ISSN 2348 – 8034
Impact Factor- 4.022

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

IMPROVED FEATURE SELECTION FORCLASSIFICATION OF SENTIMENTAL REVIEWS USING N-GRAM MACHINE LEARNING APPROACH

Amandeep Singh*[1] & Dr. Dinesh Kumar[2]
[*1]M.Tech CSE, DAVIET, Jalandhar, India
[2]Associate Professor & Head IT, DAVIET, Jalandhar

ABSTRACT

The sentiment analysis approach is used to determine the sentiment in the text content by using the keyword intensity or term frequency based approach. The keyword extraction models are used to determine the words containing the sentiment from the text data, and eliminate the remaining content based upon the selection or design of the feature extraction model. The keywords based features are then transformed to the numeric formation by using the ratio, weight or appearance based description, and further classified using the supervised classification model to identify its orientation. In this paper, the supervised machine learning approach combines the count vectorization and TF-IDF based features with Chi-square based feature selection for sentiment analysis in the IMDB review database. The proposed feature description model combines the various N-gram features, such as unigram, bigram and trigram, which signify the different aspects of sentiment contained in the text data. The proposed model has outperformed the existing model based upon the layered model using a count based method with TF-IDF. Support Vector Machine (SVM) classification method is considered as the best method after the result evaluation with the proposed feature descriptor.

Keywords: *Sentiment analysis, TF-IDF, SVM, Naïve Bayes, n-gram.*

I. INTRODUCTION

Social media Analytics is the method of collecting data from different user's communication on various social sites and blogs and then calculating that data to derive the decision related to business. Social media is a correct way to know about real-time choices of customers, wish one aims and emotions. The social media analytics can also be referred as media of listening or the media which is intelligent.

Sentiment Analysis is the process which determines whether a piece of text is positive, negative or neutral. As more and more people use internet these days, sentiment analysis has become very useful tool to find out public opinion about certain topic though social media. Different machine learning techniques are used for the process of classification. Sentiment analysis is carried out in three different levels such as document level, sentence level, and aspect level. Document level sentiment analysis is done in this paper.

Lot of people use internet to express their views about a certain movie on various social platforms like IMDB. These reviews are mostly unstructured and in the text. Thus, the stop words and other unwanted information are removed from the reviews before further analysis. These reviews goes through a process of vectorization in which, the text data are converted into matrix of numbers. These matrices are then given input to different machine learning techniques for classification of the reviews. Different parameters are then used to evaluate the performance of the machine learning algorithms.

The main contributions of this paper are as follows:

- **i.** Improved feature extraction method is proposed for the classification of movie reviews of IMDb dataset IMDb (2011) using n-gram techniques.
- **ii.** Best feature selection is done using Chi Square test to select only the best features.
- **iii.** Machine learning techniques used are Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD) are used for classification purpose using the n-gram approach.
- **iv.** The performance of existing and proposed feature selection method is compared using parameters like

 ISSN 2348 – 8034
Impact Factor- 4.022

precision, recall, f-measure, andaccuracy. The results indicate that accuracy of proposed feature selection method is more than existing method.

II. LITERATURE SURVEY

Four different machine learning algorithms such as Naive Bayes (NB), Maximum Entropy (ME), Stochastic Gradient Descent (SGD), and Support Vector Machine (SVM), which have been considered for classification of human sentiments [13]. The accuracy of different methods is critically examined in order to access their performance on the basis of parameters such as precision, recall, f-measure, and accuracy.Sentence Compression for Aspect-Based Sentiment Analysisis a framework in whichsentiment sentence compression (Sent Comp) is done before performing sentimental analysis[1]. Sent Comp is different from previous sentence compression methods as it only removes words which do not hold any sentimental value and are not necessary while determining sentiment of the sentence. Thus this method creates a shorter sentence which is easier to analyze.They have applied a discriminative conditional random field model to automatically compress sentiment sentences. Using the Chinese corpora of four product domains, Sent Comp significantly improves the performance of the aspect-based sentiment analysis.

Feature-based opinion summarization is performed in two steps [6]: Identify the features of the product that customers have expressed opinions on (called opinion features) and rank the features according to their frequencies that they appear in the reviews. They analyze number of positive and negative reviews are giver by the customers for each feature. The specific reviews that express these opinions are attached to the feature. This facilitates browsing of the reviews by potential customers. Opinion summarization system performs the summarization in two main steps: feature extraction and opinion direction identification. The inputs to the system are product name and an entry page for all the reviews of the product. The output is the summary of the reviews. Given the inputs, the system first downloads (or crawls) all the reviews, and puts them in the review database.Frequent Pattern Mining Algorithm for Feature Extraction of Customer Reviews [10] includes the following steps: 1) Pre-processing step includes stop word removal and word stemming; 2) POS tagging includes frequent feature identification; 3) Mining frequent patterns lead to potential features; 4) Pruning includes compactness pruning and redundancy pruning and thus frequent features will be extracted. Finally summary can be made including the sentences which contain potential features.

III. EXPERIMENTAL DESIGN

3.1. Feature selection

CountVectorizer and TF-IDF are two methods used for feature selection:

- **CountVectorizer:** It is a process which converts the text data into a matrix which contains count of each word in every document. A matrix is generated which contains number of rows equal to number of documents in the database and number of columns are equal to numbers of unique words in the database.
- **n-gram** is a contiguous sequence of n items from a given sample of text or speech. CountVectorizer creates matrix using n-gram range for example: unigram, bigram, trigram, unigram + bigram, bigram + trigram and unigram + bigram + trigram.
- **TF-IDF (TermfrequencyandInversedocumentFrequency):** TF-IDFconsistoftwo words:thefirst isusedtofind thenormalizedTermFrequency(TF), thenumberoftimes the word appears in a document, divided bythe total number of words that comes in document;thesecondtermis InverseDocumentFrequency (IDF),whichiscalculatedasthe logarithmofthe totalnumberofthedocumentsintheblogdividedbythenumberof documents wherethe particular term appears.

Term Frequency:

$$TF(t) = nq/nw \qquad (1)$$

Where:

nq = Number oftimes word q comesin a document

nw = Total numberofwordsin the document.

Inverse document Frequency:

ISSN 2348 – 8034
Impact Factor- 4.022

IDF (t) =loge(tn / tnq) (2)
Where:
tn = Total number ofdocuments
tnq = Number ofdocuments with word q in it

List of Functions:

- **Stop_words:** It is an in build library. stop_word = "english" will filter all words from the data except words in English library.
- **ExtractTFIDF()**: Extracts TF-IDF features from message.
- **countVectorizer():** Extracts count based features from message.
- **transformCV2TF():** Converts count based features to transform count based features to term frequency based features.
- **addFeature():** Adds features to feature matrix.

Algorithm CV_TFIDF (message)

CountVectorizer and TF-IDF based feature selection using n-gram range

1. Set n-gram range, ngram_range← (x.y)
2. Run the iteration over every message in the given dataset
   a. Remove stop words,
      stop_words= "english"
   b. Initialize label vector
   c. Extract the TF-IDF features from the target message, F1 ←ExtractTFIDF(message)
   d. Extract the count based features from the target message, C ←countVectorizer(message)
   e. Apply the TF-transformation of the count based features to transform count based features to term frequency based features,
      F2 ← transformCV2TF(C)
   f. Add the feature 1 to the feature description matrix of first feature, F1_mat ←addFeature(F1)
   g. Add the feature 2 to the feature description matrix of second feature, F2_mat ←addFeature(F2)
   h. Update the label vector for the message
3. Combine F1_mat and F2_mat

3.2 Best feature selection using chi square test

The feature selection method is based upon the Chi square test of the data, which test the validation of every column individually, which undergoes the threshold calculation. The cutoff threshold is used to determine the best features in the given data by selecting the best compatible columns from the target data. The chi square test works in the case with N data rows and multiple classes, usually determined as positive, negative and neutral tweets. Each of the column in the given dataset is considered as a features, and feature selection is all about selecting the most prominent columns or features in the given feature matrix. Technically, all of the features in the feature descriptor matrix are not distributed with equal significance.

The selection of the high significance features from the feature matrix is used to boost the classification accuracy. The chi square test is applied to each feature or column in the feature matrix, where the significance of each feature is computed by observing the expected and observed counts. The chi square test evaluates the derivative of expected counts (E) over the observed counts (O). The chi square score computes the independence of each class in the given feature, which is computed by derivative (x^2). If derivative (x^2) is observed with high value, the hypothesis is considered incorrect. The hypothesis is observed correct while the small value is computed chi square test.

The table 3.1 shows the observation table computed by the chi square test function:

ISSN 2348 – 8034
DOI: 10.5281/zenodo.1313639
Impact Factor- 4.022

Table 1: Chi square test evaluation over positive and negative classes

	Class I	Class II	Total values
Occurrence of feature X	Ax	Bx	Ax + Bx = Mx
Absence of feature X	Cx	Dx	Cx + Dx = Nx -Mx
Total values	Ax + Cx = Px	Bx + Dx = Nx - Px	

Where Ax denotes the data rows belonging to class I and Bx denotes the data rows of class II. On the other hand, the contradicting features are denoted by Cx and Dx for both of the classes. The total feature belonging to class I are computed by adding Ax and Bx values, given Mx = Ax .+Bx, and Nx-Mx = Cx + Dx contains all of the contradictive features, which doesn't belong to class I. Similarly, Px = Ax + Cx denotes the total features belonging to class II, and Nx – Px gives all non-independent features, which belong to class II.

List OfFuctions:

- **feature_selection:** selects best feature from feature matrix.
- **CombineFeatures():** Combines feature from two matrix.
- **train_test_split():** Splits features for testing and training.

Algorithm CHI_SQ (feature_matrix)

Best feature selection using Chi square test

1. Apply the best feature selection policy using Chi square test over both of the feature matrices
   a. Select the best features from the feature 1 matrix, F1fs ←feature_selection(F1_mat)
   b. Select the best features from the feature 2 matrix, F2fs ←feature_selection(F2_mat)
2. Combine both of the feature matrices to amalgamate the features, CF ←CombineFeatures(F1fs, F2fs)
3. Split the features in the training and testing data using the random feature splitter, [X_train, X_test, y_train, y_test] ←train_test_split(CF, labels)

3.3 Classification model application

The application of the classification models is targeted to discover the emotion or sentiment in the given text data. The machine learning algorithms that are used for the text classification are Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM), and Stochastic Gradient Descent (SGD). Also, the algorithm SEN_ANALYSIS() involve the overall flow of the data processing under the proposed model, which involves a number of methods together to classify sentiment. Figure 1 shows the overall workflow of the classification model.

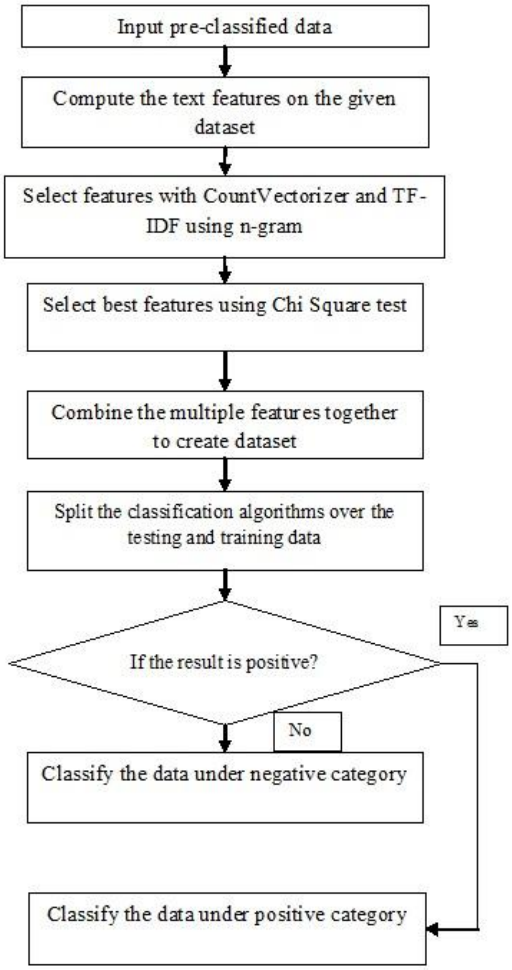

Figure 1: Workflow of the supervised classification.

List of Fuctions:

- **evalautePerformance():** Evaluates performance of the algorithm.

Algorithm SEN_ANALYSIS

1. Acquire the data from the local storage to the runtime memory
2. Synthesize the dataset, and extract text data from the input user data
3. Feature selection using **Algorithm CV_TFIDF (text data)**
4. Best feature selection using **Algorithm CHI_SQ (feature_matrix)**
5. Select the target classifier algorithm
6. Train the classifier model with training data, i.e. X_train and y_train
7. Test the classifier model with test data, i.e. X_test, and return the predictions in the array y_preds
8. Evaluate the performance of the classification algorithm, parameters ←evalautePerformance(y_test, y_preds)
9. Return the performance parameters

3.4 Performance parameters

The statistical parameters to measure the statistical errors (Type 1 and Type 2) are measured in order to evaluate the overall performance of the proposed model by evaluating the samples by the means of the programming or the manual binary classification. The proposed model evaluation is entirely based upon this statistical analysis. The following table 2 explains the significance of the type 1 and type 2 statistical errors for the evaluation of the hypothesis.

Table 2: Type 1 and type 2 statistical error

	Doesn't contain the target object or condition	Contain the target object or condition
Tests Negative or Accepted Null Condition	True Negative	False Negative
Tests Positive or Rejected Null Condition	False Positive	True Positive

i. True Positive

The true positive is when the final condition marked as matching and correct, which shows the positive condition and denies the null hypothesis. True positive is given with the symbol A. The true positive is given as the following:

$$TP = n_{11} = \text{number of such individuals} \qquad (3)$$

ii. True Negative

The true negative is when the final condition marked as non-matching and correct, which shows the negative condition and accepts the null hypothesis. True negative is given with the symbol B. The true positive is given as the following:

$$TN = n_{00} = \text{number of such individuals} \qquad (4)$$

iii. False Positive

The false positive is when the final condition marked as matching and incorrect, which shows the positive condition and denies the null hypothesis. False positive is given with the symbol C. The false positive is given as the following:

$$FP = n_{01} = \text{number of such individuals} \qquad (5)$$

iv. False Negative

The false negative is when the final condition marked as non-matching and incorrect, which shows the negative condition and accepts the null hypothesis. False negative is given with the symbol D. The false negative is given as the following:

$$FN = n_{10} = \text{number of such individuals} \qquad (6)$$

v. Recall

Recall is the test of the probability of the accuracy, which indicates the performance of the proposed model in the presence of the false negative cases. The false negative cases depict the falsely detected case from the data entries. In recall, the accuracy of the proposed model has been analyzed in the presence of false negative cases:

$$Recall \coloneqq \frac{True\ Positive}{True\ Positive\ + False\ Negative} \qquad (7)$$

vi. Precision

The precision depicts the accuracy of the model in the presence of the false positive cases. The accuracy of the model depicts the overall impact of the false positive cases, which rejects positive cases. A positive case in our case is when the data entry contains the certain set of parameters from one of the registered category, but returns the false result for such entries.

$$Precision \coloneqq \frac{True\ Positive}{True\ Positive\ + False\ Positive} \qquad (8)$$

v. F1-Measure

The F1-Measure is the cumulative parameter to assess the overall impact of the precision and recall in the case to study the overall impact of the false positive and false negative cases over the overall accuracy assessed from the preliminary statistical parameters. The F1-score value is represented in the range of 0 to 1 or 0 to 100, decided as per the maximum ranges of the precision and recall. The following equation is utilized to measure the F1-measure:

$$F1 - Measure \coloneqq 2 * \frac{(R * p)}{R + P} \qquad (9)$$

Where R is recall, and p or P is precision.

ISSN 2348 – 8034
Impact Factor- 4.022

vii. Accuracy

The overall accuracy is the analysis of the proposed model in the terms of overall accuracy, which is computed by dividing the total number of true cases (including true negative and true positive), by all of the cases.

$$Accuracy := \frac{True\ Positive\ +\ True\ Negative}{True\ Positive\ +\ True\ Negative\ +False\ Positive\ +False\ Negative} \quad (8)$$

3.5 Dataset used

The Internet Movie Database (IMDb) dataset is considered for sentiment analysis IMDb (2011). It consists of 12,500 positively labeled test reviews, and 12,500 positively labeled train reviews. Similarly, there are 12,500 negative labeled test reviews, 12,500 negative labeled train reviews. Apart from labeled supervised data, an unsupervised dataset is also present with 50,000 unlabeled reviews.

IV. RESULT ANALYSIS

The proposed model has been designed for the classification of text data using the sentiment analytics, which has been performed using multiple features. The features of count vectorization and term frequency inverse document frequency (TF-IDF) are extracted from the text data obtained from IMDB review database, which are combined in the layered fashion under the proposed model. The proposed model utilizes various types of classification algorithms, which includes Naive Bayes, Support Vector Machine (SVM), Bagging classification

and voting classification. The proposed and existing model both are evaluated using the similar classification algorithms with the connected features.

The overall analysis of all of the classification models is conducted over the different feature combination in existing and proposed model feature description models. Table 3 shows the evaluation parameter and accuracy for Naïve Bayes n-gram classifier with both existing and proposed feature selection. The performance of Naïve Bayes classifier is better with proposed method of feature selection with every n-gram range. Table 4 shows result of all performance parameters for Support Vector Machine (SVM) n-gram classifier with both existing and proposed parameters. SVM classifier also performs better with proposed feature selection method and all performance parameters show improvement in results with every n-gram range. Table 5 shows result of performance parameters for Maximum Entropy (ME) n-gram classifier. This classifier shows best results with proposed feature selection method when n-gram range is unigram + bigram + trigram. Table 6 shows result of performance parameters for Stochastic Gradient Descent (SGD) n-gram classifier. Stochastic Gradient Descent classifier shows great improvement with proposed feature selection method and performs best with bigram.

Table 7 compares accuracy of all four classifiers used in this paper using n-gram with both existing and proposed feature selection method. This table is represented in form of a graph in figure 2. Support Vector Machine algorithm achieved the highest accuracy of 91.5% with n-gram range as bigram and proposed feature selection method.

Table 3: Evaluation parameter and accuracy for Naïve Bayes (NB) n-gram classifier with existing and proposed feature selection

Method	Existing Features				Proposed Features			
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Unigram	0.89	0.88	0.90	89.5	0.92	0.89	0.91	90.5
Bigram	0.89	0.89	0.88	89	0.92	0.90	0.91	91
Trigram	0.87	0.85	0.86	86.8	0.92	0.87	0.89	89.5
Unigram + Bigram	0.90	0.85	0.87	88.5	0.91	0.86	0.87	89

Bigram + Trigram	0.88	0.84	0.88	86	0.89	0.86	0.88	88
Unigram + Bigram + Trigram	0.89	0.85	0.88	87.5	0.90	0.87	0.88	88

Table 4: Evaluation parameter and accuracy for Support Vector Machine (SVM) n-gram classifier with existing and proposed feature selection

Method	Existing Features				Proposed Features			
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Unigram	0.89	0.82	0.85	84	0.93	0.90	0.91	91
Bigram	090	0.84	0.86	85.5	0.93	0.9	0.91	91.5
Trigram	0.88	0.84	0.83	84.5	0.89	0.85	0.89	88.5
Unigram + Bigram	0.89	0.86	0.87	86.5	0.91	0.89	0.87	90
Bigram + Trigram	0.88	0.82	0.84	84	0.89	0.86	0.88	87.5
Unigram + Bigram + Trigram	0.87	0.86	0.87	85	0.89	0.86	0.88	87

Table 5: Evaluation parameter and accuracy Maximum Entropy (ME) n-gram classifier with existing and proposed feature selection

Method	Existing Features				Proposed Features			
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Unigram	0.88	0.84	0.86	85	0.88	0.86	0.89	86.5
Bigram	0.84	0.88	0.86	86.5	0.82	0.84	0.86	84
Trigram	0.85	0.84	0.84	83.5	0.84	0.74	0.76	79
Unigram + Bigram	0.87	0.84	0.85	85.5	0.88	0.84	0.87	86
Bigram + Trigram	0.85	0.78	0.81	82	0.80	0.77	0.78	78.5
Unigram + Bigram + Trigram	0.88	0.84	0.86	86	0.90	0.87	0.89	89

Table 6: Evaluation parameter and accuracy for Stochastic Gradient Descent (SGD) n-gram classifier with existing and proposed feature selection

Method	Existing Features				Proposed Features			
	Precision	Recall	F-measure	Accuracy	Precision	Recall	F-measure	Accuracy
Unigram	0.72	0.97	0.82	84	0.92	0.90	0.91	90
Bigram	0.76	0.96	0.78	81	0.91	0.90	0.88	90.5
Trigram	0.87	0.78	0.82	82.5	0.86	0.84	0.88	86
Unigram + Bigram	0.88	0.83	0.84	83.5	0.91	0.88	0.89	89.5
Bigram + Trigram	0.85	0.83	0.84	83	0.86	0.84	0.86	84.5
Unigram + Bigram + Trigram	0.86	0.82	0.85	83	0.88	0.83	0.85	85

[Singh, 5(7): July 2018] ISSN 2348 – 8034
DOI: 10.5281/zenodo.1313639 Impact Factor- 4.022

Table 7: Comparative result of accuracy of all the algorithms obtained using n-gram and existing and proposed feature selection

Classifier	Features	Unigram	Bigram	Trigram	Uni+Bi	Bi+Tri	Uni+Bi+Tri
Naïve Bayes	Existing	89.5	89	86.8	88.5	86	87
	Proposed	90.5	91	89.5	89	88	88
Support Vector Machine	Existing	84	85.5	84.5	86.5	84	85
	Proposed	91	91.5	88.5	90	87.5	87
Maximum Entropy	Existing	85	86.5	83.5	85.5	82	86
	Proposed	86.5	84	79	86	78.5	89
Stochastic Gradient Decent	Existing	84	81	82.5	83.5	83	83
	Proposed	90	90.5	86	89.5	84.5	85

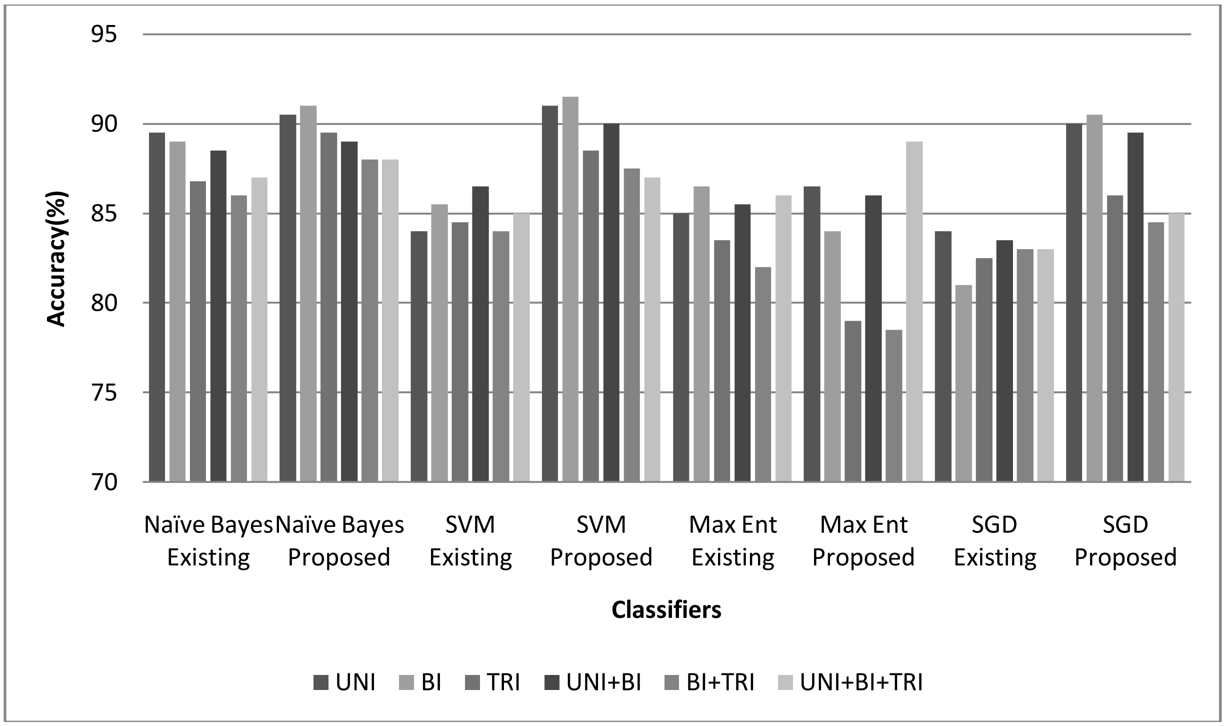

Figure 2: Comparative result ofaccuracy of all the algorithms obtained using n-gram and existing and proposed feature selection.

V. CONCLUSIONS AND FUTURE WORK

In this paper, the work is carried upon the review classification in the IMDB database, which is the customer database containing the reviews about the different movies. All of the these reviews are analyzed for the sentiment hidden in the text, which is detected by using the proposed model feature description model along with feature selection methods, which are employed along with different classification models. The feature description method is based upon the term count ratio based upon count vectorization along with term frequency& inverse document frequency (TF-IDF). Both of the features are combined in horizontal fashion in order to extend the final feature, which finally undergoes the selection based upon the Chi-square test to validate the features with high significance. The proposed feature description method is based upon the independent or various combinations of unigram, bigram and trigrams, which signifies the N-gram approach of this sentiment classification algorithm. The final features are classified using the supervised classification approaches such as SVM, Naïve Bayes, Voting and Bagging classification models. The proposed feature extraction method has outperformed the existing method, which included TF-IDF and count vectorization in layered arrangement. For the proposed feature extraction, theSupport Vector Machineclassifieris observed as best classification method with highest accuracy of 91.5% with bigram.

In this paper, we used four algorithms for classification of movie reviews. Each classifier has some advantages and some disadvantages. So for future work, we can use combinations of these algorithms to create an ensemble based classifier which has good qualities of the classifiers used and does not have bad ones. This will increase the performance and we can achieve better results.

VI. ACKNOWLEDGEMENTS

We are extremely thankful to Dr. Manoj Kumar, Principal, DAV Institute of Engineering and Technology, Jalandhar, Punjab, India, for providing infrastructure facilities to work in, without which this research would not have been possible.

REFERENCES

1. *Che, Wanxiang, Yanyan Zhao, HongleiGuo, Zhong Su, and Ting Liu. "Sentence Compression for Aspect-Based Sentiment Analysis." Audio, Speech, and Language Processing, IEEE/ACM Transactions on 23, no. 12 (2015): 2111-2124.*
2. *Yazhini, R., and Raja P. Vishnu. "Automatic summarizer for mobile devices using sentence ranking measure."In Recent Trends in Information Technology (ICRTIT), 2014 International Conference on, pp. 1-6.IEEE, 2014.*
3. *Zha, Zheng-Jun, Jianxing Yu, Jinhui Tang, Meng Wang, and Tat-Seng Chua. "Product aspect ranking and its applications." Knowledge and Data Engineering, IEEE Transactions on 26, no. 5 (2014): 1211-1224.*
4. *Esuli, Andrea, and FabrizioSebastiani. "Sentiwordnet: A publicly available lexical resource for opinion mining." In Proceedings of LREC, vol. 6, pp. 417-422. 2006.*
5. *Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews."In Proceedings of the COLING/ACL on Main conference poster sessions, pp. 483-490.Association for Computational Linguistics, 2006.*
6. *Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews."InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177.ACM, 2004.*
7. *Jin, Jian, Ping Ji, and Ying Liu. "Translating online customer opinions into engineering characteristics in QFD: A probabilistic language analysis approach." Engineering Applications of Artificial Intelligence 41 (2015): 115-127.*
8. *Jin, Jian, Ping Ji, and Ying Liu. "Prioritising engineering characteristics based on customer online reviews for quality function deployment." Journal of Engineering Design 25, no. 7-9 (2014): 303-324.*
9. *Suanmali, Ladda, Naomie Salim, and Mohammed Salem Binwahlan. "Fuzzy logic based method for improving text summarization." arXiv preprint arXiv:0906.4690 (2009).*
10. *Ghorashi, Seyed Hamid, Roliana Ibrahim, Shirin Noekhah, and NiloufarSalehiDastjerdi. "A frequent pattern mining algorithm for feature extraction of customer reviews." In IJCSI International Journal of Computer Science Issues. 2012.*
11. *Kherwa, Pooja, Anish Sachdeva, Dhruv Mahajan, NishthaPande, and Praveen Kumar Singh. "An approach towards comprehensive sentimental data analysis and opinion mining."In Advance Computing Conference (IACC), 2014 IEEE International, pp. 606-612.IEEE, 2014.*
12. *Liu, Siyuan, Xiaoyin Cheng, and Fan Li. "TASC: Topic-Adaptive Sentiment Classification on Dynamic Tweets." (2015).*
13. *Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of sentiment reviews using n-gram machine learning approach." Expert Systems with Applications 57 (2016): 117-126.*
14. *Giatsoglou, Maria, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, and Konstantinos ChChatzisavvas. "Sentiment analysis leveraging emotions and word embeddings." Expert Systems with Applications 69 (2017): 214-224.*

15. *Saif, Hassan, Yulan He, Miriam Fernandez, and Harith Alani. "Contextual semantics for sentiment analysis of Twitter." Information Processing & Management 52, no. 1 (2016): 5-19.*