ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Sentiment Analysis Using Hybrid Feature Selection Techniques

Sasan Sarbast Abdulkhaliq[1], Aso Darwesh[2]

[1]Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq, [2]Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq

ABSTRACT

Nowadays, people from every part of the world use social media and social networks to express their feelings toward different topics and aspects. One of the trendiest social media is Twitter, which is a microblogging website that provides a platform for its users to share their views and feelings about products, services, events, etc., in public. Which makes Twitter one of the most valuable sources for collecting and analyzing data by researchers and developers to reveal people sentiment about different topics and services, such as products of commercial companies, services, well-known people such as politicians and athletes, through classifying those sentiments into positive and negative. Classification of people sentiment could be automated through using machine learning algorithms and could be enhanced through using appropriate feature selection methods. We collected most recent tweets about (Amazon, Trump, Chelsea FC, CR7) using Twitter-Application Programming Interface and assigned sentiment score using lexicon rule-based approach, then proposed a machine learning model to improve classification accuracy through using hybrid feature selection method, namely, filter-based feature selection method Chi-square (Chi-2) plus wrapper-based binary coordinate ascent (Chi-2 + BCA) to select optimal subset of features from term frequency-inverse document frequency (TF-IDF) generated features for classification through support vector machine (SVM), and Bag of words generated features for logistic regression (LR) classifiers using different n-gram ranges. After comparing the hybrid (Chi-2 + BCA) method with (Chi-2) selected features, and also with the classifiers without feature subset selection, results show that the hybrid feature selection method increases classification accuracy in all cases. The maximum attained accuracy with LR is 86.55% using (1 + 2 + 3-g) range, with SVM is 85.575% using the unigram range, both in the CR7 dataset.

Index Terms: Binary Coordinate Ascent, Bag of Words, Chi-square, Logistic Regression, n-grams, Opinion Mining, Sentiment Analysis, Support Vector Machine, Twitter-Application Programming Interface, Term Frequency-Inverse Document Frequency

1. INTRODUCTION

In the past two decades, the internet and more specifically social media have become the main huge source of opinionated data. People broadly use social media such as Twitter, Facebook, Instagram to express their attitude and opinion toward things such as products of commercial companies, services, social issues, and political views in the form of short text. This steadily growing subjective data makes social media a tremendously rich source of information that could be exploited for the decision-making process [1], [2].

As mentioned before, one of the most popular and widespread social media is Twitter. It is a microblogging platform that allows people to express their feelings toward vital aspects in the form of a 280-character length text called

Access this article online

DOI: 10.21928/uhdjst.v4n1y2020.pp29-40

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Sasan Sarbast Abdulkhaliq, Department of Computer Science, University of Sulaimani, Sulaymaniyah, Iraq. E-mail: Sasan.abdulkhaliq@uhd.edu.iq

Received: 19-12-2019 Accepted: 07-01-2020 Published: 13-02-2020

tweet. Moreover, Twitter is used by almost all famous people and reputable companies around the world [3]. This has made Twitter become a very rich source for data, as it has approximately 947 million users and 500 million generated tweets each day. Hence, companies and organizations trying to benefit from this huge and useful data to find their customer's satisfaction with their products and service levels they offer, politicians wish to envisage their fans' sentiments. However, it is impractical for a human to analyze this massive data, to avoid this, sentiment analysis, or opinion mining techniques can be used to automatically discover knowledge and recognize predefined patterns within large sets of data [4].

Sentiment analysis is a natural language processing (NLP) technique for detecting or calculating the mood of people about a particular product or topic that has been expressed in the form of short text using machine learning algorithms. The main goal of sentiment analysis is to build a model to collect and analyze views of people about a particular topic and classify them into two main classes positive or negative sentiment [5].

One major step in sentiment analysis is feature extraction, which is the numerical representation of tokens in a given document. But actually, features can be noisy due to the data collection step as a consequence of data collecting technologies imperfection or the data itself which contains redundant and irrelevant information for a specific problem. This degrades the performance of the learning process, reduces the accuracy of classification models, increases computational complexity of a model, and leads to overfitting. Thus, high dimensionality problem should be handled when applying machine learning and data mining algorithms with data that have high dimensional nature. To handle this problem, feature selection techniques could be used to select the best features from available feature space for classification, regression, and clustering tasks.

Besides, one of the most important aspects of classification is accuracy. Feature selection plays a key role in improving accuracy by identifying and removing redundant and irrelevant features. Feature selection techniques are broadly classified into three categories, which are filter-based, wrapper-based, and embedded or hybrid methods [6].

In this work, Twitter-application programming interface (Twitter-API) being used to extract most recent tweet according to specific keywords such as (Amazon, Trump, Chelsea FC, CR7) and label them using lexicon-based approach into two categories which are positive and negative, followed by pre-processing step to remove irrelevant terms. Bag of word (BoW) technique and term frequency-inverse document frequency (TF-IDF) weighting scheme along with different n-gram ranges such as (Unigram, Bigram, Trigram, and Uni + Bi + Tri-gram) are used to extract features from tweets. The next step is selecting the best features to feed to our model which consists of a filter-based method Chi-square (Chi-2) to select the most relevant attributes within generated features, followed by selecting the best feature-subset within features using wrapper-based method binary coordinate ascent (BCA) to improve classification accuracy. Two supervised machine learning algorithm has been chosen for their performance and simplicity, namely, logistic regression (LR) and linear support vector machine (SVM) to perform binary sentiment classification on selected feature-subsets.

1.1. Related Works

In Zhai *et al.* [7], the authors utilize Chi-2 for feature selection with single and double words, together with SVM and Naïve Bayesian as classifiers. Obtained is that accuracy gradually increases as the number of features increase. With 1300 features accuracy hits 96%, then it remains slightly stable until 2000 features. Meanwhile, the accuracy of information (information gain [IG]) remains below Chi-2 for the whole features. Besides, we can see that the feature selection applied as combination features could also affect the performance of the classification. It extracts context-related multi-features with little redundancy which can help to reduce the internal redundancy, consequently improve the classification performance. Shortcomings of Chi-2 is also pointed, as it only considers the frequency of words within the document, regardless of the effectiveness of the word, and as result, it can cause the removal of some effective but low-frequency words during feature selection step.

In contrast, the researchers in Kurniawati and Pardede [8] proposed a hybrid feature selection technique on a balanced dataset, which composed of particle swarm optimization (PSO) plus IG together, followed by classification step using SVM, achieving a better result than using each one separately. The results are as follows: Compared to using SVM alone, the proposed method achieves 1.15% absolute improvements. Compared to IG + SVM and PSO + SVM, the method achieves 1.97% and 0.6% improvements, respectively. Overall, the system achieved 98% accuracy using area under the curve accuracy measure.

In the research done by Kaur *et al.* [9], the proposed system uses k-nearest neighbors (KNN) as a classifier for classifying

sentiments of text on e-commerce sites into positive, negative, and neutral sentiments on tweeter dataset. Features generated using n-gram before the KNN classifier took place. The performance of the proposed model was analyzed using precision, recall, and accuracy followed by comparing them to results obtained from the SVM classifier. The outcome was that the proposed system could outperform SVM classifier by 7%.

On the other hand, the work done by Zhang and Zheng [10] incorporates part of speech tagging to specify adverbs, adjectives, and verbs in the text first, then applied term frequency-inverse document frequency (TF-IDF) for generating features as a result of their corresponding word weights. Then, features were adopted for classification and fed to both SVM and extreme learning machine with kernels to classify sentiments of Hotel reviews in Chinese. They attained that essential medicines list accuracy is slightly better than SVM when introduced with the kernel and takes an effectively shorter time of training and testing than SVM.

In the work Joshi and Tekchandani [11], researchers have made a comparative study among supervised-learning algorithms of machine learning such as SVM, maximum entropy (MaxEnt), and Naïve Bayes (NB) to classify Twitter movie review sentiments using unigram, bigram, and unigram-bigram combination features.

Their study result shows that SVM reaches maximum accuracy of 84% using hybrid feature (unigram and bigram), leaving other algorithms behind. Furthermore, they observed that MaxEnt Excels NB algorithm when used with bigram feature.

In Luo and Luo [12] researchers proposed a new odds ratio (OR) + SVM-recursive feature elimination (RFE) algorithm that combines OR with a recursive SVM (SVM-RFE), which is an elimination based function. OR is used first as a filtering method to easily select a subset of features in a very fast manner, followed by applying SVM-RFE to precisely select a smaller subset of features. Their observation result emphasizes that OR + SVM-RFE attains better classification performance with a smaller subset of features.

In Maipradit *et al.* [13], a group of researchers suggests a method for classifying sentiments with a general framework for machine learning.

n-gram IDF has been used in feature generation and selection stage. As the classification stage, an automated machine learning tool has been used which makes use of auto-sklearn for choosing the best classifier for their datasets and also choosing the best parameter for those classifiers automatically. Classification is applied on different publicly available datasets (Stack Overflow, App reviews, Jira issues).

However, their study might not be feasible to be generalized for every other dataset; their datasets were specifically chosen for comments, reviews, and questions and answers. Their classification result achieved the best average model evaluation metrics F1, precision, and recall score values for all datasets in predicting class labels for positive, negative, and neutral classes for abovementioned datasets. Moreover, the highest F1-score value achieved was 0.893 in positive comments, 0.956 in negative comments of Jira issues dataset, and 0.904 F1-score value for in neutral comments of stack overflow dataset.

In work done by researchers in Rai *et al.* [14], tweets have been gathered from Twitter's API first. Later on weights for each word within review tweets have been calculated. Followed by selecting the best features using the NB algorithm, and consequently classifying the sentiment of reviews using three different machine learning classifiers, namely, NB classifier, SVM, and Random Forest Algorithm. After measuring they realized that all three algorithms are performing the same for 50 tweets, but increasing the number of tweets and adding more features changes the accuracy and other measures dramatically. As a part of their observation, they noticed that increasing the number of tweets from 50 to 250 will increase the accuracy of NB and SVM up to 83% approximately while adding more features to each algorithm gives slightly better classification accuracy up to 84% for 250 tweets.

Another group of researchers in Naz *et al.* [15] has employed another method to classify sentiments of Twitter data. The method composed of a model that employs a machine learning algorithm utilizing different feature combinations (unigram, bigram, trigram, and the combination of unigram + bigram + trigram) + SVM to improve classification accuracy. Furthermore, three different weighting approaches (TF and TF-IDF and binary) have been tried with the classifier using different feature combinations to see the effect of changing weights on classification accuracy. The best accuracy achieved by this approach was 79.6% using unigram with TF-IDF. Furthermore, sentiment score vector is created to save overall scores tweets and then associated with the feature vector of tweets, then classified them using SVM with different n-grams of features from different feature selection methods as mentioned before. The result shows that using a sentiment score vector with unigram + SVM gives the best accuracy result compared to other n-grams which were 81%.

Another research has been carried out by Wagh and Punde [16], a comparative study among different machine learning approaches have been applied by other researchers. The focus of their work was to discuss the sentiment analysis of Twitter tweets, considering what people like or dislike. They perceived that applying machine learning algorithms such as SVM, NB, and Max-Entropy on results of semantic analysis WordNet to form hybrid approach can improve accuracy of sentiment analysis classification by 4–5% approximately.

Another research has been performed by Iqbal *et al.* [17], in which multiple feature combinations are fed to (NB, SVM, and MaxEnt) classifiers for classifying movie reviews from the IMDb dataset and tweets from Stanford Twitter sentiment 140 dataset, in the term of people's opinion about them. The experiment incorporates four different sets of features, each of which are a combination of different single features as following: Combined single word features with stopword filtered word features as (set 1), unigram with bigram features as (set 2), bigram with stopword filtered word features as(set 3), and most informative unigram with most informative bigram features as(set 4). Chi-2 has been used as a supervised feature selection technique to obtain more enhanced performance by selecting the most informative features. And also, Chi-2 helps to decline the size of training data. Their result shows that combining both unigram and bigram features and subsequently feeding it to MaxEnt algorithm gives the best result in term of F1-score, precision, and recall compared to two other algorithms, and also compared to using single feature and baseline model which is SentiWordnet (SWN) method by 2–5%.

Another research was done by Rane and Kumar [18] on a dataset containing tweets of 6 US Airlines and carried out sentiment analysis to extract sentiments as (positive, negative, and neutral). The motivation of the research was to provide airline companies a general view of their customer's opinions about airline services to provide them a good level of service to them. As the first step preprocessing has been performed, followed by a deep learning concept (Doc2vec) to represent tweets as vectors which makes use of distributed BoW and distributed memory model, which preserves ordering of words throughout a paragraph, to do phrase-level sentiment analysis. The classification task has been done using seven different supervised and unsupervised learning, namely, decision tree, random forest, SVM, KNN, LR, Gaussian NB, and AdaBoost. After classification, they attained acceptable accuracy that can be used by airline companies with most of the classifiers as follows: Random forest (85.6%), SVM (81.2%), AdaBoost (84.5%), and LR (81%) are among the best classifiers as result, they concluded that the accuracy of the classifiers are high enough, that makes them reliable to be used by airline industry to explore customer satisfaction.

Another work done by Jovi *et al.* [19] to review available feature selection approaches for classification, clustering and regression tasks, along with focusing on their application aspects. Among which IG (precision) and normal separation (accuracy, F-measure, and recall) have the best performance for text classification tasks, whereas iterative feature selection (Entropy, precision) attains the best performance for text clustering. Results show that using hybrid approaches for feature selection, consisting of a combination of the best properties from filter, and wrapper methods giving out the best result by applying first, a filter method to reduce feature dimensions to obtain some candidate subsets. Then applying a wrapper method was based on a greedy approach to find the best candidate subset.

In Rana and Singh [20] authors have proposed a model for classifying movie reviews using NB classifier and Linear SVM classifiers. They realized that applying the classifiers after omitting synthetic words gives a more accurate result. Their result shows that SVM achieves better accuracy than the NB classifier. Furthermore, both algorithms distinctly performed better for genre drama, reaching 87% with SVM and 80% with the NB algorithm.

In Kumar *et al.* [21] authors have developed a classification model to classify reviews from websites such as amazon.com. After extracting reviews of three different products, namely, APPLE IPHONE 5S, SAMSUNG J7, and REDMI NOTE 3 from the website automatically, they applied NB, LR, and SWN algorithms for classifying reviews in the term of positive and negative. After using quality measure metrics (F1 score, recall, and precision), NB has achieved the best result among three classifiers with F1-scores: 0.760, 0.857, and 0.802 for three above-mentioned datasets, respectively.

In Iqbal *et al.* [22] researchers proposed a hybrid framework to solve scalability problems that appear when feature set grows in sentiment analysis. Using genetic algorithm (GA) based technique to reduce feature set size up to 42% without effecting accuracy. Comparing their proposed (GA) based feature reduction technique against two other well-known techniques: Principal Component Analysis (PCA) and Latent Symantec Analysis, they affirmed that GA based technique had 15.4% increased accuracy over PCA and up to 40.2% increased accuracy over latent semantic analysis. Furthermore, they employed three different methods of sentiment analysis, which are SWN, Machine Learning, and

Machine Learning with GA optimized feature selection. In all cases, the SWN approach has lower accuracy than two other mentioned approaches achieving its best accuracy of 56%, which impractical for real-time analysis. Their developed model which incorporates GA results in reducing feature size by 36–43% in addition to 5% increased efficiency when compared to the ML approach due to reduced feature size. They have tested their proposed model using six different classifiers on different datasets, the classifiers are, namely, J48, NB, PART, sequential minimal optimization, IB-k, and JRIP. Among all classifiers, NB classifier has shown the highest accuracy (about 80%) while using GA based feature selection on Twitter and reviews dataset, on the other hand, IB-k outperformed other classifiers with accuracy 95% while applying on the geopolitical dataset. Another evaluation is done for the scalability and usability of their proposed technique using execution time comparison. They found that the system showed a linear speedup with the increased dataset size. However, the technique consumed 60–70% of the aggregate execution time on customer reviews dataset, but it results in a speedup of modeling the classifiers up to 55% and remains linear, confirming that proposed algorithm is fast, accurate, and scales well as the dataset size grows.

1.2. Problem Statement

Thus, twitter is one of the richest sources of opinionated data; there is a big demand on analyzing twitters' data nowadays for the process of decision making. However, these data are unstructured and contain a lot of irrelevant and redundant information that leads to high-dimensionality of feature space, consequently analyzing them properly and accurately by machine learning, and data mining techniques is a big challenge. High dimensionality degrades the performance of the learning process, reduces the accuracy of classification models, increases computational complexity of a model, and leads to model overfitting. To overcome this problem, a hybrid of two feature selection methods proposed to remove the redundant and irrelevant features to select the best feature subset for classification task automatically. In this work, we use Chi-2 to calculate the correlation between attributes and the class labels. Low correlation of a particular feature means that the feature is irrelevant to the class label and needs to be removed prior to classification. In this way features are reduced. but there is still the problem of redundant features. The redundant features were removed by applying BCA, which uses an objective function to selects optimal feature subset from features were selected by Chi-2. As result irrelevant and redundant features were removed, which lead to solve high dimensionality problem in model building, and eventually classification accuracy is improved.

2. METHODOLOGY

The main objective of this study is to select optimal or sub-optimal feature-subset to perform Twitter sentiment analysis throughout utilizing filter-based method (Chi-2) and hybrid filter + wrapper method, namely, Chi-2 + BCA to improve the accuracy of the classification model. The work was implemented on Acer vn7-591g series laptop, Intel(R) Core(TM) i7-4710HQ CPU at 2.50 GHz (8 CPUs), 16-GB RAM, Windows 10 Home 64-bit. Fig. 1 depicts the flow diagram of our proposed system model and the following subsections explain each step of developing the proposed model in detail:

2.1. Data Collection

Twitter-API with python code is used to automatically download most recent tweets about (Amazon, Trump, Chelsea FC, and CR7) keywords, respectively, and lexicon rule-based method being utilized to assign positive and negative scores for each tweet, then protecting it in Comma Separated Value (.csv) file. Table 1 illustrates the details of each keyword dataset and Fig. 2 shows a sample of collected tweets.

2.2. Pre-processing

Text pre-processing is the first step in Twitter sentiment analysis, the tweets should go through some pre-processing step such as removing duplicate tweets, converting to lowercase, replacing emoji's with their meaning, removing URLs, usernames, and expand contractions such as (can't → cannot), replacing slang word like (omg → oh my god), reducing repeated character to only two character, removing numbers, special characters, punctuation marks, multiple space, tokenizing, removing stop-words, and lemmatizing, followed by removing duplicate tweets after pre-processing. Finally persisting the cleaned in another (.csv) file. Table 2 shows the emoji's and their meaning.

TABLE 1: Dataset size description

Keyword	Positive	Negative	Total # of tweets
Amazon	432	791	1223
Trump	1054	1200	2254
Chelsea FC	2239	761	3000
CR7	2816	1184	4000

TABLE 2: Emoji's and their meaning

Emoji	Meaning
:), :-D, :-j, =p, :3	Positive
:(, :|, :^), +o(, :=&	Negative

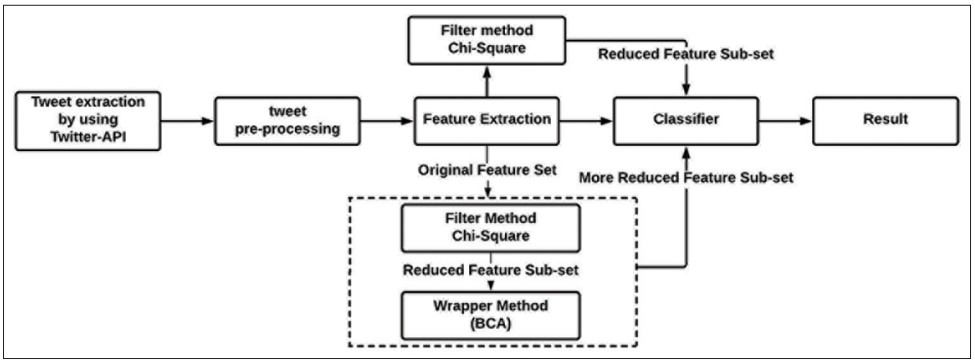

Fig. 1. Flow diagram of proposed system model.

	text	label
0	Trump Says Beijing Playing Vicious Game in Tar...	-1
1	Take comfort, people in Florida, knowing that ...	1
2	Once again placing his personal business befor...	-1
3	@JoeBiden As we said to Clinton - if you GENUI...	-1
4	Texans are funny like that... https://t.co/vMy...	1

Fig. 2. Sample of collected tweets.

2.3. Feature Extraction

Feature extraction is the process of converting the text data into a set of features or numerical representations of words or phrases. The performance of the machine learning process depends heavily on its features, so it is crucial to choose appropriate features for your classification model. On the other hand, applying different n-grams which are a different combination of words within the document gives out different accuracy results. We used (unigram, bigram, trigram, and combination of all) as the most commonly used ranges to see the impact of each of them on classification results. The followings are two Feature Extraction Methods used by the proposed model

1) TF-IDF: TF-IDF stands for Term Frequency – Inverse Document Frequency, it is a simple and effective metric that represents how "important" a word is to a document in the document set. It has many uses; one of its common uses is for automated text analysis. It is very useful for scoring words in machine learning algorithms in NLP.
   TF-IDF for a word in a document is calculated by multiplying two different metrics:
   - TF: Calculating how many times a term occurs in a document. The reason behind using it is that words that frequently occur in a document are probably more important than words that rarely occur. The result is then normalized by dividing it by the number of words in the whole document. This normalization is done to prevent a bias toward longer documents.
     TF(t) = (Number of times term t appears in a document)/(Total number of terms in the document).
     And its mathematical formula is:

$$tf_{td} = \frac{n_{td}}{\sum_k n_{kd}} \quad (1)$$

     Where n_{td} is the number of times that term t occurs in document d, and n_{kd} is the number of occurrences of every term in document d.
   - IDF: Measures how important term is by taking the total number of documents in the corpus and divide it by the number of documents where the term appears. It is calculated by:
     IDF(t) = log(Total number of documents/Number of documents with term t in it).
     And its mathematical formula is:

$$idf_t = \log \frac{|D|}{|D_t|} \quad (2)$$

     Where $|D_t|$ is the total number of documents in the corpus, $|D_t|$ is the number of documents where the term t appears in it.
     Hence, the TF-IDF is the multiplication of eq. 1 and eq. 2. Which is:

$$tf - idf_t = \frac{n_{td}}{\sum_k n_{kd}} \log \frac{|D|}{|D_t|} \quad (3)$$

2) BoW: One of the simplest types of feature extraction models is called BoW. The name BoW refers to the fact that it does not take the order of the words into account. Instead one can imagine that every word is put into a bag. It simply counts the number of occurrences of each word within a document and keeps the result in a vector which is known as count-vector.

2.4. Feature Selection

The main goal of feature selection is to select an optimal group of features for learning algorithms from the original feature set to retain the most useful features as possible and removes useless features that do not affect classification result. In this way, feature selection reduces the high dimensionality of data by eliminating irrelevant and redundant features. Thus, improves the model accuracy, reduces computation, and training time. It also reduces storage requirement, and avoids overfitting. Feature selection methods mainly divided into three categories which are Filter Methods, Wrapper Methods, and Hybrid or embedded methods. In general, feature selection methods are composed of four main steps, namely, feature subset generation, subset evaluation, stopping criterion, and result validation. Fig. 3 illustrates the basic steps of the feature selection process.

In this work, we are using Chi-2 filter-based method in conjunction with BCA wrapper-based method to form a hybrid feature subset selection technique to select the best subset for our classification models. First, we employed Chi-2 to remove irrelevant features, leading to produce reduced feature set. Then applied BCA for selecting more optimal subset of features that are more reduced feature subset. The operation details of both methods are described below:

2.5. Chi-2

Chi-2 is a type of filter-based feature selection method; it is used to select informative features and ranking them to remove irrelevant features with low ranks. In statistics, the Chi-2 test is used to examine the independence of two events. The events, and are assumed to be independent if:

$$p(XY) = p(X)\,p(Y) \tag{4}$$

In text feature selection, these two events correspond to the occurrence of a particular term and a class, respectively. Chi-2 can be computed using the following formula:

$$Chi-2(t,C) = \sum_{t\in\{0,1\}}\sum_{C\in\{0,1\}} \frac{(N_t,C-E_t,C)^2}{E_t,C} \tag{5}$$

N is the observed frequency, and E is expected frequency for both of term t and Class C. CHI2 is a measure of how much expected count E and observed count N deviate from each other. A high value of Chi-2 indicates that the hypothesis of independence is not correct. The occurrence of the term makes the occurrence of the class more likely if the two events are dependent. Consequently, the regarding term is relevant as a feature. The Chi-2 score of a term is calculated for individual classes. This score can be globalized over all classes in two ways. The first way is to compute the weighted average score for all classes and the second way is to choose the maximum score among all classes. In this paper, the former approach is preferred to globalize the Chi-2 value for all classes in the corpus.

$$\sum P(C_i).\text{Chi}-2(t,C_i) \tag{6}$$

Where $P(C_i)$ is the class probability and Chi-2 (t, C_i) is the class-specific Chi-2 score of term t.

2.6. BCA

BCA is a wrapper based feature selection method, was introduced by Zarshenas and Suzuki [23] in 2016. The goal of the BCA algorithm is to choose optimal or sub-optimal sub-set from available features from feature space that makes machine-learning algorithms the highest possible performance for a specific task, such as classification. The BCA algorithm iteratively adds and removes features to and from the selected subset of features based on the objective function values starting from an empty sub-set. At each iteration, the BCA checks whether the existence of a particular feature, in a given subset of features, improves or degrades the classification performance. If a feature was included in or removed accidentally from the feature sub-set, the BCA algorithm will be capable of correcting the wrongly taken decisions in the proceeding scans to approximate the optimal solution as much as possible. Compared to Sequential Feature Selection (SFS) and Sequential Forward Floating Selection (SFFS) are two of the most popular wrapper-based FSS techniques and filter-wrapper Incremental Wrapper Subset Selection (IWSSr), the BCA is more efficient than

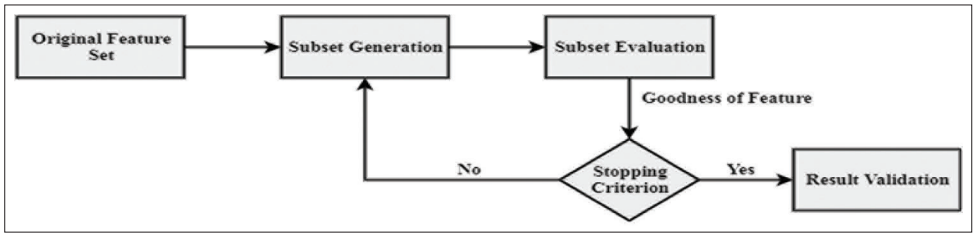

Fig. 3. A general framework of feature selection.

both of them in terms of processing time and classification accuracy. We came to the point that this algorithm is an effective feature selection method for the classification of datasets with a high number of initial attributes. Fig. 4 shows the work of the algorithm.

2.7. Hybrid Feature Selection Method

Filter methods are fast because they use mathematics and statistics for selecting features. The proposed model uses Chi-2 as filter-based method to remove irrelevant features and producing a reduced feature subset from the original feature set. In addition, wrapper-based methods are more accurate because they work as a part of the classification algorithms to evaluate usefulness of a particular feature, but they are computationally slow when applied on original feature set. Hence, the proposed model takes advantage of characteristics of both feature selection methods by first removing irrelevant features from the original feature set using Chi-2, and then applying BCA to the features those are selected by Chi-2 to select more optimal feature subset to enhance classification accuracy.

2.8. Classification Algorithms

In sentiment analysis classification essentially means categorizing data into different classes based on some calculation to determines the sentiment of the text. In our study, we applied two machine learning algorithms, namely, Linear SVM (LSVM) and LR for binary classification (positive and negative) of Twitter data.

SVM is a non-probabilistic machine learning algorithm. It is primarily used for classification in machine learning and could be fine-tuned for using with regression. The aim of SVM is to find the optimal decision boundary between classes

By transforming our data with the help of mathematical functions called Kernels. The best decision boundary created is called a hyperplane.

With a linearly separable data linear kernel is used. Since our class labels are linear (only positive and negative), we will perform classification with "linear SVM."

LR is a statistical machine learning algorithm for predicting classes which have dichotomous nature. Dichotomous mean having just two possible classes, binary by another mean. The term logistic mean logit function (a probabilistic function which returns values just in [0,1]).

3. RESULTS AND DISCUSSION

Based on the results attained from the two classifiers: TF-IDF with SVM and BoW with LR along with different n-grams, five-fold cross-validation, using Chi-2 and Hybrid Chi-2 + BCA feature selection methods, we achieved accuracy levels illustrated in the following Figs. 5-12:

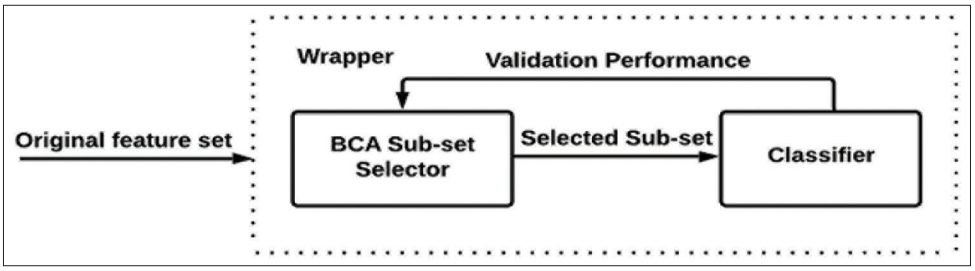

Fig. 4. Binary coordinate ascent.

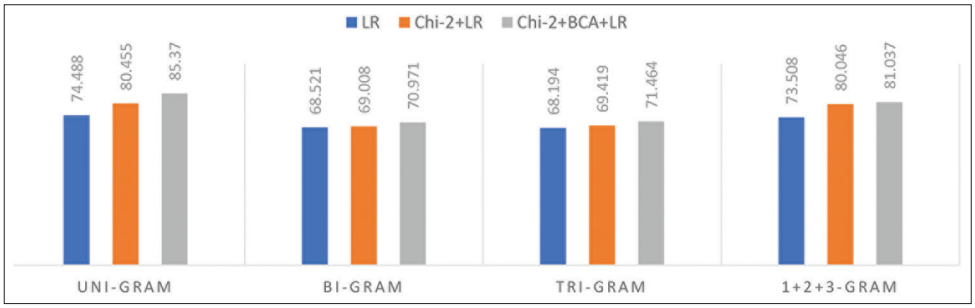

Fig. 5. Accuracies of Amazon dataset.

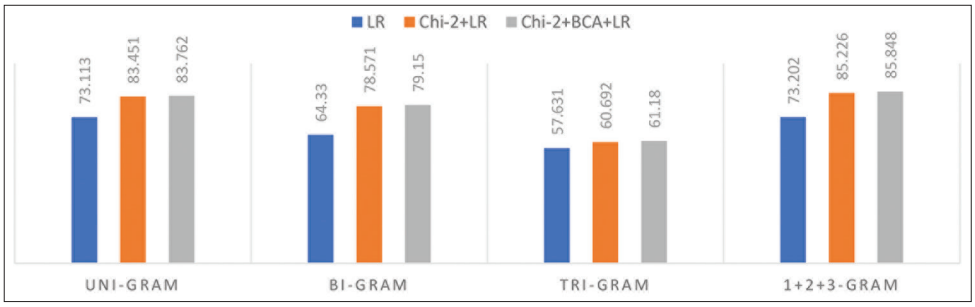

Fig. 6. Accuracies of Trump data set.

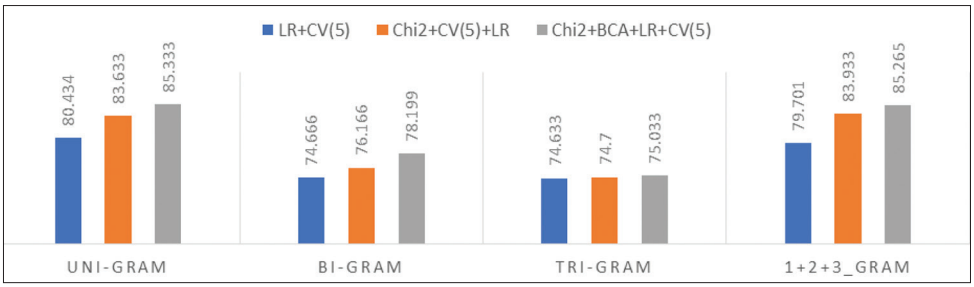

Fig. 7. Accuracies of Chelsea FC data set.

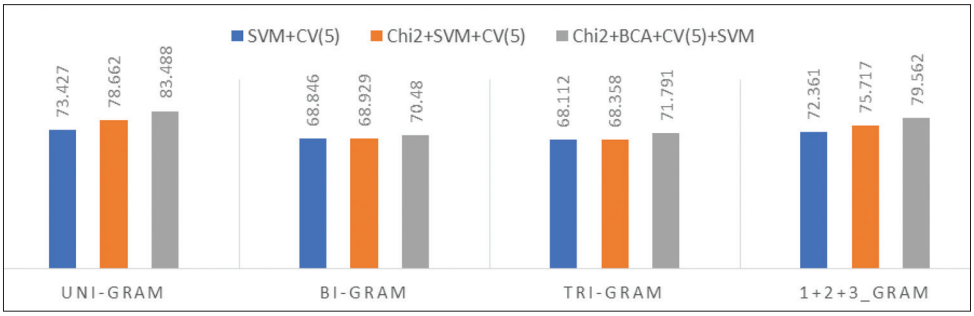

Fig. 8. Accuracies of CR7 data set.

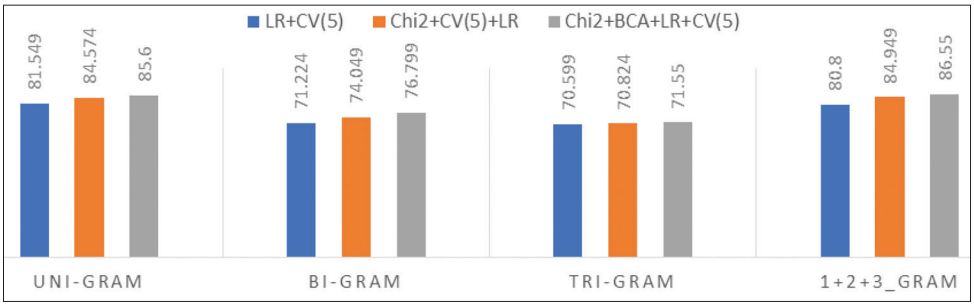

Fig. 9. Accuracies of Amazon dataset.

The graph shows that LR classifier attains best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively. However, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 5%, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 5% with unigram, followed by the rest three approaches bigram and trigram and 1 + 2 + 3-g with a slight increase.

The graph shows that LR classifier attains best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively, which is in all cases less than Chi-2 result, that achieve big raise in accuracy with unigram by 10% bigram

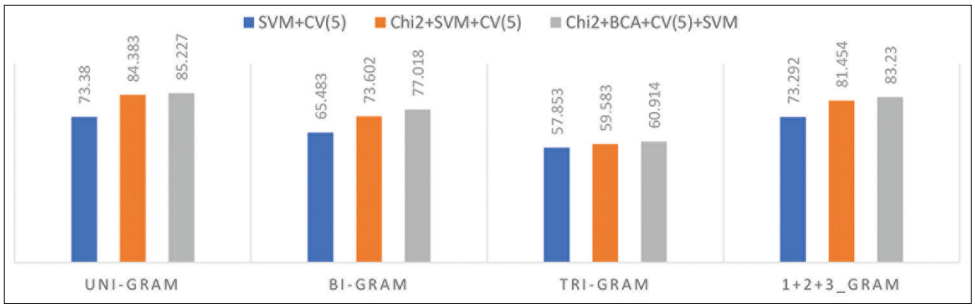

Fig. 10. Accuracies of Trump dataset.

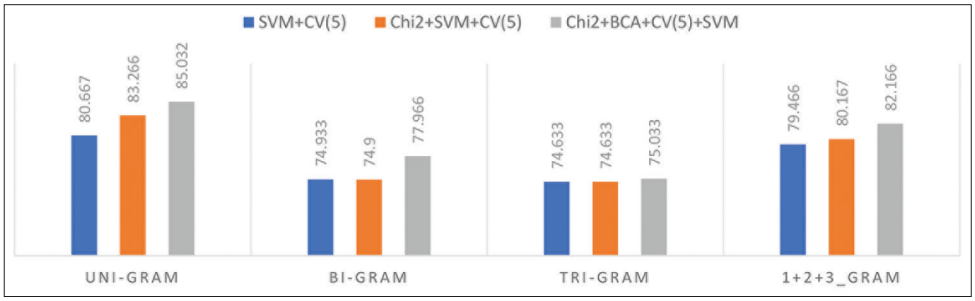

Fig. 11. Accuracies of Chelsea FC dataset.

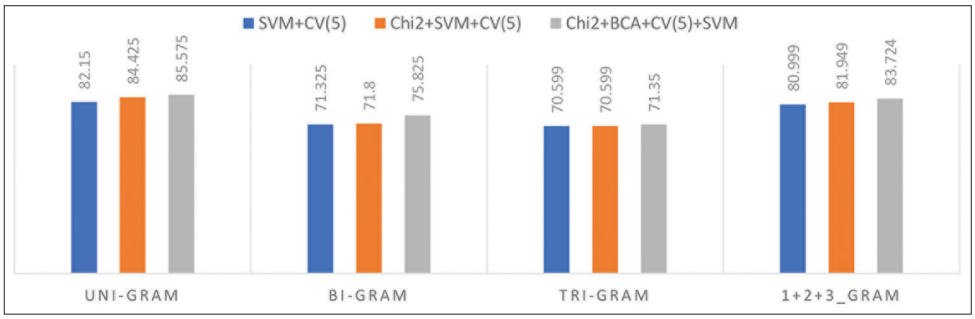

Fig. 12. Accuracies of CR7 dataset.

14% and 1 + 2 + 3-g with 10% followed by trigram with 3% increase. Finally, applying BCA accuracy has a slight raise in all cases with less than 1%.

The graph shows that LR classifier attains best accuracy with unigram range followed by 1 + 2 + 3-g, bigram, and trigram, respectively. After applying Chi-2, unigram and 1 + 2 + 3-g accuracy increased with more than 3%, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a better accuracy rate increase in all cases. In unigram bigram, and 1 + 2 + 3-g accuracy increased by more than 1.5%, followed by trigram with a small increase.

The graph shows that LR classifier attains best result with 81.549% in unigram range followed by 1 + 2 + 3-g, bigram, and trigram respectively. Consequently, after applying Chi-2, unigram, 1 + 2 + 3-g, and bigram accuracy increased by more than 3%, followed by and trigram with a small increase. Finally, applying BCA achieves even more accuracy rate increased by approximately 1.5% with unigram, bigram, and 1 + 2 + 3-g, and a small change can be observed with trigram.

All bar charts illustrate that the accuracy attained from Hybrid Chi-2 +BCA outperforms the accuracy of LR, LR when applied only with features selected by Chi-2, in all n-gram ranges. Moreover, unigram and (1 + 2 + 3-g) achieve higher results than bigram and trigram in (Chi-2+BCA) feature selection. Results also show that with the growth of datasets, the accuracy of the classifier increases.

The graph shows that SVM classifier attains the best result in unigram range followed by 1 + 2 + 3-g, bigram, and trigram,

respectively. Then, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 3% and 5%, respectively, followed by bigram and trigram with a slight increase. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 5% with unigram and (1 + 2 + 3), followed by bigram and trigram with a slight increase.

The graph shows that SVM classifier attains the best result in unigram and 1 + 2 + 3-g, followed by bigram and trigram, respectively. Then, Applying Chi-2, unigram, bigram, and 1 + 2 + 3-g accuracy dramatically increased with more than 10%, 8%, and 7%, respectively, followed by trigram with a slight increase. Finally, applying BCA to features selected by Chi-2 achieves a dramatical accuracy rate increase by approximately 3%, and more than 1% with unigram and (1 + 2 + 3), followed by and trigram with a slight increase.

The graph shows that SVM classifier achieves the best result in unigram range, followed by 1 + 2 + 3-g, bi-gram, and tri-gram, respectively. However, after applying Chi-2, unigram accuracy dramatically increased with more than 3%, followed by 1 + 2 + 3-g, bigram, and trigram with a slight increase. Finally, applying BCA to features selected by Chi-2 achieves a dramatical accuracy increase by approximately 3% with unigram, followed by 1 + 2 + 3, bigram, and trigram with a slight increase.

The graph shows that SVM classifier attains the best result in unigram range, followed by 1 + 2 + 3-g, bigram, and trigram, respectively. However, after applying Chi-2, unigram, and 1 + 2 + 3-g accuracy increased with more than 2% and 1%, respectively, followed by bigram with a slight change, while trigram increase remained same. Finally, applying BCA achieves a dramatical accuracy rate increase by approximately 4% with bigram, 1% with unigram, and more than 2% with 1 + 2 + 3, followed by trigram with a slight increase.

All Bar charts illustrate that the accuracy achieved from Hybrid Chi-2+BCA outperforms the accuracy of SVM, SVM when applied only with features selected by Chi-2, in all n-gram ranges. Moreover, unigram and 1 + 2 + 3-g achieve higher results than bigram and trigram in most cases of (Chi-2+BCA) feature selection. Results also show that with the growth of datasets, the accuracy of the classifier increases, same as with LR.

4. CONCLUSION

In the context of our work, we developed a sentiment classification model for classifying tweets into positive and negative based on the sentiment of the author. As the amount of data becomes huge, the task of classifying them becomes more challenge and the need for reducing the number of features arises to improve classification accuracy. We proposed a hybrid feature selection method by incorporating a filter-based method Chi-2, followed by wrapper-based method BCA for reducing the number of irrelevant features and selecting optimal or sub-optimal features respectively, from features generated by BoW and TF-IDF, each of which used with a different classifier. After training our model with different n-gram ranges and five-fold cross-validation, we conclude that applying our proposed hybrid feature selection method (Chi-2+BCA) reduces features and improves classification performance in the term of accuracy up to 11.847% compared to using original feature set with linear SVM and 10.882% with LR classifiers, both with unigram range. Moreover, the maximum improvement of Chi-2+BCA over using only Chi-2 was 4.915% and 4.826% for LR and SVM, respectively.

4.1. Future Work

Using the same system with a greater number of tweets to inspect the effectiveness of BCA with the growth of the dataset. Using BCA as a feature subset selection algorithm with deep learning algorithms such as LSTM and RNN. Applying BCA to other feature generation techniques such as word2vec or doc2vec. Hybridizing BCA with other filter methods.

REFERENCES

[1] H. P. Patil and M. Atique. "Sentiment Analysis for Social Media: A Survey". 2015 *IEEE 2nd International Conference Information Science Secur*, 2016.

[2] M. K. Das, B. Padhy and B. K. Mishra. "Opinion Mining and Sentiment Classification: A Review". *Proceeding International Conference Inventory System Control*, pp. 4-6.

[3] A. S. Al Shammari. "Real-time Twitter Sentiment Analysis using 3-way classifier". *21st Saudi Computer Society National Computer Conference's*, pp. 1-3, 2018.

[4] R. D. Desai. "Sentiment Analysis of Twitter Data". *Proceeding 2nd International Conference Intelligence Computing Control System* no. Iciccs, pp. 114-117, 2019.

[5] P. M. Mathapati, A. S. Shahapurkar and K. D. Hanabaratti. "Sentiment Analysis using Naïve Bayes Algorithm". *International Journal of Computational Science and Engineering*, vol. 5, no. 7, pp. 75-77, 2017.

[6] N. Krishnaveni and V. Radha. "Feature Selection Algorithms for Data Mining Classification: A Survey". *Indian Journal of Science and Technology*, vol. 12, no. 6, pp. 1-11, 2019.

[7] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao. "A Chi-square Statistics Based Feature Selection". *2018 IEEE 9th Internatinal Conference Software Engineering Services Science*, pp. 160-163, 2018.

[8] I. Kurniawati and H. F. Pardede. “Hybrid Method of Information Gain and Particle Swarm Optimization for Selection of Features of SVM-Based Sentiment Analysis”. *2018 Internatinal Conference Information Technology System innovation*, pp. 1-5, 2019.

[9] S. Kaur, G. Sikka and L. K. Awasthi. “Sentiment Analysis Approach Based on N-gram and KNN Classifier”. *ICSCCC 2018 1st International Conference Security Cyber Computer communication*, pp. 13-16, 2019.

[10] X. Zhang and X. Zheng. “Comparison of Text Sentiment Analysis Based on Machine Learning”. *Proceeding 15th Internatioanl Symposium Parallel Distributed Computing ISPDC 2016*, pp. 230-233, 2017.

[11] R. Joshi and R. Tekchandani. “Comparative Analysis of Twitter Data Using Supervised Classifiers”. *Proceeding International Conference Invention Computer Technology ICICT 2016*, vol. 2016, 2016.

[12] M. Luo and L. Luo. “Feature Selection for Text Classification Using OR+SVM-RFE”. *2010 Chinese Control Decision Conference CCDC 2010*, pp. 1648-1652, 2010.

[13] R. Maipradit, H. Hata and K. Matsumoto. “Sentiment classification using N-gram IDF and automated machine learning”. *IEEE Software*, vol. 7459, pp. 10-13, 2019.

[14] S. Rai, S. M. Shetty and P. Rai. “Sentiment Analysis of Movie Reviews using Machine Learning Classifiers”. *International Journal of Computer Applications*, vol. 182, no. 50, pp. 25-28, 2019.

[15] S. Naz, A. Sharan and N. Malik. “Sentiment Classification on Twitter Data Using Support Vector Machine”. *Proceeding 2018 IEEE/WIC/ACM International Confernce Web Intell. WI 2018*, pp. 676-679, 2019.

[16] R. Wagh and P. Punde. “Survey on Sentiment Analysis Using Twitter Dataset”. *Proceeding 2nd International Conference electronic communications Aerospace Technology ICECA 2018*, No. Iceca, pp. 208-211, 2018.

[17] N. Iqbal, A. M. Chowdhury and T. Ahsan. “Enhancing the Performance of Sentiment Analysis by Using Different Feature Combinations”. *International Conference Compututing Communication IC4ME2 2018*, pp. 1-4, 2018.

[18] A. Rane and A. Kumar. “Sentiment Classification System of Twitter Data for US Airline Service Analysis”. *Proceeding International Computing Software APPL Conference*, vol. 1, pp. 769-773, 2018.

[19] A. Jovi, K. Brki and N. Bogunovi. “A Review of Feature Selection Methods with Applications”. *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*, pp. 25-29, 2015.

[20] S. Rana and A. Singh. “Comparative Analysis of Sentiment Orientation Using SVM and Naive Bayes Techniques”. *Proceeding 2016 2nd Interenational Confernce Next General Computer Technologies*, *2016*, pp. 106-111, 2017.

[21] K. L. S. Kumar, J. Desai and J. Majumdar. “Opinion Mining and Sentiment Analysis on Online Customer Review”. *2016 IEEE Interenatioanl Conference Computing Intelligence computing Research ICCIC 2016*, 2017.

[22] F. Iqbal, J. Maqbool, B. C. M. Fung, R. Batool, A. M. Khaytak, S. Aleem and P. C. K. Hung. “A hybrid framework for sentiment analysis using genetic algorithm based feature reduction”. *IEEE Access*, vol. 7, pp. 14637-14652, 2019.

[23] A. Zarshenas and K. Suzuki. “Binary coordinate ascent: An efficient optimization technique for feature subset selection for machine learning”. *Knowledge-Based Systems*, vol. 110, pp. 191-201, 2016.