Performance Evaluation of Several Machine Learning Classification Algorithms with Combined Feature Selection Methods for Sentiment Analysis

Premnarayan Arya, Amit Bhagat

Abstract*: *Sentiment analysis (SA) is broadly studied to extract opinions from on line reviews and several methods have been proposed in current works. SA algorithms are used to classifying reviews in positive and negative. SA or machine learning classification algorithms apply directly on online review data sets without using feature selection methods (FSMs) leads poor performance. Towards deal with this problem, we proposed a model to improved performance of sentiment classification methods. This paper investigated performance of five machine learning classification algorithms like Naïve Bayes, k-Nearest Neighbor, Support Vector Machine, Logistic Regression, and Random Forest with different FSMs Unigram, Bigram, Information Gain, Chi-Square, Gini Index. Our method implemented on two data sets, first, electronics product data sets and second, movie review data sets. In starting, applying individually FSMs to extract features then applying combined FSMs to generate feature vector score. The feature selected by their feature vector score ranking. In last, the classification algorithms used popular feature vector for classifying the reviews into positive and negative. The performance measured of the classifier through Precision, recall, and F-measure. The best results achieved by all classification algorithms with the combination of FSMs (Unigram, Bigram, IG, GI, CS), and the highest F-score achieved by RF algorithm.

Index Terms*: *Sentiment analysis, Machine learning classification algorithm, Feature selection method.

I. INTRODUCTION

Sentiment analysis (SA) has sentiments in different forms like people opinions, emotions, and attitudes etc. This interesting topic increased commerce and culture [1]. SA can observed right decision basis on user feedback. Different social media use SA like Twitter, Facebook, Instagram, Google+, shopping, movie industries, politics, and healthcare. The movie industries decide star rating and percentage of a particular movie based on public review. The people write review about any movie by their personal knowledge and experience. People purchased any product online then they used that product. After that, they post a review about that particular product good or bad, we should purchase or not purchased after reading that post. So, these reviews increased product sell in market.

Researchers work continuously on machine learning to solve any problem quickly in minimum time. Machine learning, deep learning, and artificial Intelligence comes under computer science to trained machines and write programming to work smart [2]. For classification purpose used supervised machine learning to classify sentiments in positive and negative. By training data sets to trained a model then by testing data sets to test that model work according previous training. Unlabeled data used in unsupervised machine learning to classify document by own intellect. Labeled data and unlabeled data used by semi supervised machine learning to classify a document.Machine learning field have many challenges, like text mining, sentiment classification, image processing or classification problem quiet remain to solve in this area. SA mostly has three types such as document based sentiment classification, sentence based sentiment classification and aspect based sentiment classification. First, document based SA, to find out about whole document is positive of negative. For example, a review related to a mobile product contains twenty lines, and then we observed whole review not an individual sentence or words. Second, sentence based SA to find out sentiment one by one sentence in a whole review. And third, aspect based SA to find out a review about product features. Remaining part of the research paper contained following; related work, this section mentioned previous existing work of researchers. Proposed approach, this section explained sentiment detection. Methodology, this section explained feature selection methods. Classification, this section mentioned about classification which used in this paper. Experiments and results, section explores results by tables and graphs. And last, conclusion concludes about research paper.

II. RELATED WORK

Zhang et al. [4] used sentiment classification algorithms and proposed a method for word2vec and SVM^{perf}. The authors used Chinese language cloth product comment to make cluster for similar features word2vec and classified texts through SVM^{perf}. Authors also used feature selection method such as lexicon based and part of speech to generate training case.

Revised Manuscript Received on April 10, 2019.

Premnarayan Arya, Maulana Azad National Institute of Technology, Bhopal, India.
Dr. Amit Bhagat, Maulana Azad National Institute of Technology, Bhopal, India.

They got classification accuracy 87.10% of word2vec and 90.30% SVM^{perf}.

Rahman et al. [5] research on ensemble machine learning like Gradient Boost, Bagging Classifier, Random Forest, Extra Tree, and Ada Boost apply with unigram and bigram separately. The authors got highest accuracy through combined bigram+tf-idf compare to separate unigram and bigram. Performance of Extra Tree is 100% accuracy with combined.

Ghosh et al. [6] proposed research by using classification methods like MNB, SMO, RF, and LR with feature selection method like information gain, chi square and gini index separate-separate. After that they also apply together on imdb data sets and kitchen data sets. Precision, recall, F-measure and ROC used for performance calculation. SMO classifier obtained highest accuracy 92.31.

Zhang et al. [7] proposed framework for sentiment classification. Authors experiment by SentiWordNet for sentiment lexicon and product review data like dvd, kitchen, books, and electronics. F-measure and accuracy achieved by sentiment lexicon method is good than BAT and SOCAL method on data sets.

Singh et al. [8] proposed a methodology and mainly focus on OneR machine learning algorithm as well as BF tree, NB, and J48 algorithms used on review data of imdb and Amazon for training and testing. Authors used feature selection method like Document frequency, Mutual Information, and Information Gain. The experimental results obtained highest percentage by OneR got F-score 97.0%, BF tree got F-score 72.1%, NB got 81.2%, and J48 got 91.7%.

Elmurngi et al. [9] study on four supervised machine learning algorithms to classify the sentiment that is compared using three different Amazon reviews datasets like Clothing, Shoes and Jewelry reviews, Baby reviews and Pet supplies reviews. These methods also detect unfair positive and unfair negative reviews. This research is main goal to classify Amazon reviews datasets into fair reviews or unfair reviews with the use of SA algorithms and supervised learning techniques. The experimental results show their accuracy and performance of four sentiment classification algorithms in order to detect unfair reviews. Experiment result found that the LR algorithm is best classifier with highest accurate as compared to the NB, SVM and Decision Tree (DT-J48) algorithms, not merely text classification but in unfair reviews detection as well.

Wang et al. [10] worked on ensemble methods like Boosting, Random subspace, and Bagging. These methods combined with base classification methods like NB, ME, DT, KNN, and SVM for classifying sentiments. After investigation, authors found are that ensemble methods performance best than base classification methods.

Ghosh et al. [11] proposed model for SA. Authors used support vector machine, multinomial naïve bayes, k-nearest neighbor, and maximum entropy with feature selection methods like information gain, gini index, and chi square applied on reviews data sets. They did experiment on classifier and feature selection method by different types. Overall Support vector machine produce good accuracy with information gain.

Catal et al. [12] study on classification algorithms like support vector machine, naïve bayes, and bagging. Authors applied classifier individually as well as their fusion on turkish movie reviews. And, experimental results achieved highest accuracy by fusion classifier than individually classifier.

Solai et al. [13] worked on RMSE and MAE for real life data sets like Amazon and epinions. This methods works on according to buyer requirements such as flavor, faith and tong of mouth (review) spread by the social media facebook, twitter etc.

III. PROPOSED APPROACH

There are four steps involved in our proposed approach. These steps are describing following as:

Step1: Collection of data: we collect data like electronics product review data sets and movie review data sets for sentiment classification.
Step2: Pre-processing of data: preprocessing means to make raw data into right format. It removed redundancy from data.
Step3: Selection features: Feature selections also known as attribute selection or variable selection. In feature selection method identify words frequency for classifying sentiment. In SA have noninheritable a big role in increasing classification accuracy and distinguishing relevant attributes. [14].
Step4: Classification algorithms: it is used to classify documents or reviews in positive and negative sentiment. We used five machine learning classification algorithms for classification the data set.

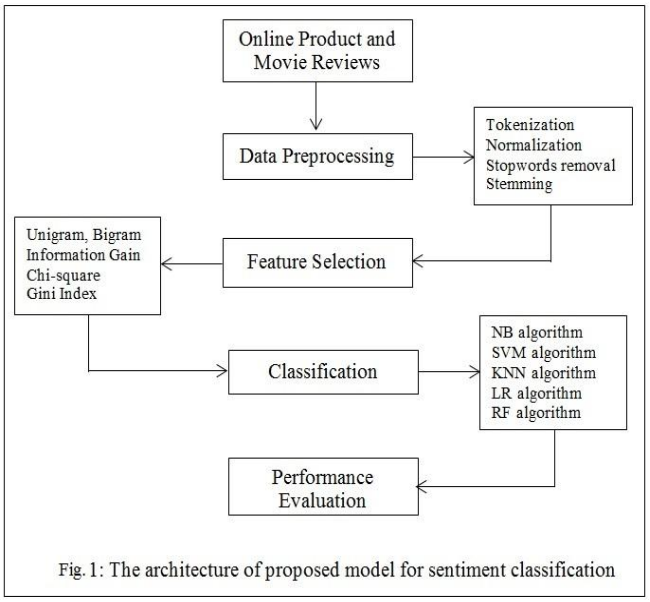

Fig. 1: The architecture of proposed model for sentiment classification

IV. METHODOLOGY

Latest field introduced of sentiment level classification [15], on the other hand, text classification introduced later [16]. This study discovered five machine learning classification algorithms performance by individually and joint feature sets.

A. Planning for data set

The experiments work applied on reviews of movie data [17]. This movie data organized by Pang and Lee in the year of 2014 [18]. Our research work implemented used by electronics product and movie review data sets for classifying sentiments or opinion. These review data sets are standard type and used by many authors and researchers in their research work. The movie review cover 2000 reviews where 1000 positive and 1000 negative. The data set (electronics product) is available on Amazon website [19]. This data set also used by authors Blitzer et al. [20] for their research work.

B. Pre-processing

Tokenization splits document in a small tokens like word, special character, number etc. after that the document used for processing. The words after convert tokenization make small letter or capital letter, this process called normalization. Then remove stop words such as new line, white space, prepositions, and special character. Identify the root of word is known as stemming.

C. Ngrams

Ngram model consists of a contiguous sequence of n words of a given review data set. Mostly models applied with 1-gram sequence, 2-gram sequence, and 3-gram sequence, and sometimes the sequence can be extended. For text sample data, "something is better than nothing". Unigrams (1 gram or n=1), "something", "is", "better", "than", "nothing". Bigrams, 2 gram or n=2, "something is", "is better", "better than", "than nothing". Trigrams, 3 gram or n=3, "something is better", "is better than", "better than nothing".

D. Feature selection methods

FSMs select specific features or attributes from user or customer review data sets. Selection of significant feature is very essential task for accurate and correct classification of user review sentiments. These selected features play important role for training and testing data sets [21]. This process can improve performance of our proposed model and generate good accuracy during the experiment. There are three FSMs used in our research work, which describes as follows:

a. *Information Gain:* The IG is employed to pick the rending attribute in every node within the tree. The attribute with the very best info gain is chosen because the rending attribute for this node. It's wont to choose necessary options supported the category attribute rules of features classification. The IG value of every term will live the amount of bits of data non inheritable for sophistication prediction by knowing the presence or absence of that term within the document.

$$IG(f) = \{ \sum_{x=1}^{m} P(c_x) \log P(c_x) \} + \{ P(f) \sum_{x=1}^{m} P(c_x \mid f) \log P(c_x \mid f) \} + \{ P(f') \sum_{x=1}^{m} P(c_x \mid f) \log P(c_x \mid f) \} \quad (1)$$

b. *Chi-square (χ^2):* Chi-square may be a terribly unremarkably applied mathematics takes a look at, which might quantify the association between the feature or term f and its connected category C_x. It tests a null-hypothesis that the 2 variables feature and sophistication are fully freelance of every alternative. The CHI worth of feature f for sophistication C_x is higher and therefore the nearer relationship exists between the variables feature f and sophistication C_x. The options with the very best χ^2 values for a class ought to perform for classifying the texts.

$$\chi_c^2 = \Sigma \frac{(O_i - E_i)}{E_i} \quad (2)$$

Where, subscript "c" is that the degrees of freedom. "O" is discovered value and E is mean. There are some variations on the chi-square data point. That one utilize depends upon however you collected the information and which hypothesis is being tested. However, all of the variations use the identical plan that is you just are scrutiny expected values with the values really collect.

c. *Gini Index:* GI measures the options to discriminate between categories. This methodology was in the main projected to be used for call tree formula supported associate degree impurity split method. the most principle of GI is to contemplate D as a dataset of the sample having m variety of various categories $C_{i=1}^{m}$ = {c_1, c_2...c_n}. Per the category level, the sample set will be splitted into n set (Si, x=1, 2…n).

Gini Index measures the features ability to discriminate between classes. This method was mainly proposed to be used for decision tree algorithm based on an impurity split method. The main principle of Gini Index is to consider D as a dataset of the sample having m number of different classes $C_{i=1}^{m}$ = {c_1, c_2...c_n}. According to the class level, the sample set can be splitted into n subset (Si, i=1, 2…n).

$$GI(D) = 1 - \sum_{x=1}^{n} P_x^2 \quad (3)$$

Where, probability P_x of any sample belongs to category C_x, may be computed by S_i/S [22]. GI for a feature may be calculable severally for binary classification. The researchers adopted Gini index Text (GIT) technique for scheming the feature score that was bestowed in [23].

V. MACHINE LEARNING CLASSIFICATION ALGORITHMS

A. *Naïve Bayes*

A Naïve mathematician categoryifier could be an easy probabilistic classifier supported Bayes theorem with robust independence assumption that the presence of a feature in a very class doesn't rely on the presence or absence of another feature. In text classification, the given document is assigned a category

$$C = \text{argum max}_c\ p(c|d) \quad (4)$$

Its underlying likelihood model is represented as an "independent feature model". The NB classifier uses the Bayes' rule as follows:

$$p(c|d) = \frac{p(c)p(d|c)}{p(d)} \quad (5)$$

Where, $p(d)$ plays no role in choosing C. To estimate the term $p(d|c)$, Naïve mathematician decomposes it by assumptive the f_i is not absolutely freelance given d's category as

$$p_{NB}(c|d) = \frac{p(c)\ (\prod_{i=1}^{m} p(f_i|c)n_i(d)}{p(d)} \quad (6)$$

Where, m is that the no of options and f_i is the feature vector. Contemplate a preparation technique consisting of relative-frequency estimation p(c) and $p(f_i|c)$, [24].

B. *Support Vector Machine*

SVMs are a supervised learning principle for classification and multivariate analysis for binary classification in each linear and nonlinear [25]. SVMs are supported the concept of finding the most effective doable hyper plane that best divides a dataset into two classes.

Objective function= Regularization function + Loss function

$$\text{Min}_w\ \lambda\ ||w||^2 + \sum_{i=0}^{n}(1 - y_i(x_i i, w)) \quad (7)$$

Where x_i the input sample, y_i is that the output label and w is represent weight vector and λ is the regularization parameter.

C. *K-Nearest Neighbor*

The KNN algorithmic rule works by examining the k nearest instances within the preparation data set and creating a function for feature f_i ,and class c feature $f_{i,c}$ (d, c) are often outlined as follows;

$$f_{x,c}(d, c) = \begin{cases} 1 & Nx(d) > 0,\ c' = 0 \\ 0 & otherwise \end{cases} \quad (8)$$

Where, function $f_{i,c}$ (d, c) is to feature f_x, and class cN_x(d) indicates the occurrence of feature x in document d. The feature-class pair which occurs very frequently in document d, having high frequency, is the strong indicator for class c. The function which holds a strong orientation will be set to 1; otherwise it will be 0.

D. *Logistic Regression*

LR could be a classification algorithmic program, (also known as logit function) for associated analysing a dataset during which there are one or additional freelance variables that confirm an outcome [26]. It works on binary variable (two classes) and multi variable (multiclass). It finds the simplest fitting model to explain the connection between variable quantity and a collection of freelance variables. The logit function represent as follows;

$$\text{logit}(p) = a_0 + a_1W_1 + a_2W_2 + \ldots\ldots\ldots + a_nW_n \quad (9)$$

Where, possibility 'p' occurrence of a features. The logit alteration is defined as the logged odds. The logged odds represent as follows;

$$\text{odds} = \frac{p}{1-p} = \frac{possibility\ of\ presence\ of\ features}{possibility\ of\ absence\ of\ features} \quad (10)$$

$$\text{logit}(p) = ln\left(\frac{p}{1-p}\right) \quad (11)$$

E. *Random Forest*

RF is associate ensemble learning methodology used for classification and regression issues. RF uses a fabric approach to form a bunch of call trees with random set of the information. A model is trained many times on random sample of the dataset to attain smart prediction performance from the random forest algorithmic rule. During this ensemble learning methodology, the output of all the choice trees within the random forest, is combined to create the ultimate prediction. The ultimate prediction of the random forest algorithmic rule springs by polling the results of every call tree or simply by going with a prediction that seems the foremost times within the decision trees. To make RF model with coaching dataset D as follows;

$$\text{Training dataset (D)} = \{(f_x, C_x)\ N_{x=1} \mid f_x \in R^F, C \in \{1, 2, \ldots, c\}\} \quad (12)$$

Where f_x are options, C_x is that the set of categories and N represents the amount of preparation samples. Sample the coaching set D with replacement to make bagged samples D1, D2 ,…,D_p and every call tree is fully grown from these bagged sample set. In every call tree, for each node we have a tendency to take into account a random and separate set of prognosticative options as candidate feature for cacophonous the node. The category prediction of RF model with p variety of trees will outline as follows. Let assume $\hat{C}^p$ be the prediction of tree T_p given input *f*.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

C = common elective $\{\hat{C}^p\}_1^p$ (13)

VI. EXPERIMENTS AND RESULTS

A. Experimental setting

We used Anaconda Python 3.5.2 Windows-x86_64 version to conduct all the experiments. The CSV file load in python. First step preprocessing of raw text is done using NLTK. For feature choice and organization purpose, we tend to used scikit-learn, numPy, SciPy python library to reinforce and extend the basic python abilities.

B. Evaluation Parameters

ML algorithmic rule is calculated the term or components of confusion matrix on a collection of take a look at information. The confusion matrix consists of 4 terms, like, True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). In keeping with the worth of those components, the analysis matrices precision, recall, and F-score are determined to assessment the performance score of every classifier.

Precision: $\frac{TP}{TP+FP}$ (14)

Recall: $\frac{TP}{TP+FN}$ (15)

F-score: $\frac{2 * Precision * Recall}{TPrecision + Recall}$ (16)

C. Results and discussion

The experimental result shows from table 1-4. Table 1 shows five classifiers NB, KNN, SVM, LR, and RF with individually feature selection method using electronics product reviews data sets then produced Precision, Recall and F-score values. NB classifier with Chi method generate f-score 84.93. KNN classifier with GI method generates f-score 85.29. SVM classifier with IG method generates f-score 86.92. LR classifier with GI method generates f-score 89.04, and RF classifier with Chi method generates highest f-score 89.97.

Table 2 shows five classifiers NB, KNN, SVM, LR, and RF with combined feature selection method using electronics product reviews data sets then produced Precision, Recall and F-score values. NB classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 86.28. KNN classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 87.80. SVM classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 92.10. LR classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 92.84, and RF classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates highest f-score 94.49.

Table 3 shows five classifiers NB, KNN, SVM, LR, and RF with individually feature selection method using movie reviews data sets then produced Precision, Recall and F-score values. NB classifier with GI method generate f-score 85.69. KNN classifier with IG method generates f-score 85.79. SVM classifier with GI method generates f-score 88.54. LR classifier with Chi method generates f-score 90.15, and RF classifier with GI method generates highest f-score 93.20.

Table 4 shows five classifiers NB, KNN, SVM, LR, and RF with combined feature selection method using electronics product reviews data sets then produced Precision, Recall and F-score values. NB classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 88.94. KNN classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 89.92. SVM classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 92.29. LR classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates f-score 93.88, and RF classifier with combined FSMs Unigram+Bigram+IG+CS+GI generates highest f-score 94.98.

The summary of experimental result is that the highest f-score generates by combined FSMs with classifier than individually FSMs with classifier. The f-score value increased lowest to highest from table1-4. While comparing the performance of classifier algorithm RF produces the best result with movie review data sets than electronics product review data sets. The table 4 is represented highest f-score obtained by RF 94.98 with combined method Unigram+Bigram+IG+CS+GI.

	Classifier														
Method	NB			KNN			SVM			LR			RF		
	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Unigram	82.4	80.7	81.54	83.3	79.7	81.46	85.4	83.1	84.23	87.1	85.8	86.45	88.1	86.3	87.19
Bigram	82.0	78.9	80.42	81.2	79.3	80.24	85.9	82.7	84.27	87.0	84.4	85.68	88.0	86.6	87.29
Info Gain	83.1	82.6	82.85	85.8	83.6	84.69	88.5	85.4	**86.92**	88.7	86.2	87.43	89.5	87.8	88.64
Chi-square	86.3	83.6	**84.93**	84.7	82.8	83.74	87.3	86.2	86.75	89.3	88.6	88.95	91.6	88.4	**89.97**
Gini Index	85.0	82.7	83.83	86.2	84.4	**85.29**	87.8	84.6	86.17	89.8	88.3	**89.04**	90.0	87.7	88.84

Table 1: Experimental result of different classifier with individually feature selection methods for electronics product review data set.

	Classifier														
Method	NB			KNN			SVM			LR			RF		
	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Uni+Bigram	76.3	75.8	76.05	77.5	75.4	76.44	85.1	83.8	84.44	86.5	84.7	85.59	88.2	86.3	87.24
Uni+Bi+Info_Gain	86.1	83.4	84.73	87.8	84.3	86.01	91.6	87.3	89.4	88.7	86.3	87.48	89.2	87.8	88.49
Uni+Bi+Chi_Square	87.2	84.1	85.62	86.5	83.7	85.08	88.9	86.5	87.68	90.8	88.2	89.48	92.6	90.4	91.49
Uni+Bi+Gini_Index	85.8	83.6	84.69	86.9	84.1	85.48	90.1	88.7	89.39	92.3	90.1	91.19	93.5	90.8	92.13
Uni+Bi+IG+CS+GI	87.6	85.0	**86.28**	88.3	87.3	**87.80**	92.4	91.8	**92.10**	93.7	92.0	**92.84**	95.4	93.6	**94.49**

Table 2: Experimental result of different classifier with combined feature selection methods for electronics product review data set.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Method	Classifier														
	NB			KNN			SVM			LR			RF		
	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Unigram	83.1	82.5	82.80	84.1	82.8	83.44	85.8	83.4	84.58	88.3	86.6	87.44	89.5	87.3	88.39
Bigram	82.7	80.2	81.43	82.9	80.3	81.58	84.7	82.1	83.38	87.8	85.4	86.58	88.3	88.3	88.30
Info Gain	84.3	82.8	83.54	86.9	84.7	**85.79**	88.3	85.6	86.93	90.6	87.3	88.92	91.3	91.3	91.30
Chi-square	85.2	83.4	84.29	86.2	83.6	84.88	88.8	87.1	87.94	92.3	88.1	**90.15**	92.1	92.1	92.10
Gini Index	86.6	84.8	**85.69**	86.4	84.5	85.44	89.3	87.8	**88.54**	91.8	88.3	90.02	93.2	93.2	**93.20**

Table 3: Experimental result of different classifier with individually feature selection methods for movie review data set.

Method	Classifier														
	NB			KNN			SVM			LR			RF		
	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score	Prec	Recall	F-Score
Uni+Bigram	77.8	75.3	76.53	78.2	79.6	78.89	85.6	83.0	84.28	87.1	84.9	85.99	88.6	87.1	87.84
Uni+Bi+Info_Gain	88.3	85.8	87.03	88.6	85.7	87.13	92.4	90.0	91.18	90.2	87.5	88.83	92.3	90.0	91.14
Uni+Bi+Chi_Square	87.5	85.1	86.28	89.8	86.9	88.33	94.8	92.3	93.53	93.6	91.0	92.28	94.7	91.3	92.97
Uni+Bi+Gini_Index	88.2	86.2	87.19	89.2	87.1	88.14	92.4	90.8	91.59	94.7	91.6	93.12	95.3	92.8	94.03
Uni+Bi+IG+CS+GI	90.0	87.9	**88.94**	91.6	88.3	**89.92**	93.2	91.4	**92.29**	95.1	92.7	**93.88**	96.4	93.6	**94.98**

Table 4: Experimental result of different classifier with combined feature selection methods for movie review data set.

The classifier performance is shown in figure 2-5. The figures 2 and 3 are shows precision, recall and f-score separate-separate of NB, KNN, SVM, LR, and RF classifier using electronics product reviews data sets. In figure 2, the NB obtained maximum precision value 86.3, maximum recall value 83.6, and F-score maximum value 84.93 with Chi-square method. The KNN obtained maximum precision value 86.2, maximum recall value 84.4, and F-score maximum value 85.29 with Gini Index method. The SVM obtained maximum precision value 88.5, maximum recall value 85.4, and F-score maximum value 86.92 with Information Gain method. The LR obtained maximum precision value 89.8, maximum recall value 88.6 (with Chi-square method), and F-score maximum value 89.04 with Gini Index method. The RF obtained maximum precision value 91.6, maximum recall value 88.4, and F-score maximum value 89.97 with Chi-square method.

In figure 3, the NB obtained maximum precision value 87.6, maximum recall value 85.0, and F-score maximum value 86.28 with combined Unigram+Bigram+IG+CS+GI method. The KNN obtained maximum precision value 88.3, maximum recall value 87.3, and F-score maximum value 87.80 with combined method. The SVM obtained maximum precision value 92.4, maximum recall value 91.8, and F-score maximum value 92.10 with combined method. The LR obtained maximum precision value 93.7, maximum recall value 92.0, and F-score maximum value 92.84 with combined method. The RF obtained maximum precision value 95.4, maximum recall value 93.6, and F-score maximum value 94.49 with combined method.

The figures 4 and 5 are shows precision, recall and f-score separate-separate of NB, KNN, SVM, LR, and RF classifier using movie reviews data sets. In figure 4, the NB obtained maximum precision value 86.6, maximum recall value 84.8, and F-score maximum value 85.69 with GI method. The KNN obtained maximum precision value 86.9, maximum recall value 84.7, and F-score maximum value 85.79 with IG method. The SVM obtained maximum precision value 89.3, maximum recall value 87.8, and F-score maximum value 88.54 with GI method. The LR obtained maximum precision value 92.3, maximum recall value 88.3 (with GI method), and F-score maximum value 90.15 Chi method. The RF obtained maximum precision value 93.2, maximum recall value 90.8, and F-score maximum value 93.20 with GI method.

In figure 5, the NB obtained maximum precision value 90.0, maximum recall value 87.9, and F-score maximum value 88.94 with combined Unigram+Bigram+IG+CS+GI method. The KNN obtained maximum precision value 91.6, maximum recall value 88.3, and F-score maximum value 89.92 with combined method. The SVM obtained maximum precision value 93.2, maximum recall value 91.4, and F-score maximum value 92.29 with combined method. The LR obtained maximum precision value 95.1, maximum recall value 92.7, and F-score maximum value 93.88 with combined method. The RF obtained maximum precision value 96.4, maximum recall value 93.6, and F-score maximum value 94.98 with combined method.

Finally, the performance of Random Forest classification method is best than other classification algorithms NV, KNN, SVM, and LR for electronics products data sets as well as movie review data sets.

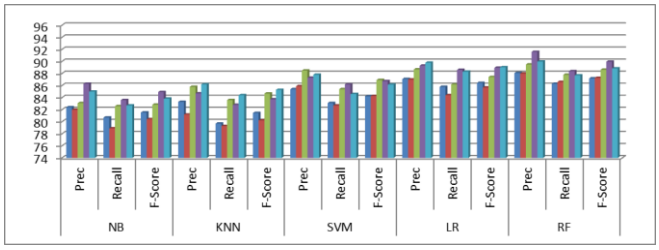

Fig. 2: Performance comparison of five classifier for electronics product review data set.

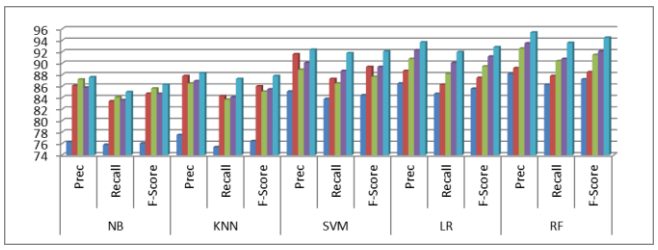

Fig. 3: Performance comparison of five classifier for electronics product review data set.

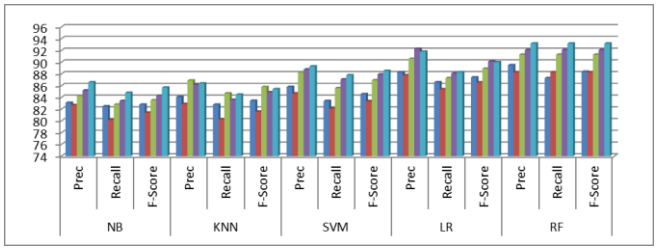

Fig. 4: Performance comparison of five classifier for movie review data set.

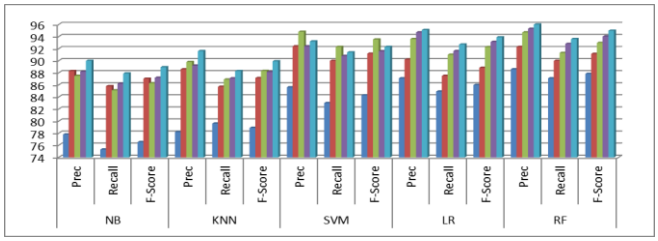

Fig. 5: Performance comparison of five classifier for movie review data set.

D. Performance Evaluation

The experiments include five machine learning classification algorithms NV, KNN, SVM, LR, and RF used with individually and combined FSMs applied on electronics product review data sets and movie review data sets. This study evaluated that well trained machine learning algorithm to perform correctly classification on reviews SA. The RF is best classification algorithm for all tests, as it accurately classified 88.84, 94.49 on electronics products data sets and 93.20, 94.98 on movie review data sets.

This study [27], applied ML with FSMs on Malay language sentiment. It had been incontestable that enriched feature choices resulted in higher performance in Malay sentimental. They approached 3 FSMs (IG, Gini Index, and Chi) to reinforce the performance of 3 ML classifiers (SVM, NB, and KNN). The movie reviews data set utilized in Malay language. The outcomes displayed of SVM classifier and IG-base technique established the most effective classifier with 85.33% accuracy and feature size is 300. They too find the usage of FSMs upgraded outcomes of original classifier. In this study [28], on SA, author have used SVM, NB, and ME classifier with ngram technique like unigram and bigram as well as their combination on movie review data sets. They achieved the accuracy of 82.7, 81.2, and 81.0 for the classifiers SVM, NB, and ME, respectively. In this study [29], observed how classifiers SVM, NB, and KNN work with different feature sizes of movie review data set. Topmost feature selected by IG FSMs from training data sets. This research investigated, the SVM performed best than Naive Bayes and KNN approaches with highest accuracy of 81.71. In this research [30], explore ML classifiers, like, NB, SVM, ME, and SGD, to analyzed sentiments by ngram of movie reviews data set.

VII. CONCLUSION

Main goal of this research is to discover performance of ML classification. There are five classification algorithms used with different FSMs on electronics product review data sets and movie review data sets. First, we applied the classification algorithms with individually unigram, bigram, IG, CS, GI on electronics product review data sets. Second, we applied the classification algorithms with combined Unigram+Bigram, Unigram+Bigram+IG, Unigram+Bigram+CS, Unigram+Bigram+GI, Unigram+Bigram +IG+CS+GI on electronics product review data sets. Third, we applied the classification algorithms with individually unigram, bigram, IG, CS, GI on moview review data sets. And fourth, we applied the classification algorithms with combined Unigram+Bigram, Unigram+Bigram+IG, Unigram+Bigram+CS, Unigram+Bigram+GI, Unigram+Bigram +IG+CS+GI on movie review data sets. The performance evaluation measured by precision, recall, and F-measure. The main purpose of FSMs used to improve the performance of classification algorithms. Highest score produce used with combined FSMs and classifier than individually FSMs. Hence, the classification algorithms give highest F-score with combined FSMs (Unigram+Bigram +IG+CS+GI) like NB F-score is 86.28, KNN F-score is 87.80, SVM F-score is 92.10, LR F-score is 92.84 and RF F-score is 94.49 for electronics product review data sets, NB F-score is 88.94, KNN F-score is 89.92, SVM F-score is 92.29, LR F-score is 93.88, and RF F-score is 94.98 for movie review data sets.

For future work, we wish to extend this work to use more machine learning classification with different feature selection methods on different SA datasets.

REFERENCES

1. Bing Liu, "Sentiment Analysis, Mining opinions, Sentiments and Emotions," Book, June 2015.
2. Yagang Zhang, Taiwo Oladipupo Ayodele, Types of machine learning algorithms, new advanced in machine learning, InTech, 2010.
3. Liu B. Sentiment analysis and subjectivity. In: Indurkhya N, Damerau FJ, editors. Invited chapter for the handbook of natural language processing. 2nd ed. England: Taylor & Francis; 2010.
4. Dongwen Zhang, Hua Xu, Zengcai Su, and Yunfeng Xu, Chinese comments sentiment classification based on word2vec and SVMperf, Expert Systems with Applications, vol. 42(4), March 2015, pp. 1857-1863.
5. Sheikh Shah Mohammad Motiur Rahman, Md. Habibur Rahman, Kaushik Sarker, Md. Samadur Rahman, Nazmul Ahsan, and M. Mesbahuddin Sarker, Supervised Ensemble Machine Learning Aided Performance Evaluation of Sentiment Classification, IOP Conf. Series: Journal of Physics: Conf. Series 1060 (2018), 012036 doi:10.1088/1742-6596/1060/1/012036.
6. Monalisa Ghosh and Goutam Sanyal, An ensemble approach to stabilize the features for multi-domain sentiment analysis using supervised machine learning, Journal of Big Data, Springer Open, Nov 2018, https://doi.org/10.1186/s40537-018-0152-5.
7. Hongyu Han, Yongshi Zhang, Jianpei Zhang, Jing Yang, Xiaomei Zou, Improving the performance of lexicon-based review sentiment analysis method by reducing additional introduced sentiment bias, PLoS ONE 13(8): e0202523. https://doi.org/10.1371/journal.pone.0202523
8. Jaspreet Singh, Gurvinder Singh and Rajinder Singh, Optimization of sentiment analysis using machine learning classifiers, Human-centric Computing and Information Sciences, Springer Open, Dec 2017, DOI 10.1186/s13673-017-0116-3.
9. Elshrif Ibrahim Elmurngi and Abdelouahed Gherbi, Unfair reviews detection on Amazon reviews using Sentiment Analysis with supervised learning techniques, in Journal of Computer Science, 2018, pp. 714-726.
10. Gang Wang, Jianshan Sun, Jian Ma, Kaiquan Xu, Jibao Gu. Sentiment classification: The contribution of ensemble learning, Decision Support Systems, Vol 57(1), Jan 2014, pp. 77-93,

https://doi.org/10.1016/j.dss.2013.08.002
11. Monalisa Ghosh and Goutam Sanyal, Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis, Hindawi Applied Computational Intelligence and Soft Computing, Volume 2018, https://doi.org/10.1155/2018/8909357
12. Cagatay Catal, Mehmet Nangir, A Sentiment Classification Model Based On Multiple Classifiers, Applied Soft Computing Journal, Vol. 50, Jan 2017, pp. 135-141.
13. S.V. Solai Ananth1, Chandu PMSS, Live Twitter Knowledge as a Corpus for Sentiment Analysis and Opinion Mining , International Journal of Engineering Science and Computing, January 2017.
14. Koncz, P. and J. Paralic, 2011. An approach to feature selection for sentiment analysis. proceedings of the 15th International Conference on Intelligent Engineering Systems, Jun. 23-25, IEEE Xplore Press, Poprad, Slovakia, pp: 357-362. DOI: 10.1109/INES.2011.5954773
15. Kolog EA, Montero CS, Toivonen T. Using machine learning for sentiment and social influence analysis in text In: Advances in intelligent systems and computing. Cham: Springer; 2018.
16. Hatzivassiloglou V, McKeown KR. Predicting the semantic orientation of adjectives. In Proceedings of the 8th conference on European chapter of the association for computational linguistics. 1997, pp. 174–81.
17. https://www.kaggle.com/nltkdata/movie-review#movie_reviews.zip.
18. Pang B, Lee L. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: Proceedings of the 42nd annual meeting of the ACL. Barcelona; 2004.
19. http://www.cs.jhu.edu/~mdredze/datasets/sentiment. Accessed 7 Jan 2017.
20. Blitzer J, Dredze M, Pereira F. Biographies, bollywood, boom-boxes and blenders: domain adaptation for sentiment classification. In: Proc. assoc. computational linguistics. Austin: ACL Press; 2007. p. 440–7.
21. Li JJ, Yang H, Tang H. Feature mining and sentiment orientation analysis on product review. In: Management information and optoelectronic engineering. 2015.
22. Shang W, Huang H, Zhu H, Lin Y, Qu Y, Wang Z. A novel feature selection algorithm for text categorization Expert Syst Appl. 2007, 33(1):1–5.
23. Park H, Kwon S, Kwon MF. Complete Gini-index text (git) feature-selection algorithm for text classification. In: Proceedings of software engineering and data mining (SEDM). Piscataway: IEEE; 2010, pp. 366–1.
24. Pang, B., Lee, L., Vaithyanathan, S. (2002). 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.
25. C.W. Hsu, C.C. Chang, and C.J. Lin, A practical guide to support vector classification, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada, 2005.
26. Kantardzic, M. "Data Mining: Concepts, Models, Methods and algorithms," New York: Wiley-IEEE Press, 2011.
27. T. Al-Moslmi, S. Gaber, A. Al-Shabi, M. Albared, and N. Omar, Feature selection methods effects on machine learning approaches in malay sentiment analysis, in Proceedings of the 1st ICRIL International Conference on Innovation in Science and Technology (lICIST '15), 2015, pp. 444–447.
28. B. Pangand L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL'04), Association for Computational Linguistics, Barcelona, Spain, July 2004.
29. P. Kalaivani and K. L. Shunmuganathan, Sentiment classification of movie reviews by supervised machine learning approaches, Indian Journal of Computer Science and Engineering (IJCSE), vol. 4, 2013, pp. 286–292.
30. Tripathy, A. Agrawal, and S. K. Rath, Classification of sentiment reviews using ngram machine learning approach, Expert Systems with Applications, vol. 57, 2016, pp. 117-126.

AUTHORS PROFILE

Premnarayan Arya, currently pursuing Ph.D degree in the Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India, He received his M.Tech in Computer Science degree from Devi Ahilya University, Indore in 2011. He received his Master of Computer Application degree from Rajeev Gandhi Technical University, Bhopal, MP, India, in 2008, and He receive his Graduation degree from Makhanlal University, Bhopal, MP, India, in 2005. He has more than eight year of teaching, research experience. His research interest includes sentiment analysis, opinion mining, and data mining, machine learning, natural language processing.

Dr. Amit Bhagat, received Ph.D degree in Data Mining from the Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal, India, in 2013. He received his Master's degree and bachelor's degree in Computer Applications from MCU, Bhopal, MP, India, in 2003 and 2001 respectively. He has more than sixteen year of teaching, research and industry experience. He has various publications in national and international journals and conferences. His current interest includes Big data analytics, data mining and Web mining.