Indonesian Journal of Electrical Engineering and Computer Science
Vol. 22, No. 2, May 2021, pp. 1041~1051
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v22.i2.pp1041-1051

Analyzing impact of number of features on efficiency of hybrid model of lexicon and stack based ensemble classifier for twitter sentiment analysis using WEKA tool

Sangeeta Rani, Nasib Singh Gill, Preeti Gulia
Department of Computer Science & Applications, M D University, Haryana, India

Article Info

Article history:

Received Feb 8, 2021
Revised Mar 24, 2021
Accepted Apr 11, 2021

Keywords:

Ensemble classifier
IBK
Lexicon based classifier
Meta stacking
REPTree
SMO
Voting ensemble

ABSTRACT

Twitter is used by millions of people across the world, so the data collected from Twitter can be highly valuable for research and helpful in decision support. Here in this paper 'Twitter US Airline data' from Kaggle data repository is used for sentiment classification of customers' reviews. The current research aims to implement various machine learning classifiers, Stack-based ensemble classifiers and hybrid of lexicon classifier with other classifiers. 11 different classification models are implemented for different sized feature sets. Also, all the 11 models are re-implemented by adding sentiment score of lexicon based classifier as one of the features in the feature set. Results are analyzed by varying number of input feature variables used in the classification. Four different size feature sets having 301,501, 701, and 1301 number of features are used to analyze the variations in the final findings. Chi-Square and Information gain techniques are used for feature selection. The results show that an increase in the number of features increases the accuracy up to 701 features. After that, accuracy is stable or decreases with increase in feature set size. Also, the cost of adding sentiment score of lexicon classifier to the input feature set is nominal, but the results are improved consistently. WEKA and R Studio tools are used for analysis and implementation. Accuracy and Kappa are used for representing and comparing the efficiency of models.

Corresponding Author:

Sangeeta Rani
Department Computer Science and Application
Maharishi Dayanad University
Rohtak, Haryana, India
Email: Sangeeta.yogi@gmail.com

1. INTRODUCTION

Social networking sites provide a huge repository of data for various researches and analysis. Customers give their reviews about various products and services used by them on different social media platforms over the internet. The analysis of this data can be very useful in various decision support and policy making. Sentiment classification is an area of machine learning and artificial intelligence used to find the polarity of content at different sentiment levels like negative, positive or neutral. Twitter sentiment classification [1]-[2] deals with finding the sentiment polarity of Twitter data. Billions of tweets are posted daily by Twitter users related to different topics worldwide. This information can be analyzed and helpful in decision making related to education, business, medical science, sports, natural disasters, social problems and politics. Twitter data is attractive due to limited size of tweet, more probability of single sentiment in one tweet and vast number of users worldwide.

Sentiment classification of raw data leads to poor efficiency. Tweets preprocessing and cleaning helps to remove unnecessary noise and prepare the data for further processing [3]-[4]. In Twitter data, URL's, Hashtags, emoticons, numbers, and punctuation symbols are generally removed to clean data. As the data size involved in analysis is big, a large number of features are involved as input variables for classification model development. These large number of features leads to developing a complicated model without even much enhancing the efficiency of the model. The cost of using a large number of features is not proportionate to the enhancement of model accuracy and even sometimes the accuracy of the model may also reduce. Feature ranking and selection is used to select the minimum set of features without much affecting the efficiency of the model. Filter based, wrapper based and ensemble based are various feature selection techniques that are used to select the most appropriate features to develop an efficient model [5]-[8].

Basically, sentiment analysis is a classification technique that categorizes the tweets in various sentiment classes. Different classification algorithms as Machine learning [9]-[10], Lexicon based [11], Ensemble classifiers [12] and Neural network based classifiers are used by researchers. Ensemble classifiers are very effective to enhance the overall efficiency of a model. The hybrid of lexicon and ensemble based classifiers is also effective to improve accuracy further with negligible overhead.

In the present paper, various machine learning and hybrid classification algorithms are used for different sized feature set. 'Twitter US Airline data' is used for implementation and analysis. The impact of number of features on overall efficiency of the model is analyzed for various classification models. Two feature selection techniques and four different size feature sets are implemented by using 11 different machine learning and hybrid classification models. Stack-based ensemble classification models are also implemented by using SMO, NB, REPTree and IBK machine learners. Chi-Square and IG are used for feature ranking and selection. Both techniques are compared to analyze any effect on the efficiency of the model.

Section 1 of the paper gives the introduction to Twitter sentiment analysis. A literature survey of related work in the area of sentiment analysis is mentioned in Section 2. Feature selection is summarized in Section 3. Summary of the classification algorithm is given in Section 4. Section 5 gives the details of the data set used in the research. Simulation tools used are mentioned in Section 6. The methodology used in the research is mentioned in Section 7. Section 8 discusses the result and analysis part. The research work is concluded in Section 9.

2. RELATED WORK

Designing an efficient classification model considers several factors like preprocessing of data, feature selection technique, number of features involved, classification model and data itself. Selecting relevant features and the right number of features is very important for the development of an efficient model. The effect of using various classification models and different sized feature set are mentioned in various researches.

E. M. Karabulut *et al.* [13] shows the effect of various feature selection techniques: information gain, gain ratio, symmetric uncertainty, One-R, Relief-F, chi-square by using naïve bayes (NB), artificial neural network (ANN) as multilayer perceptron (MLP) and J48 decision tree as classification algorithms. WEKA is used for implementation of various models. As per research, chi-square and information gain performed better as compared to other feature selection methods. An improvement of maximum 15% in accuracy is observed and MLP is the most sensitive classification algorithm to feature selection. It is also observed for NB classifier the gain ratio, for MLP the chi-square and for J48 the Information Gain is the most positively effective feature selection algorithms.

V. Sugumaran *et al.* [14] performed their research to observe the effect of number of input feature variables on model accuracy by using SVM and PSVM models. The results show that increasing the number of features enhances the accuracy up to a point and after that accuracy is decreased or almost stable for both SVM and PSVM classifiers. So using more number of features after a limit may not increase accuracy further but only complicate the model.

Yap Bee Wah *et al.* [15] in their research compared filter and wrapper feature selection techniques to improve model efficiency. The filter-based feature selection techniques used are IG and correlation-based feature selection. Sequential backward and sequential forward elimination are used under wrapper based techniques. As per results, wrapper methods performed better as compared to the filter method. R studio is used for simulation of model. The results also show the impact of feature set size on the accuracy of the classifier. M. Cherrington *et al.* [16] mentioned various challenges involved in filter-based feature selection method.

Troussas *et al.* [17] Implemented boosting, bagging, voting and stack based ensemble classification models on three different data set. Stack based ensemble of SVM, NB, C4.5 and KNN is implemented using LR Meta classifier. The result demonstrates that stack-based model performed better than other classifiers.

M. Naz *et al.* [18] also implemented an ensemble of NB and KNN for sentiment prediction. Accuracy of the classification model is enhanced further by using NB, KNN and SVM based ensemble.

Doing variations in feature selection techniques, number of features used, and use of different ensemble classifiers can have an effect on efficiency of the model. Y. Emre Isik *et al.* [19] in their research implemented ensemble at two levels: one at feature selection level and the other at classifier level. An efficient feature selection can enhance efficiency of model without degrading the accuracy and classifier ensemble can also enhance accuracy of model. Both ensemble of feature selection and ensemble of classifiers are used to enhance the overall accuracy of the model. The technique shows good results as compared to other machine learning classifiers.

Joseph D. Pursa *et al.* [20] implemented ensemble classifiers in combination with various feature selection techniques. Changlin Zhou *et al.* [21] also used stack-based heterogeneous ensemble model along with wrapper-based feature ranking technique to resolve over fitting issue and enhance productivity prediction. The results show that hybrid wrapper based feature selection strategy reduces data complexity, improves comprehensiveness of model without reducing the accuracy of data. The technique used also performed better as compared to other base line models.

In the present research also, we are implementing various machine learning, lexicon and ensemble classifiers by using different number of features selected by chi-square and IG feature ranking and selection techniques. All the models are compared to see the effect on efficiency in terms of accuracy and Kappa values of all models.

3. FEATURE SELECTION TECHNIQUES

Feature selection is very important in machine learning classification while working on huge data sets having a large number of input features. Huge feature set makes the model complicated without even enhancing the efficiency prominently. The overall motive should be to select the minimum set of features without much affecting the efficiency of the model. Feature selection provides a way of reducing the number of input feature variables while developing a model. It reduces the dimensionality of the data set and simplifies the model [22]-[24]. Feature selection also reduces over fitting, reduces training time and may enhance accuracy of the model.

Statistical based feature selection methods find relationship between feature variables and target class. These methods select features with the strongest relationship with the class of target variable and have more influence while predicting the target variable. Feature selection methods are generally classified as supervised and unsupervised. The target class variable is also involved in removing irrelevant features in case of supervised feature selection methods. In unsupervised method, the target class is ignored for feature selection and it only finds the correlation between input variables to reduce the redundant variables. There are various statistical feature selection methods which are used for feature selection and reduction as mentioned:

3.1. Filter based ranking and feature selection

The filter method of feature selection calculates the relevance of features by using the intrinsic properties of features. It is a supervised technique that uses statistical techniques to calculate relation between input and target variables. A subset of the highly relevant features is selected and correlation matrix may be used for finally filtering out the features [25]-[26]. The filter method can be used as an attribute evaluator and ranker to rank all the features in the input feature set. The features with high rank can be selected for higher efficiency of the model.

The model starts with a complete set of features and various statistical techniques like ANOVA, chi-square, mutual information, IG, ReliefF and Pearson's correlation can be used to filter out the most relevant features [27]. As the filter-based feature selection does not depend upon classifier, so it is fast as compared to the wrapper method [28]. The disadvantage of this method is that it totally ignores the role of the classification algorithm in feature selection. In the present research we have used information gain and Chi-square techniques for feature selection.

3.1.1. Information gain

Information gain (IG) is used to select the splitting attribute at every node of the tree and the feature with highest information gain is selected. It actually calculates the reduction in entropy. IG can be used for feature ranking and selection by evaluating the Information gain of each input feature variable in the context of the target class variable. The input variables that maximize the information gain are selected which in turn minimizes the entropy and best splits the dataset into groups for efficient classification. Information gain is very effectively used in various researches for Twitter sentiment classification also [29]-[30] but Information

gain is biased for the input feature with higher number of distinct values. The (1) gives the formula for IG calculation as given under [31]:

$$I(A) = H(P(B)-H(P(B/A)) \quad (1)$$

A = Input Feature Variable
B = Target Class Variable
I (A) = Information gain of A with respect to target variable B.
P (B) = Marginal distribution of B assuming that it is independent of feature A.
P (B/A) = Distribution of B assuming that it is dependent of feature A.

3.1.2. Chi-square

Chi-Square is used when the features are categorical. The target class variable is categorical in case of Twitter sentiment classification as it holds the class as negative, positive and neutral in the data set used in present research. In case of feature selection, Chi-Square measures the relation between a feature and target variable in terms of $\chi 2$ value as mentioned in (2). A zero $\chi 2$ value means that there is no correlation between input feature variable and the target class. Higher the χ 2 value, higher is the role of input feature variable in predicting the target class. The value of Chi-Square is calculated between each input feature variable and the target class. Input features with the best Chi-Square values were selected as the final set of features. $\chi 2$ is one of the most effective feature selection techniques in sentiment analysis as it is quite robust as per the distribution of data and ease of computation [32], [33].

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (2)$$

O_i = Number of observation of class.
E_i = Number of expected observation of class if features and the target are independent.

3.2. Wrapper based feature selection method

The wrapper based feature selection work on greedy strategy that aims to find the best feature subset from a set of input feature variables by building a classification model by using a machine learning classifier [34]. Actually, we are trying to fit the machine learning classifier on the given data set. Multiple models are created in wrapper feature selection method by using different subsets of input features. Usefulness of a subset of feature is actually measured by training a model on it. The feature set with the best performing model as per performance metric is selected. This approach is relatively expensive compared to filtering methods due to repeated learning steps and cross-validation validation. Wrapper feature selection is comparatively more efficient in terms of accuracy of predicting class. There are various strategies used for the implementation of wrapper based feature selection:

a) Forward selection: Forward selection starts with a single feature. Features are added incrementally to the feature subset every time to increase the performance of the model. Features are added till the best results are obtained.
b) Backward elimination: In Backward selection, initially we start with the complete set of features and keep on removing features in such a manner to enhance the model efficiency. This is repeated to get the best feature subset for a given machine learning algorithm.
c) Recursive feature elimination: Recursive feature elimination or RFE tries to fit in a model by recursively removing weakest features from the input feature set at each iteration until the desired number of features is reached. The model is repetitively created and the next model is created with the remaining features of the previous model to further prune the least important features. Features are ranked based on elimination order.

3.3. Embedded feature selection method

Embedded feature selection methods are a combination of filter and wrapper based feature selection methods. These methods include feature interaction while keeping the computational cost reasonable. The embedded technique implement learning algorithm that have it's own built in feature selection methods. Regularization methods are most commonly used embedded methods which finalize the features on the basis of coefficient threshold. Decision tree, LASSO and RIDGE regression are examples of embedded feature selection. LASSO, RIDGE regression both have inbuilt penalization functions to reduce over fitting. The embedded methods includes the benefits of wrapper method like interaction of features, are fast like filter methods, have higher efficiency and are less prone to over fitting.

4. CLASSIFICATION TECHNIQUES

Classification in machine learning is a problem of identifying to which of a set of target classes a new observation belongs, on the basis of the input data set. Three classification techniques are involved in the research of the present paper as mentioned:

4.1. Lexicon based classification technique

In lexicon based or dictionary based classification technique, lexicon score of tweets or a piece of writing is calculated by using the sentiment dictionary [35]. In this approach, the polarity score of each word in BOG is calculated by using already present polarity dictionaries like WorldNet and SentiWordNet. The average polarity score of all words in a tweet is calculated. Average polarity score above a threshold value decides the polarity of a tweet as positive. This method is simple but does not provide as good results as machine learning and ensemble classifiers. Kolchyna *et al.* [36] compared lexicon based approach with Naive Bayes and SVM supervised classifiers. The supervised classifier gives nearly 7% better results than lexicon-based classification technique.

In the present research we have used a hybrid of lexicon-based approach with machine learning and ensemble classifiers. Although results of lexicon based classifiers are not as good as other classifiers, but a hybrid of lexicon based classifier with machine learning and ensemble classifiers can enhance accuracy with little overhead.

4.2. Machine learning classification techniques

Machine learning classifiers use supervised approach and need training examples which can be labeled manually or obtained from online sources. naive bayes (NB) [37], support vector machines [38], [39], decision tree [40], [41], AdaBoost, regression logistic regression, J48, Simple CART, random tree are some commonly used machine learning based classifiers. Kolchyna *et al.* [36] analyzed various machine learning classifiers. SVM and Naive Bayes perform better than other classifiers. It is also observed that machine learning approach performs better than lexicon-based approach. In the present research NB, SMO, IBK and REPTree are used as machine learning classifiers. Sequential minimal optimization (SMO) is an improved implementation of SVM. IBK is KNN and REPTree is fast decision tree in WEKA tool.

4.3. Ensemble classification techniques

Ensemble classifiers are used in various researches to further increase the accuracy of classification by using multiple classifiers together. Bagging, boosting and random forest, stack based ensemble and voting based ensemble are commonly used ensemble approaches. The idea is to combine several classifiers to give better results. Various researches done in the area of ensemble technique show that ensemble classifier mostly increases classification accuracy as compared to other two approaches [42]-[48].

5. DATA SET

For the present paper, the Dataset 'US Airline Sentiment data Corpus' used for the implementation and analysis is taken from standard Kaggle Data set repository. The Twitter US Airline Sentiment data Corpus has a total of 14640 tweets with sentiment target class. 3099 tweets are neutral, 2360 tweets are positive and 9178 tweets are negative. 70% data is used for testing and 30% for training purpose. The sample tweets in the data set are collected for six different US Airlines named United, Southwest, Delta, American, US Airways and Virgin America [49].

6. SIMULATION TOOL

WEKA and R studio tools are used in the present research. WEKA is an open source tool that can be used for data preprocessing, feature selection, clustering, regression, visualization, implementing several machine learning and ensemble classifiers and in various data mining tasks. WEKA is simple but quite effective simulation tools used by researchers. In the present research, WEKA is used for creating BOW, selection of features and implementing different classification models. At initial level, R Tool is also used for cleaning and removing noise from the data. Then the cleaned CSV files are used in WEKA for further processing [50], [51].

7. METHODOLOGY

For the implementation and analysis of the present research, "Twitter US Airline Data Corpus" is obtained from Kaggle data repository. This data set holds 14640 tweets and their corresponding class as

negative, positive and neutral. Tweets are cleaned and preprocessed by using R Tool, and the final CSV file is passed to WEKA for further processing. Tweets are represented as feature set and research is performed on different sized feature set for different machine learning and hybrid classifiers.

Data cleaning is done for removing irrelevant content from the tweets to improve efficiency of the model. R Tool is used to clean the data by removing URL's, Hashtags, numbers, emoticons, duplicate tweets and punctuation symbols. After cleaning, the cleaned data is passed to WEKA for further processing. This cleaned data is represented in the form of a feature set having huge dimensionality that needs to be reduced to simplify the model. Feature selection is performed by using Chi-Square and information gain feature selection methods and research is performed on feature sets of different sizes. To monitor the effect feature set size, feature sets with 301, 501, 701 and 1301 features are implemented for all machine learning and hybrid classifiers used in the present research. A further improvement is done by using lexicon-based classifier with different ensemble classifiers. Sentiment score retrieved from lexicon based classification algorithm is added in the feature set to get enhanced feature set. All the machine learning and ensemble classification model are again implemented for this enhanced feature set, also by varying again the number of features in the feature set. The results are monitored to observe any kind of improvement in all the models after adding sentiment score extracted from lexicon based classifier.

We are implementing SMO, NB, IBK, REPTree as machine learning algorithms. Meta bagging with REPTree is implemented for all size feature sets. In the first stack base ensemble model SMO, NB, REPTree are used as base classifiers and in the second stack base ensemble model SMO, IBK, REPTree are used as base classifiers. Random forest is used as meta learner in both stack based ensembles. Four different voting based ensemble models are implemented with average of voting and majority voting techniques. 70 percent of the input data is used for training purpose and the rest 30 percent is for testing the models. Results are analyzed and represented by using accuracy and Kappa metrics. All the finding and analysis is discussed in the next section of result and analysis. The methodology used in research is explained in Figure 1.

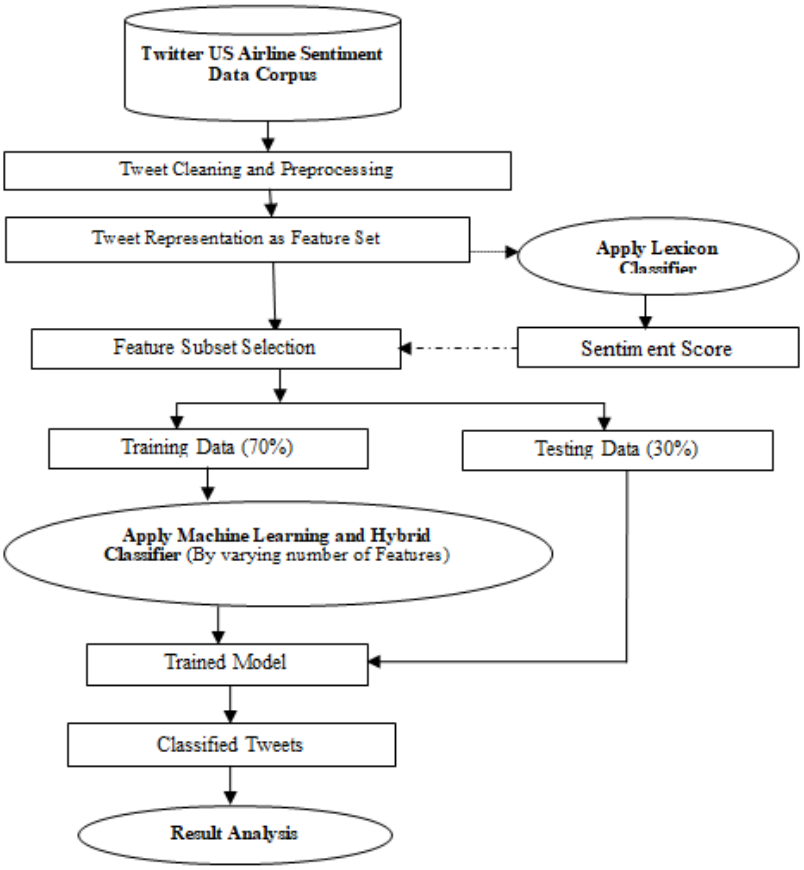

Figure 1. Work flow for various tweet sentiment classification models

8. RESULT AND ANALYSIS

In the present research, various machine learning and ensemble based classifiers are implemented on different sized feature set. Also, a hybrid of lexicon and stack-based ensemble classifiers is implemented. The results show that up to limit results are improved as the size of the feature set increases, but after that, accuracy is stable or decreases. Also, by using a hybrid of lexicon-based classifier with stack-based ensemble improves the results. The cost of getting and adding sentiment score extracted from lexicon based classifier is negligible, but it consistently increases the accuracy for all the classification models. Voting based classifiers are better than NB, IBK and REPTree machine learning classifiers. Accuracy and Kappa for various classification models using Information Gain feature selection is given in Table 1 and Table 2. Accuracy and Kappa for various classification models using Chi-Square feature selection is given in Table 3 and Table 4.

Findings of research and analysis:

a) Increase in the number of features from 301 to 701 enhances the accuracy, but accuracy is either stable or decreases by further increase in the number of input feature variables to 1301.
b) For all the classification models, machine learning or ensemble, adding lexicon score as one of the features in the feature set enhances the accuracy of classification. The hybrid of lexicon classifier with stack base classifier also enhances the accuracy of the model.
c) Stack based ensemble classifiers are efficient than machine learning and voting-based ensemble classifiers.
d) Stack based ensemble of SMO, NB, REPTree performed better than stack ensemble of SMO, IBK, REPTree.
e) SMO is comparatively better than other machine learning classifiers and accuracy is further increased by adding sentiment score of lexicon classifier in the input feature set.
f) IBK performance is less, almost in all cases for all size feature sets.
g) Chi- Square feature selection performed a little better than information gain in several findings.
h) All the above findings are uniform for both Chi- Square and information gain feature selection methods.

Table 1. Accuracy for various classification models using information gain feature selection method

ACCURACY for Different Classification Models using IG Feature Selection Method									
		Without Sentiment Score of Lexicon Classifier				With Sentiment score of Lexicon Classifier as one feature in Feature Set			
Sr. No	No of Features →	301	501	701	1300	302	502	702	1301
	Classification Method								
1	NB	72.9964	73.0191	75.1821	*75.136*	74.0665	74.2031	75.3643	75.2732
2	SMO	77.8461	78.6202	79.9863	78.847	78.4153	79.3033	80.4645	79.1439
3	REPTREE	71.8352	72.2222	72.3361	72.1767	72.4044	72.3588	71.3115	71.357
4	IBK	68.9891	68.3743	47.9053	47.4271	72.4044	64.3443	51.0018	49.6812
5	Meta Bagging + REPTree as Base Classifier	74.7951	74.5219	74.6585	74.4536	74.1576	74.1576	74.5219	74.4536
6	Meta Stacking using NB, SMO, REPTree (RF as Meta Classifier)	78.2332	79.326	80.601	79.418	79.212	80.123	81.162	79.504
7	Meta Stacking using IBK, SMO, REPTree (RF as Meta Classifier)	77.78	78.351	79.247	77.892	78.128	78.871	78.78	78.848
8	Voting using NB, SMO, REPTree Average Probability	75.9335	76.5027	77.3224	8.1193	76.5483	76.7987	78.1193	78.1193
9	Voting using IBK, SMO, REPTree Average Probability	75.7058	75.5237	73.5428	70.0137	72.8825	72.7231	68.1694	70.0137
10	Voting using NB, SMO, REPTree / Majority voting	75.8652	76.4117	77.5956	77.6184	76.184	76.4572	78.1193	78.1193
11	Voting using IBK, SMO, REPTree / Majority voting	75.2732	75.6375	77.0036	76.275	76.7304	76.8443	76.3434	75.6831

Table 2. Kappa for various classification models using information gain feature selection method

KAPPA for Different Classification Models using IG Feature Selection Method									
		Without Sentiment Score of Lexicon Classifier				With Sentiment score of Lexicon Classifier as one feature in Feature Set			
Sr. No	No of Features →	301	501	701	1300	302	502	702	1301
	Classification Method								
1	NB	0.4398	0.4404	0.5067	*0.5058*	0.4763	0.4792	0.509	0.507
2	SMO	0.5861	0.5998	0.6218	0.6028	0.5938	0.6099	0.630	0.607
3	REPTREE	0.481	0.4846	0.4343	0.4353	0.4651	0.4643	0.423	0.424
4	IBK	0.4527	0.4457	0.2348	0.232	0.4651	0.3821	0.250	0.242
5	Meta Bagging + REPTree as Base Classifier	0.5186	0.5135	0.4923	0.4791	0.4787	0.4794	0.480	0.479
6	Meta Stacking using NB, SMO, REPTree (RF as Meta Classifier)	0.5802	0.6020	0.6241	0.6050	0.6001	0.6180	0.636	0.607
7	Meta Stacking using: IBK, SMO, REPTree (RF as Meta Classifier)	0.5701	0.5800	0.6010	0.5720	0.5801	0.5920	0.590	0.592
8	Voting using NB, SMO, REPTree Average Probability	0.5345	0.5441	0.5416	0.5566	0.5335	0.5384	0.556	0.556
9	Voting using IBK, SMO, REPTree Average Probability	0.5374	0.5344	0.5105	0.4671	0.4906	0.4902	0.443	0.467
10	Voting using NB, SMO, REPTree Majority voting	0.5375	0.5458	0.5516	0.553	0.53	0.5333	0.562	0.562
11	Voting using IBK, SMO, REPTree Majority voting	0.5375	0.543	0.5647	0.5543	0.5543	0.5558	0.553	0.543

Table 3. Accuracy for various classification models using chi-square feature selection method

ACCURACY for Different Classification Models using Chi-Square Selection Method									
		Without Sentiment Score of Lexicon Classifier				With Sentiment score of Lexicon Classifier as one feature in Feature Set			
Sr. No	No of Features →	301	501	701	1300	302	502	702	1301
	Classification Method								
1	NB	74.795	74.8179	74.840	74.977	75.4781	75.3871	75.3415	75.0228
2	SMO	78.210	79.0301	79.394	79.007	78.3925	79.6676	79.8497	79.2122
3	REPTREE	72.586	71.9262	72.017	72.085	71.6758	71.6758	71.6302	71.0383
4	IBK	53.620	52.4362	48.793	48.315	54.2805	54.0984	50.5464	51.8215
5	Meta Bagging + REPTree as Base Classifier	74.658	74.0209	74.430	74.840	74.2031	74.408	74.2942	74.6129
6	Meta Stacking using NB, SMO, REPTree (RF as Meta Classifier)	78.643	79.5770	80.214	80.260	79.7360	80.4871	81.3901	81.1390
7	Meta Stacking using IBK, SMO, REPTree (RF as Meta Classifier)	77.869	78.5620	79.189	79.007	78.1880	78.9390	79.2370	78.9202
8	Voting using NB, SMO, REPTree /Average Probability	76.844	77.1403	77.140	77.527	77.2769	77.6639	77.5956	77.8233
9	Voting using IBK, SMO, REPTree /Average Probability	73.497	77.1403	73.178	72.085	68.3743	68.602	67.4863	69.515
10	Voting using NB, SMO, REPTree / Majority voting	76.889	77.1858	77.367	77.777	77.5273	77.8461	77.9372	77.9144
11	Voting using IBK, SMO, REPTree / Majority voting	75.9107	76.2295	76.184	75.5237	75.1138	76.0474	75.8652	75.5009

Table 4. Kappa for various classification models using chi-square feature selection method

KAPPA for Different Classification Models using Chi-Square Selection Method									
		Without Sentiment Score of Lexicon Classifier				With Sentiment score of Lexicon Classifier as one feature in Feature Set			
Sr. No	No of Features →	301	501	701	1300	302	502	702	1301
	Classification Method								
1	NB	0.5017	0.5016	0.5002	0.4993	0.5165	0.5146	0.5107	0.5002
2	SMO	0.5859	0.6018	0.6115	0.6058	0.5894	0.6139	0.6193	0.6094
3	REPTREE	0.4363	0.4292	0.4299	0.4327	0.4308	0.4372	0.423	0.4142
4	IBK	0.2797	0.2682	0.239	0.2383	0.2693	0.2697	0.2369	0.2519
5	Meta Bagging + REPTree as Base Classifier	0.4924	0.4816	0.4894	0.4924	0.4776	0.482	0.4789	0.4846
6	Meta Stacking using NB, SMO, REPTree (RF as Meta Classifier)	0.5870	0.6081	0.6202	0.6210	0.6110	0.6251	0.6401	0.6351
7	Meta Stacking using IBK, SMO, REPTree (RF as Meta Classifier)	0.5721	0.5880	0.5981	0.5940	0.5780	0.5941	0.6010	0.5941
8	Voting using NB, SMO, REPTree / Average Probability	0.5362	0.5414	0.5405	0.5442	0.5414	0.5488	0.5451	0.5476
9	Voting using IBK, SMO, REPTree / Average Probability	0.4992	0.5414	0.5019	0.4843	0.4289	0.4345	0.4228	0.4517
10	Voting using NB, SMO, REPTree / Majority voting	0.5404	0.5455	0.5478	0.5533	0.554	0.5606	0.5585	0.5561
11	Voting using IBK, SMO, REPTree / Majority voting	0.5348	0.5433	0.5479	0.5376	0.5217	0.5414	0.5383	0.5356

9. CONCLUSION

As Twitter is frequently used social networking site across the world, so the data collected from Twitter can be highly valuable for doing research. Here in this paper 'Twitter US Airline data' is used for sentiment analysis of customer's opinion. The research aims to see the effect of different feature selection techniques and different number of features on various machine learning and ensemble based classifiers. A hybrid of lexicon-based classifier with stack based ensemble classifier is also implemented for different sized feature set to observe the variations in finding. Results are improved with increase in feature set up to a limit; after that, accuracy is stable. Also, accuracy is increased by using the hybrid of lexicon classifier with other classification models. The present research is very useful in predicting the opinion of customers precisely and improving the service quality.

REFERENCES

[1] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, "Sentiment Analysis of Twitter Data," In *Proceedings of the workshop on Languages in Social Media*, Columbia University, New York, vol. 96, no. 1, pp. 30-38, 2011.

[2] S. Maheshwari, S. Shukla, D. Kumari, "Twitter Opinion Mining Using Sentiment Analysis, in World Scientific News," *An International Scientific Journal*, pp. 78-87, 2019.

[3] J. Zhao and X. Gui, "Comparison research on text pre-processing methods on Twitter sentiment analysis," *IEEE Access*, pp. 2870-2879, Feb. 2017, doi: 10.1109/ACCESS.2017.2672677.

[4] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," in *Proc. Research Gate Conference*, July 2016, doi: 10.1109/IISA.2016.7785373.

[5] B. Venkatesh and J. Anuradha, "A Review of Feature Selection and Its Methods", in *Journal of Cybernetics And Information Technologies*, vol 19, no 1, 2019, doi: 10.2478/cait-2019-0001.

[6] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Impact of Feature Selection Techniques for Tweet Sentiment Classification," *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference*, pp. 299-304, 2015.

[7] R. Ahujaa, A. Chuga, S. Kohlia, S. Guptaa, and P. Ahujaa, "The Impact of Features Extraction on the Sentiment Analysis," *International Conference on Pervasive Computing Advances and Applications-PerCAA 2019*, *Procedia Computer Science*, vol. 152, 2019, pp. 341-348. Class Association, doi: 10.1016/j.procs.2019.05.008.

[8] Z. Xu, J. Liu, Z. Yang, G. An, and X. Jia, "The impact of feature selection on defect prediction performance: An empirical comparison," In *Software Reliability Engineering (ISSRE), 2016 IEEE 27th International Symposium on*, pp. 309-320. IEEE, 2016, doi: 10.1109/ISSRE.2016.13.

[9] N. Rajput and S. Chauhan, "Analysis of Various Sentiment Analysis Techniques," *International Journal of Computer Science and Mobile Computing*, vol. 8, no. 2, pp. 75-79, Feb 2019, doi: 10.1109/ICICV50876.2021.9388525.

[10] S. Kurnaz and M. A. Mahmood, "Sentiment Analysis in Data of Twitter using Machine Learning Algorithms," *International Journal of Computer science and Mobile Computing*, vol. 8, no. 3, pp. 31-35, March 2019.

[11] D. Ray, "Lexicon Based Sentiment Analysis of Twitter Data," *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, vol. 5, no. x, Oct 2017, doi: 10.22214/ijraset.2017.10130.
[12] K. Lakshmi Devi, P. Subathra, and P. N. Kumar, "Tweet Sentiment Classification Using an Ensemble of Machine Learning Supervised Classifiers Employing Statistical Feature Selection Methods," *Proceedings of the 5th International Conference on "Fuzzy and Neuro Computing (FANCCO-2015)"*, vol. 415, pp. 1-13, Nov 2015, doi: 10.1007/978-3-319-27212-2_1.
[13] Esra Mahsereci Karabulut, Selma Ayşe Özel, and Turgay İbrikçi, "A comparative study on the effect of feature selection on classification accuracy", In *The Journal of Procedia Technology*, vol. 1, pp. 323-327, 2012, doi: 10.1016/j.protcy.2012.02.068.
[14] V. Sugumaran and K. I. Ramachandran, "Effect of number of features on classification of roller bearing faults using SVM and PSVM", in *Journal of Expert Systems With Applications*, vol. 38, no. 4, pp. 4088-4096, 2011, doi: 10.1016/j.eswa.2010.09.072.
[15] Yap Bee Wah, Nurain Ibrahim, Hamzah Abdul Hamid, Shuzlina Abdul-Rahman, and Simon Fong, "Feature Selection Methods: Case of Filter and Wrapper Approaches for Maximizing Classification Accuracy", in *The Pertanika Journal of Science & Technology*, vol. 26, no. 1, pp. 329-340, 2018.
[16] M. Cherrington, F. Thabtah, J. Lu, and Q. Xu, "Feature Selection: Filter Methods Performance Challenges," *2019 International Conference on Computer and Information Sciences (ICCIS)*, Sakaka, Saudi Arabia, pp. 1-4, 2019, doi: 10.1109/ICCISci.2019.8716478.
[17] C. Troussas, A. Krouska, and M. Virvou, "Evaluation of ensemble-based sentiment classifiers for Twitter data," in *7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Chalkidiki, pp. 1-6, 2016, doi: 10.1109/IISA.2016.7785380.
[18] M. Naz, K. Zafar, and A. Khan, "Ensemble Based Classification of Sentiments Using Forest Optimization Algorithm," *Big Network Inference, Integration and Analysis for Precision Medicine (BigDataNetAnalysis)*, vol. 4, no. 2, pp. 1-13, May 2019, doi: 10.3390/data4020076.
[19] Y. Emre Isik, Y. Görmez, O. Kaynar, And Z. Aydin, "NSEM: Novel Stacked Ensemble Method for Sentiment Analysis," *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)*, Malatya, Turkey, pp. 1-4, 2018, doi: 10.1109/IDAP.2018.8620913.
[20] Joseph D. Pursa, Taghi M. Khoshgoftaar, and N. Amri, "Using Feature Selection in Combination with Ensemble Learning Techniques to Improve Tweet Sentiment Classification Performance", *27th International Conference on "Tools with Artificial Intelligence (ICTAI)", IEEE*, pp 186-193, ISSN :1082-3409, 9-11 Nov. 2015, doi: 10.1109/ICTAI.2015.39.
[21] Changlin Zhou, *et al*, "A Novel Stacking Heterogeneous Ensemble Model with Hybrid Wrapper-Based Feature Selection for Reservoir Productivity Predictions", in *Hindawi special issue Complexity in Deep Neural Networks*, vol. 2021, Jan 2021, doi: 10.1155/2021/6675638.
[22] M. M. Fouad, T. F. Gharib, and A. S. Mashat, "Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble," in *International conference on Advanced Machine Learning and Applications*, vol. 723, pp. 517-527, Jan 2018, doi: 10.1007/978-3-319-74690-6_51.
[23] Deng, X., Li, Y., Weng, J. *et al.*, "Feature selection for text classification: A review," *Multimedia Tools and Appl*ications, pp. 3797-3816, 2019, doi: 10.1007/s11042-018-6083-5.
[24] R. Mansour, M. F. A. Hady, E. Hosam, H. Amr, and A. Ashour, "Feature Selection for Twitter Sentiment Analysis: An Experimental Study", *International Conference on Intelligent Text Processing and Computational Linguistics, Springer*, pp. 92-103, 2015, doi: 10.1007/s11042-018-6083-5.
[25] Lee, G. H. Lushington, and M. Visvanathan, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," *Journal of clinical Bioinformatics*, vol. 1, no. 11, pp. 1-8, 2011, doi: 10.1186/2043-9113-1-11.
[26] Daniel Mesafint and Manjaiah D. H, "Feature Selection Methods For Prediction Of The Individual's Status Of Hiv/Aids From Edhs Dataset-A Filter Approach", *IJREAT International Journal of Research in Engineering & Advanced Technology*, vol. 7, no. 3, June-July, ISSN: 2320-8791, 2019.
[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157-1182, 2003, doi: 10.1162/153244303322753616.
[28] B. Ghotra, S. McIntosh, and A. E. Hassan, "A Large-Scale Study of the Impact of Feature Selection Techniques on Defect Classification Models," *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*, Buenos Aires, pp. 146-157, 2017, doi: 10.1109/MSR.2017.18.
[29] Ni Made Gita Dwi Purnamasari, M. Ali Fauzi, Indriati, and Liana Shinta Dewi, "Cyber bullying identification in twitter using support vector machine and information gain based feature selection", *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 18, no. 3, pp. 1494-1500, ISSN: 2502-4752, June 2020, doi: 10.11591/ijeecs.v18.i3.pp1494-1500.
[30] Changki Lee and Gary Geunbae Lee, "Information gain and divergence-based feature selection for machine learning-based text categorization", *Information Processing & Management*, vol. 42, no. 1, pp. 155-165, ISSN 0306-4573, 2006, doi: 10.1016/j.ipm.2004.08.006.
[31] Samir Ifzarne, Hiba Tabbaa, and Imad Hafidi, Nidal Lamghari, "Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks," *Journal of Physics: Conference Series*, vol. 1743, 2021.
[32] Ikram Sumaiya Thaseen and Cherukuri Aswani Kumar, "Intrusion detection model using fusion of chi-square feature selection and multi class SVM", In *journal of Journal of King Saud University - Computer and Information Sciences*, vol. 29, no. 4, pp. 462-472, ISSN 1319-1578, 2017, doi: 10.1016/j.jksuci.2015.12.004.

[33] Sanur Sharma and Anurag Jain, "Hybrid Ensemble Learning With Feature Selection for Sentiment Classification in Social Media", *International Journal of Information Retrieval Research*, vol. 10, no. 2, April-June 2020, doi: 10.4018/IJIRR.2020040103.
[34] Ghosh, M., Guha, R., Sarkar, and R., Abraham A, "A wrapper-filter feature selection technique based on ant colony optimization", in *Journal of Neural Computation & Application*, pp. 7839-7857, 2020, doi: 10.1007/s00521-019-04171-3.
[35] M. Ahmad, M. Ferdy Octaviansyah, A. Kardiana, and K. Fadli Prasetyo, "Sentiment Analysis System of Indonesian Tweets using Lexicon and Naïve Bayes Approach," *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, Indonesia, pp. 1-5, 2019, doi: 10.1109/ICIC47613.2019.8985930.
[36] Olga Kolchyna, Th´arsis T. P. Souza, Philip C. Treleaven, and Tomaso Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination", Cornell University Library, 18 Sep 2015, arXiv :1507.00955.
[37] N. Ardhianie, R. Andreswari, and M. A. Hs, "Sentiment Analysis Of 'Indonesian No Dating Campaigns' on Twitter Using Naïve Bayes Algorithm," *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, Indonesia, pp. 116-120, 2019, doi: 10.1109/ISEMANTIC.2019.8884331.
[38] S. Naz, A. Sharan and N. Malik, "Sentiment Classification on Twitter Data Using Support Vector Machine," *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, Santiago, pp. 676-679, 2018, doi: 10.1109/ISEMANTIC.2019.8884331.
[39] R. N. Chory, M. Nasrun, and C. Setianingsih, "Sentiment Analysis on User Satisfaction Level of Mobile Data Services Using Support Vector Machine (SVM) Algorithm," *2018 IEEE International Conference on Internet of Things and Intelligence System*, Bali, pp. 194-200, 2018, doi: 10.1109/IOTAIS.2018.8600884.
[40] F. J. J. Joseph, "Twitter Based Outcome Predictions of 2019 Indian General Elections Using Decision Tree," *2019 4th International Conference on Information Technology (InCIT)*, Bangkok, Thailand, pp. 50-53, 2019, doi: 10.1109/INCIT.2019.8911975.
[41] R. Bibi, U. Qamar, M. Ansar and A. Shaheen, "Sentiment Analysis for Urdu News Tweets Using Decision Tree," *2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA)*, Honolulu, HI, USA, pp. 66-70, 2019, doi: 10.1109/SERA.2019.8886788.
[42] Joseph Prusa, Tahhi M. Khoshgoftaar, and David J. Dittman, "Using Ensemble Learners to Improve Classifier Performance on Tweet Sentiment Data," *Information Reuse and Integration (IRI), IEEE International Conference*, pp. 252-257, INSPEC Accession Number: 15556647, 13-15 Aug. 2015, doi: 10.1109/IRI.2015.49.
[43] Yun Wan and Qigang Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," *IEEE 15th International Conference on Data Mining Workshops*, 978-1-4673-8493- 3/15, pp 1318-1325, 2015, doi: 10.1109/ICDMW.2015.7.
[44] Nadia F. F. da Silva, Eduardo R. Hruschka, Estevam R. Hruschka Jr, "Tweet Sentiment Analysis with Classifier Ensembles," *ELSEVIER Journal On Decision Support System*, vol. 66, pp. 170-179, October 2014, doi: 10.1016/j.dss.2014.07.00.
[45] T. Subbulakshmi and R. Regin Raja, "An Ensemble Approach For Sentiment Classification: Voting For Classes and Against Them," *ICTACT Journal On Soft Computing*, vol. 06, no. 04, ISSN: 2229-6956 (ONLINE), July 2016, pp. 1281-1286, doi: 10.21917/ijsc.2016.0175.
[46] M. Hagen, M. Potthast, M. Büchner, and B. Stein, "Twitter Sentiment Detection via Ensemble Classification Using Averaged Confidence Scores", *Advances in Information Retrieval*, vol. 9022, ISSN: 0302- 9743, pp. 741-754, 2015, doi: 10.1007/978-3-319-16354-3_81.
[47] M. Ghiassi, J. Skinner, and D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network," in *Expert System with Applications, ScienceDirect*, vol. 30, no. 16, pp. 6266-6282, Nov 2013, doi: 10.1016/j.eswa.2013.05.057.
[48] J. J. Bird, A. Ekárt, C. D. Buckingham, and D. R. Faria, "High Resolution Sentiment Analysis by Ensemble Classification," In: *Arai K., Bhatia R., Kapoor S. (eds) Intelligent Computing. CompCom 2019. Advances in Intelligent Systems and Computing*, Springer, Cham, vol. 997, June 2019, doi: 10.1007/978-3-030-22871-2_40.
[49] Rane, A. and Kumar, A, "Sentiment classification system of Twitter data for US airline service analysis," In *Proceedings of the 42nd IEEE Computer Software and Applications Conference, COMPSAC 2018, Tokyo, Japan*, pp. 769-773, 2018, doi: 10.1109/COMPSAC.2018.00114.
[50] Al-Humoud, Sarah, AlBuhairi, Tarfa, and Altuwaijri, Mawaheb, "Arabic Sentiment Analysis using WEKA a Hybrid Learning Approach," *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, vol. 1, KDIR, pp. 402-408, 2015, doi: 10.5220/0005616004020408.
[51] Jain P. K and Pamula R, "Sentiment Analysis in Airline Data: Customer Rating Based Recommendation Prediction Using WEKA," In: *Das S., Das S., Dey N., Hassanien AE. (eds) Machine Learning Algorithms for Industrial Applications*. Studies in Computational Intelligence, vol 907. Springer, Cham, 2021, doi: 10.1007/978-3-030-50641-4_4.