_{ij} = (l) Err_j Output_i \text{ ---------------- } (12)$$

$$w_{ij} = w_{ij} + \Delta w_{ij} \text{ --------------- } (13)$$

When weights are updated, set learning rate l be 0.3. Momentum are updated based on the learning rate is calculated as follows

$$\Delta\theta_j = (l) Err_j \text{ --------------- } (14)$$

$$\theta_j = \theta_j + \Delta\theta_j \text{ -------------- } (15)$$

Weights and bias are updated until the calculated weight is smaller than the previous weight.

2) Naïve Bayes

Naïve Bayes are statistical classifier. They can predict class membership probabilities. Let R be the movie review dataset with class label positive and negative. Each review dataset R is represented by an n- dimensional feature vector, X={X_1, X_2 ... X_n} as input.

Calculate probability of each class label negative and positive, P(C_i) can be computed based on the training tuples as follows

$$P\left(C_{i=p}\right) = \frac{Number\ of\ positive\ words}{Total\ number\ of\ words} \text{ -------- } (16)$$

$$P\left(C_{i=n}\right) = \frac{Number\ of\ negative\ words}{Total\ number\ of\ words} \text{ -------- } (17)$$

Calculate conditional probabilities P(X|Ci) for each feature vector to maximize the P(X|C_i) P(C_i)

$$P(X|C_p) = P(X_1|C_p) * P(X_2|C_p) *..* P(X_i|C_p) \text{ -------}(18)$$

$$P(X_i|C_n) = P(X_1|C_n) * P(X_2|C_n) *..* P(X_i|C_n) \text{ ------}(19)$$

Naïve Bayes Classifier predicts class label for each feature vector by maximizing P(X|C_{pos})*P(C_{pos}) and P(X|C_{neg})*P(C_{neg}).

3) Support Vector Machine

Support Vector Machine is a new method for classifying for both linear and nonlinear data. Let R be the movie review dataset with class label positive and negative. Each review dataset is represented by non-dimensional feature vector X={X_1, X_2, X3, ... , X_n} and Y={ +1, -1}. +1 represents positive label and -1 represents negative label.

For positive class label, W_0 +W_1X_1+W_2 X_2>1

For negative class label, W_0 +W_1X_1+W_2 X_2<1

Combining positive and negative class label,

+1/1(W_0+W_1X_1+W_2X2 +................+W_nX_n)>1 ------- (20)

If any training words in movie review dataset satisfy the above condition then that training tuples are called support vectors. To calculate the Euclidian distance of W by sqrt (W.W). Based on a lagrangian multiplier, calculate maximal marginal hyperplane as follows

$$d(X') = \sum_{i=1}^{l} Y_i a_i X_i X' + b_i \text{ ------------- } (21)$$

X_i - set of Training words in movie review dataset, Yi-class label (pos/neg) of support vector, X'- Testing set of Movie Review Dataset and a_i, b_i parameter that are determined automatically by optimized support vector.

If d(X') sign is positive, then X' falls on or above Maximal Marginal Hyper plane and so the SVM predicts that X' belongs to positive class otherwise X' belongs to negative class.

D. *Measure and validate the Performance*

After learning the sentiment classifier, measure and validate the performance for the review dataset. Tenfold cross validation is performed to cross validate the review dataset. For 100 dataset, 90 dataset for training dataset and remaining dataset for test dataset in 100 iteration. Confusion matrix predicts how prediction on instances are tabulated. From confusion matrix, measure the performance based on accuracy, precision and recall.

		Actual value	
		Positive	Negative
Predicted Value	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 1: Confusion Matrix

Accuracy is defined as proportion of total number of predictions were correct is calculated as follows

$$accuracy = \frac{Tp+TN}{TP+TN+FP+FN} \text{ -------- } (20)$$

Precision is defined as proportion of predicted positive cases and is calculated as follows

$$precision = \frac{TP}{TP+FP} \text{ ----------------- } (21)$$

Recall is defined as proportion of positive cases that were correctly identified and is calculated as follows

$$recall = \frac{TP}{TP+TN} \text{ ----------------------- } (22)$$

F- Measure is defined as proportion of precision and recall of correctly identified instances

$$F_{Measure} = \frac{2*precision*recall}{precision+recall} \text{--------------- } (23)$$

IV. PERFORMANCE ANALYSIS

In this Section, we experimentally evaluate our approach on polarity classification on movie review datasets, three classification algorithm, feature selection and feature reduction algorithm.

A. Data Source

For Polarity Classification, we use movie review dataset taken from Amazon.com. Each of the review rated by the customers from star1 to star 5. The review with star1 and star 2 are labelled as negative and the review with star3, star 4 and star5 are labelled as positive. Review dataset contain 1000 positive review and 1000 negative review. In our experiments, reviews in each category are randomly split up into ten folds (with 9 folds serving as training data and with 1 fold serving as test data). All of the following results are calculated in terms of an averaged accuracy of tenfold cross validation which can be simulated using Weka tool.

Data set	Model	No. of Review	+ve Review	-ve Review	Feature Type
Movie Review Data Set	M – I	2000	1000	1000	Unigram
	M – II	2000	1000	1000	Unigram +Bigram
	M - III	2000	1000	1000	Unigram +Bigram +Trigram

Table 2: Properties of Data Source

B. Feature Selection and Reduction

Feature Selection is used to reduce the dimensionality of feature vector. Using Weka tool, feature selection such as Information Gain, Principal Component Analysis can be carried out for different combinations of unigram, bigram and trigram to analyze which combination gives the reduced feature vector. Information Gain is best among two feature selection algorithm for different combination of unigram, bigram and trigram. Table 3 shows the reduced featured set for PCA by setting variance covered 0.4 and threshold value 0.002 and IG by setting threshold value 0.002. When IG compared to PCA, IG shows slightly reduced from PCA feature vector.

No. of Features	Model – I	Model – II	Model – III
Without IG/ PCA	1160	1138	1144
With IG	168	153	150
With PCA	184	169	164

Table 3: No. of Features in Movie Review Dataset

C. Measure and Analyze the Performance

Measure and Analyze the performance based on accuracy, precision, recall and F-Measure for different combination of unigram, bigram and trigram using Neural Network Learning Classifier with two different feature selection algorithm (Information Gain and Principal Component Analysis).

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	81.3	79.5	79.65
Precision	81.3	79.6	79.7
Recall	81.3	79.6	79.7
F- Measure	81.3	79.5	79.6

Table 4: Performance (%) Level of Movie Review dataset using PCA + Naïve Bayes

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	83.05	82.2	80.4
Precision	83.1	82.2	80.4
Recall	83.1	82.2	80.4
F- Measure	83	82.2	80.4

Table 5: Performance (%) level of Movie Review using PCA + Neural Network Learning

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	83.85	83.55	82.55
Precision	83.9	83.6	82.6
Recall	83.9	83.6	82.6
F- Measure	83.8	83.5	82.5

Table 6: Performance (%) level of Movie Review Dataset using PCA + Support Vector

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	79.6	81.2	81.75
Precision	79.6	81.2	81.8
Recall	79.6	81.2	81.8
F- Measure	79.6	81.2	81.7

Table 7: Performance (%) level of Movie Review Dataset using IG + Naïve Bayes

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	78.45	80.3	80.4
Precision	79.5	80.5	80.4
Recall	78.5	80.3	80.4
F- Measure	78.3	80.3	80.4

Table 8: Performance (%) level of Movie Review Dataset using IG + Neural Network

Model/ Performance	M- 1	M- 2	M- 3
Accuracy	84.15	81.95	83.65
Precision	84.2	82	83.7
Recall	84.2	82	83.7
F- Measure	84.1	81.9	83.6

Table 9: Performance (%) level of Movie Review Dataset using IG + Support Vector

Analyzing Table 1, Table 2, and Table 3, Table 3 shows better performance when PCA compared to SVM of 83.85% for the combination of unigram feature vector than other classifiers (Neural Network and Naïve Bayes). Table 4, Table 5 and Table 6 shows the performance for different classifier (SVM, Naïve Bayes and Neural Network) for different combinations of unigram, bigram and trigram feature vector. Analyzing Table 4, table 5 and Table 6, Table 6 shows better performance when IG compared with SVM of 84.15% than other classifiers (Neural Network and

Naïve Bayes). Analyzing SVM compared with IG and PCA, IG with SVM shows higher performance of 84.15%.

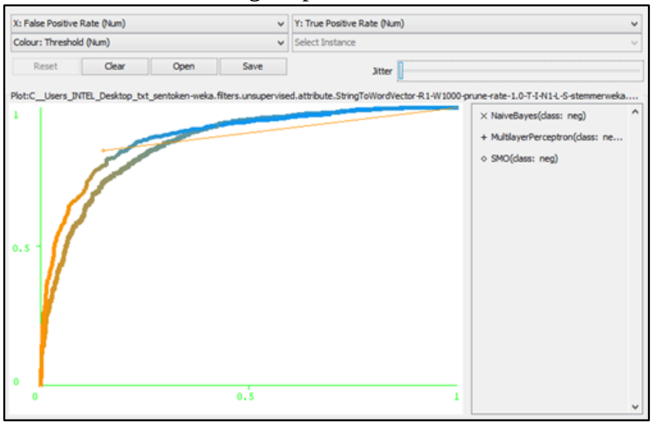

Fig. 1: Shows Performance level for IG with NN, SVM and NB. SVM shows Greater performance than other classifier. X axis denotes the False Positive Rate and Y denotes the True Positive Rate.

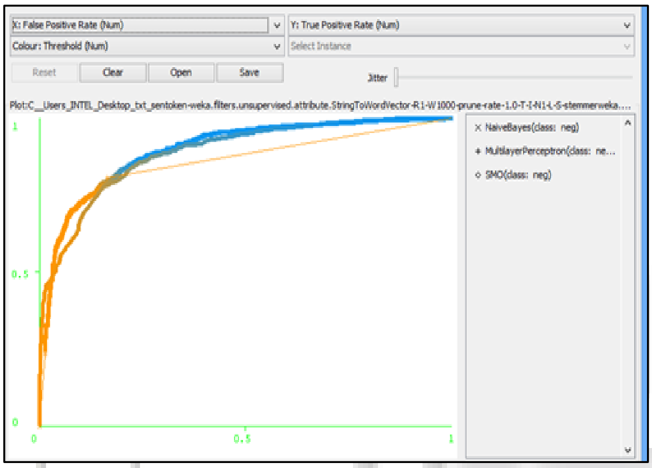

Fig. 2: Shows Performance level for PCA with NN, SVM and NB. SVM shows Greater performance than other classifier. X axis denotes the False Positive Rate and Y denotes the True Positive Rate.

V. Conclusion

In this proposed work, we analyze and compare the performance for various combination of unigram, bigram and trigram feature vector with three classifier (Support Vector Machine, Neural Network and Naïve Bayes) and Feature selection (IG) and Feature Reduction (PCA). Support Vector Machine show better performance than other two classification algorithm (Naïve Bayes and Neural Network). Feature Selection (IG) and Feature reduction (PCA) are used to reduce the dimensionality of feature vector for large dataset. It doesn't fit for small dataset. IG provides better performance than PCA. In addition to this work, we further use the movie review dataset by comparing it with ensemble techniques to provide better result. It still needs to improve polarity shifting problem.

References

[1] Kishori K. Pawar, Pukhraj P. Shrishrimal, R. R. Deshmukh, "Twitter Sentiment Analysis: A Review", International Journal of Scientific and Engineering Research, 2015.

[2] Budhika, H. Kasthuriarachchy, Kasun De Zoysa and H . L. Premaratne, "Enhanced Bag of Word Model for phrase level sentiment analysis",2014

[3] Hyeoncheol Lee, Youngsub Han, Kwangmi Ko Kim, "Sentiment Analysis on Social Network Using Probability Model", the sixth international conference on Advances in Future Internet, 2014

[4] Pappu Rajan, S. P. Victor, "Web Sentiment Analysis for Scoring Positive and Negative words using Tweeter", International Journal of Computer Applications, 2014

[5] Bishan Yang, Claire Cardie, "Context aware Learning Sentiment Analysis with Posterior regularization@, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2014

[6] Erik Cambria, Bjourn Schuller, Yunqing Xia, Catherine Havasi, "New Avenues in Opinion Mining and Sentiment Analysis", knowledge based approaches to concept level sentiment analysis published by IEEE Computer Soceity, 2013

[7] Anuj Sharma, Shubhamoy Dey, "Performance Investigation of Feature Selection Methods and Sentiment Lexicons for Sentiment Analysis", International Journal of Computer Applications, 2012

[8] Wilrud Kesler, Henrich Schutze, "Classification of Inconsistent Sentiment Words Using Syntactiv Construction", Proceedings of COLING, 2012

[9] S. ChandraKala and C. Sindhu, "Opinion Mining and Sentiment Classification: A Survey", International journal of Soft Computing, 2012

[10] Nidhi Mishra, C. K Jha, "Classification of Opinion Mining Techniques", International Journal of Computer Applications, 2012

[11] Y. Choi and C. Cardie, "Learning with compositional semantics as structural inference for subsentential sentiment analysis," in Proc. Conf.EmpiricalMethodsNaturalLanguageProcess.,2008.

[12] E. Agirre, and D. Martinez, "Exploring automatic word sense dis- ambiguation with decision lists and the web," in Proc. COLING Workshop Semantic Annotation Intell. Content, 2000, pp. 11–19.

[13] M. Gamon, "Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analy- sis," in Proc. Int. Conf. Comput. Linguistics, 2004, pp. 841–847.

[14] T. Nakagawa, K. Inui, and S. Kurohashi. "Dependency tree-based sentiment classification using CRFs with hidden variables," in Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics, 2010, pp. 786–794. .

[15] Lette T. Li, V. Sindhwani, C. Ding, and Y Zhang, "Bridging domains with words: Opinion analysis with matrix tri-factorizations," in Proc. 10th SIAM Conf. Data Mining, 2010, pp. 293–302.

[16] A. McCallum and K. Nigam, "A comparison of event models for naive Bayes text classification," in Proc. AAAI Workshop Learn. Text Categorization, 1998, pp. 307–314.

[17] V. Ng, S. Dasgupta and S. Arifin, "Examining the role of linguistic knowledge sources in the automatic identification and classifica- tion of reviews," in Proc. Int. Conf. Comput. Linguistics Annu. Meet- ing Assoc. Comput. Linguistics, 2006, pp. 611–618.

[18] Z. Hai, K. Chang, J. Kim, and C. C. Yang, "Identifying features in opinion mining via intrinsic and extrinsic domain relevance," IEEE Trans. Knowl. Data Eng., vol. 26, no. 3, pp. 447–462, Mar. 2014.
[19] X. Yu, Y. Liu, X. Huang, and A. An, "Mining online reviews for predicting sales performance: A case study in the movie domain," IEEE Trans. Knowl. Data Eng., vol. 24, no. 4, pp. 720–734, Apr. 2012.
[20] M. Koppel and J. Schler, "Using neutral examples for learning polarity," in Proc. Int. Joint Conf. Artif. Intell., 2005, pp. 1616–1617.
[21] Rui Xia, Feng Xu, Chengqing Zong, Qianmu Li, Yong Qi, and Tao L, "Dual Sentiment Analysis : Considering Two Sides of One Reviews", IEEE Transaction in Knowledge and Data Engineering, vol : 27, No : 8, Augest 2015.
[22] G. Vinodhini, R. M. Chandrasekaran, "Opinion Mining Using Principal Component Analysis based Ensemble Model for e- commerce application", CSI Publications, November 2014.
[23] John C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines", Microsoft Research, Technical Report MSR- TR- 98-14, April 21, 1998.