Analysis and Evaluation of Two Feature Selection Algorithms in Improving the Performance of the Sentiment Analysis Model of Arabic Tweets

Maria Yousef[1]
Department of Computer Science
Al-al Bayt University
Mafraq, Jordan

Abdulla ALali[2]
Department of Computer Science
Isra University
Amman, Jordan

***Abstract*—Recently, Sentiment analysis from Twitter is one of the most interesting research disciplines; it combined data mining technologies with natural language processing techniques. The sentiment analysis system aims to evaluate the texts that are posted on social platforms to express positive, negative, or neutral feelings of people regarding a certain domain. The high dimensionality of the feature vector is considered to be one of the most popular problems of Arabic sentiment analysis. The main contribution of this paper is to solve the dimensionality problem by presenting a comparative study between two feature selection algorithms, namely, Information Gain (IG), and Chi-Square to choose the best one which may lead to improve the classification accuracy. In this paper, the Arabic Jordanian sentiment analysis model is proposed through four steps. First, a preprocessing step has been applied to the database and includes (Remove Non-Arabic Symbols, Tokenizing, Arabic Stop Word Removal, and Stemming). In the second step, the TF-IDF algorithm is used as a feature extraction method to represent the text into feature vectors. Then, we utilized IG and Chi-Square as feature selection steps to obtain the best subset of features and decrease the total number of features. Finally, different algorithms have been used in the classification step such as (SVM, DT, and KNN) to classify the views people have shared on Twitter, into two classes (positive, and negative). Several experiments were performed on Jordanian dialectical tweets using the AJGT database. The experimental results show the following: 1) The information acquisition algorithm outperformed the Chi-Square Algorithm in the feature selection step, as it was able to reduce the number of features from 1170 to 713 and increase the accuracy of the classifiers by 10%, 2) SVM classifier shows the greatest classification performance among all the classifiers tested which gives the highest accuracy of 85% with IG algorithm.**

***Keywords*—Sentiment analysis; Information Gain (IG); Chi-Square; AJGT database**

I. Introduction

In the past few years, people have been able to interact and share comments across a variety of social media platforms. People all around the world use social media websites to share their ideas and feelings, exchange information, and share news in an easy way. Twitter is a social media website that was launched in 2006 and quickly grew in popularity throughout the world, with over 313 million monthly active users and more than 40 languages supported [1].

The Twitter platform allows users to upload short posts known as tweets to share their feelings, opinions, and thoughts on a wide range of topics. The number of Twitter users in the Arab world is growing rapidly, and a huge amount of Arabic tweets are generated daily. Because of the Extensive use of the Twitter website, media organizations and companies are interested in learning what people think and feel about their commodities, services, and products through their tweets by using sentiment analysis systems.

Sentiment analysis (SA), often defined as opinion mining, is a Natural Language Processing (NLP) process that contains automatically detecting an attitude from a text that is connected to a specific issue. Oueslati et al. in [2] described sentiment analysis as a "method used to determine favorable and unfavorable opinions toward specific products and services using large numbers of textual data sources". The main purpose of SA is to discover people's feelings, opinions, beliefs, emotions, and attitudes expressed in natural language whether positive, negative, or neutral about a person, an organization, a product, a location, or an event, and how this changes over time.

Machine learning and lexicon-based techniques are the two main strategies used in the literature to create SA systems. The lexicon-based technique uses a predetermined vocabulary with weighted terms and their sentiment inclination to estimate the sentiment tendency of text data. In this strategy, the process of classifying the text is done using its set dictionary as described in [3]. While the Machine Learning (ML) strategy uses well-known ML algorithms to solve SA as a classification problem [3]. In several literatures, we discovered that the two methodologies were blended to create a hybrid approach.

Sentiment analysis research in English has made significant progress, but it is still restricted in the Arabic language. The Arabic language has a complicated structure, according to its ambiguity and extensive morphological properties. This, along with having wide range of dialects and resource scarcity, makes progress in Arabic SA research difficult. Because of the significance of the Arabic language, this research focusing on assessing the sentiment of Arabic tweets shared by a huge number of users on Twitter by using machine learning approach.

One of the difficult issues that machine learning methods encounter in sentiment analysis and more generally in text classification is a large number of the features in the using dataset. Where a feature can have a favorable or negative impact on classification performance based on its relevance and redundancy with regard to the class labels. Therefore, feature selection approaches are necessary to enhance the accuracy and efficiency of classification algorithms by picking the most relevant and discriminating feature vectors. Feature selection methods determine the most useful features by measuring the quality of a features subset with which the best performance can be obtained.

This work examined the impact of two feature selection techniques (Chi-Square, and Information Gain (IG)) on the performance of three classifiers (DT, KNN, and SVM) that trained on Jordanian Arabic Tweets to create a sentiment analysis model. The Arabic Jordanian General Tweets (AJGT) Dataset is used in this study to classify Jordanian tweets in two classes (negative, and positive). The AJGT dataset involves 1,800 tweets that are categorized as (900 positives, 900 negatives) represent people's opinions on different topics.

The study is conducted as follows. The associated study is discussed in Section II. The technique of the proposed model is explained in Section III. Section IV discusses the experiments and their outcomes. Section V concludes with a summary of the study's findings.

II. Related Work

In the sentiment analysis literature, a variety of machine learning techniques have been utilized [4-6]. In the classification task, the feature selection approach is used to increase performance by reducing dimensionality. There is a substantial amount of published work in the English language that uses several feature selection techniques to increase the performance of the sentiment analysis model [7]. On the converse, the influence of feature selection on dialectal Arabic sentiment analysis has received less attention.

Duwairi and El-Orfaili [8] examined the effect of feature correlation, stemming, and n-gram models on sentiment analysis of Arabic text. For text representation, several N-gram models of words and characters were utilized. In this research, two datasets were utilized to conduct experiments. The first dataset is named the Politics dataset it contains 300 reviews of which164 are categorized as positive reviews and 136 are categorized as negative reviews. The authors gathered these reviews from Aljazeera webpage. While the second dataset is known as the Movie dataset and is open to the public. Three classification algorithms were employed to classify the reviews (SVM, Naïve Bayes, and K-NN). From the result, SVM and Naïve Bayes classifiers achieve better performance. Substantial improvement has been achieved by using the top 1200 correlated features and word N-gram with Naïve Bayes algorithm to produce the most increased accuracy with 97.2%.

Baker et al., [9] introduced the first research of epidemic illnesses based on tweets in the Arabic language by developing a novel sentiment analysis system that aims to identify Influenza from Arab countries' tweets. The authors of this work obtained, labeled, filtered, and analyzed Arabic-language influenza-related tweets. The following algorithms were used to evaluate the system's quality and performance: DT, K-NN, NB, and SVM. From the tests, the best accuracy value of detecting influenza was 83.20%, which was achieved in the NB algorithm.

Altaher Taha [10] proposed a novel technique for sentiment analysis of Arabic tweets using features weighting and deep learning. In this work, stop word removal, tokenization, and stemming are utilized as preprocessing methods, followed by the use of two feature weighting methods (Chi-Square and Information Gain) to give high weights to the most important features of Arabic tweets. Second, the deep learning approach is used to efficiently and correctly classify Arabic tweets as either positive or negative. The suggested methodology was compared to the performance of different classification algorithms such as Neural Networks (NN), Support Vector Machine (SVM), and Decision Tree (DT) using the same data gathered from Arabic tweets. The suggested method exceeds the others, achieving the maximum precision and accuracy of 93.7% and 90%, respectively.

El Rahman et al., [11] developed a model that can classify tweets as negative, positive, or neutral using different techniques to improve the classification accuracy. The authors of this study used Twitter API to extract Tweets on two topics: KFC and McDonald's, in order to determine which restaurant is more popular. This presented methodology combined the use of both unsupervised and supervised machine learning techniques. Firstly lexicon-based algorithms have been used to create previously unlabeled data. After that data was input into several supervised models for training purposes, including (SVM, NB, DT, and RF). Different testing measurements, such as f-score and cross-validation, were used to test the outcomes of the suggested models. As a consequence, both positive and negative reviews show that McDonald's is more popular than KFC.

Ghallab et al., [12] presented a deep survey of the current literature related to Arabic sentiment analysis (ASA). The major aims of this review are to encourage research and to identify new areas for future study in ASA, Analyze and compare the methods used in each study and its results, also to make it easier for other researchers to find similar works. After filtering, this review focused to analyzed 108 published studies from 22 conferences and 11 journals proceedings. The conclusions of the review indicate that there is little study on creating standard datasets and using promising classifiers. Furthermore, the review shows limited research interested in designing a novel feature representation that appropriates for the Arabic language characteristics. Finally, the review highlights future research in the development of recommender systems in several areas with the goal of an improved framework for ASA.

III. Proposed Methodolgy

In this part, we will clarify the parts of the suggested sentiment analysis model for Twitter in more detail. As seen in Fig. 1, the suggested model is divided into two parts: Training and Classification. The goal of the training part is to create a classification model that can discriminate between negative and positive tweets based on input labeled tweet collections.

Then during the classification part, the previously trained classification model will apply to assign a negative or positive label to the new unlabeled tweets. This proposed model includes four steps: Preprocessing, Feature Extraction, Feature Selection, and the Classification Model for Sentiment Analysis. The overall steps of the proposed model are shown in Fig. 1.

A. *Arabic Jordanian General Tweet (AJGT) Dataset*

Arabic Jordanian General Tweets is a new dataset Created 6in 2017 by Alomari et al., [13] for the purpose of sentiment analyses, and it contains 1,800 tweets that are categorized as (900 positives, 900 negatives) represent people's opinions on different topics. The AJGT has been written in Jordanian dialect and MSA, and it's publicly available as a Github project.

B. *Preprocessing*

The major goal of this step is to process the input tweet text by using natural language processing approach to prepare it for the next step of appropriately extracting the features. In this study, there are different preprocessing steps adapted in the data like filtering, tokenization, and stemming. These steps are clarified in more detail as below:

1) Remove non-arabic symbols: Most Arabic tweets on Twitter involve noise, like elongations, diacritical marks, mixed language, and special characters. Therefore, one of the most crucial stages in preparing Arabic text for Twitter is to clean up the tweets by eliminating this kind of noise [14]. AJGT dataset has some special characters like (!, ., ?, %). To eliminate the noise from the text, non-Arabic characters are removed by using the cleaning method. As illustrated in Fig. 2, The cleaning procedure checks if each character in the text of the tweet belongs to the Arabic alphabet by reading it character by character. If a character belongs to the Arabic alphabet, it is chosen; otherwise, it is replaced with white space by.

2) Tokenizing: Tokenization is only a segmentation technique of the sentences. The goal of this process is to divide sentences into smaller pieces called "tokens", whether that is words or phrases [15]. There are several ways for dividing the phrases in RapidMiner, including using the n-gram approach. The n-gram is a type of graph that is based on the number of letters in a word (n-number-of-tokens) and is utilized to keep the sentences at its meaning [16]. When n=2, it's referred to as diagrams, and when n=3, it's referred to as trigrams. For n=3, a sequence of three consecutive words (tokens) is created for every Arabic tweet in the dataset. In this study, we used a unigram option which means dividing the sentence into single words, each part representing a word.

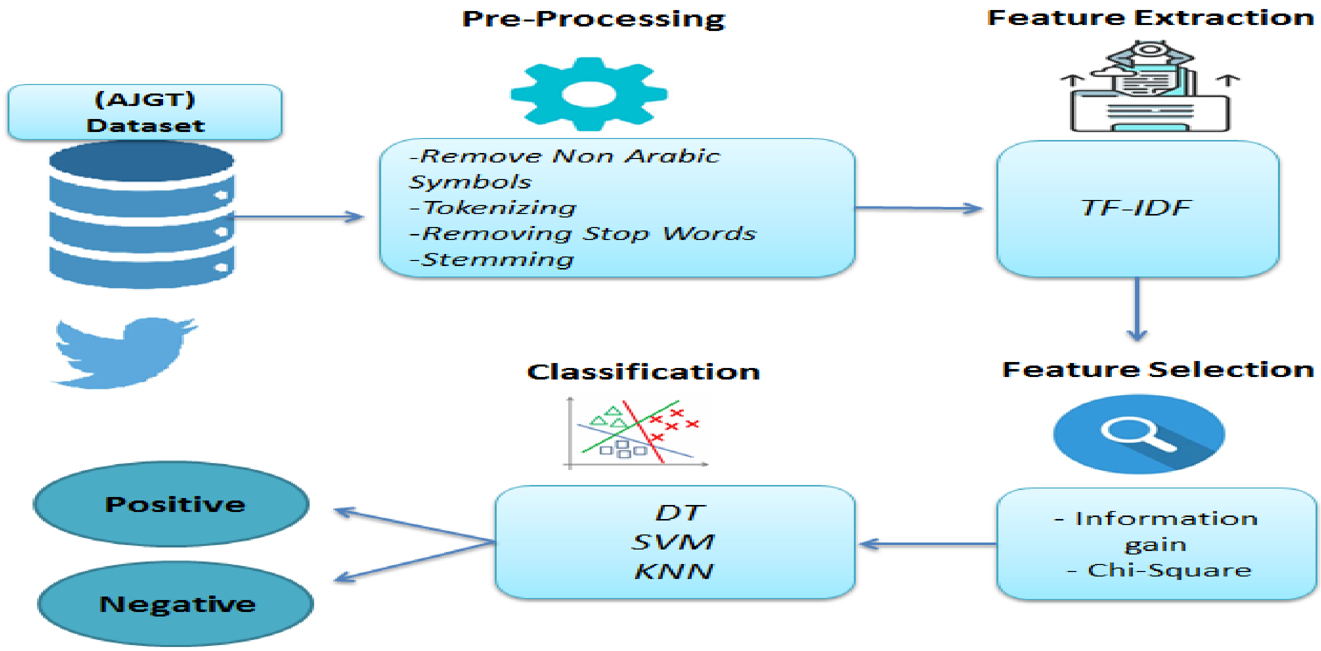

Fig. 1. The Proposed Model Framework.

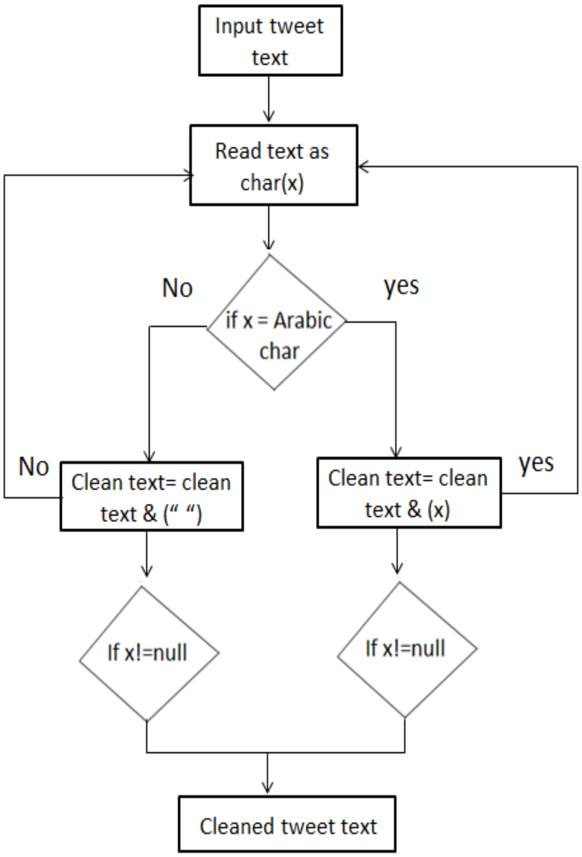

Fig. 2. Cleaned Method.

3) Arabic stop word removal: This step is one of the most popular used preprocessing techniques in different NLP applications. This task is employed to eliminate unnecessary and meaningless words [17]. Stop words refer to phrases that appear frequently in Arabic writings, such as (i.e. بعد, من , الى , لهم , كما, قد). In this step, each token is compared to the stop word list that already exists. If it's on the list, it will be removed.

4) Stemming: Replacing a word with its stem means returning it to its original form. this strategy has a significant influence on reducing storage requirements and eliminating redundant terms. Because many Arabic words have the same stem, converting words with the same stem will eliminate redundant terms, for example, the words تعمل, يعمل, يعملون have the same stem which is عمل so all of these words will be replaced with the word عمل. There are two ways to stem a word: reducing it to its three-letter root and light stemming, which eliminates frequent suffixes and prefixes but does not reduce it to its root. In this study, we used root stemming (Porter's Stemmer algorithm).

C. Feature Extraction

After the preprocessing step, the text should be converted into a numeric representation that the ML classifier can interpret. A vector model, often called a feature model, is a model that is represented by a matrix of word weights. Weighting Word is a method of assigning a score to the frequency with which a phrase occurs in a text document [18]. TF-IDF (Term Frequency-Inverse Document Frequency) is one of the popular methods for weighting words. The idea of TF-IDF is that it calculates the frequency of each token in a tweet [19]. Because each tweet is varied in length, a term may occur more frequently in a long tweet than in a short tweet. As a result, term frequency is frequently split by the document's length (the total words in the document). The TF-IDF value shows how important a token is to a document in the tweets.

D. Feature Selection

The classification's performance is heavily influenced by the feature vector. Notably, that features if, by excluding or including them, the performance would improve or decrease. The relevant features are critical to the training phase because it has an informative aspect that would enhance the classification. On the other hand, the irrelevant ones are less informative, thus including them may have an adverse effect on performance. Determine which of the features are irrelevant or relevant is the objective of this step [20]. In this section, two feature selection strategies were used to extract the most relevant and meaningful features which are Information Gain and Chi-Square (IG).

1) Information Gain (IG): IG is a useful feature selection method that assesses how informative a feature is about the class. IG denotes the reduction of uncertainty in selecting a category by knowing when the value of the feature. IG is a ranking score algorithm which can be computed for a term as shown in (1) [21]:

$$IG = \sum_{i=0}^{m} P(c_i)\log\big(P(c_i)\big) + P(t)\sum_{i=0}^{m} P(c_i\,|t)\log\big(P(c_i|t)\big) + P(\bar{t})\sum_{i=0}^{m} P(c_i\,|\bar{t})\log P(c_i|\bar{t}) \quad (1)$$

c_i Represent the ith category, the probability of the i th category is $P(c_i)$. The probability that the word t will occur or not appear in the documents are represented by $P(t)$ and $P(\bar{t})$. $P(c_i|t)$ Represents the probability of the i th category if the term t appeared, while $P(c_i|\bar{t})$ represents the likelihood of the i th category if the word t did not exist.

2) Chi-Square: The Chi-Square of independence is a statistical hypothesis test that calculates the difference between predicted and observed frequencies for two events [22]. The purpose of this test is to see if a difference between observed and predicted data is due to chance or if there is a link between the events. In feature selection, the two events to know the relationship between are the occurrence of the term and the occurrence of the class. Few studies have looked into the impact of utilizing the Chi-Square feature selection approach in Arabic sentiment analysis, such as [23]. In this study, The value for each term concerning the value of the class is calculated as shown in (2).

$$X^2(t,c) = \frac{D \times (PN - MQ)^2}{(P+M)\times(Q+N)\times(P+Q)\times(M+N)} \quad (2)$$

Where D represented the total number of tweets, P denotes the count of tweets in class c that include the term t. Q is the number of Arabic tweets containing t occurring without c.

While the number of tweets in class c that occur without t is M. While, N is the number of tweets from other classes that do not include term t.

E. Classification Algorithms

1) Decision Tree (DT): It is a hierarchical tree that divides data into groups depending on attribute value conditions. according to another definition, it is the technique of recursively partitioning training data into smaller pieces based on a series of tests that are displayed at each branch of the tree [24]. Every node in the tree represented a feature training test, and each branch dropping from the node matched the feature value. For example, to categorize an instance, start with the parent node, verify its feature, and then move down the tree branch to the value of the feature for the specific instance which knows the leaf node (or terminal node) and represents the class label [25]. In the text situation, decision tree nodes are usually represented by words in tweets. several techniques are employed in the decision tree to improve classification accuracy. We used a DT with a maximum depth of 10 in this work. This algorithm is based on a gain ratio-based selection criterion for deciding which characteristics to separate.

2) K-Nearest neighbor (K-NN): It's a ranking approach that uses a method based on the number of nearest neighbors and the distance between both the training data and the target data. [26]. Therefore, to make a prediction task, kNN utilizes similarity measures to make a comparison between a particular test entry and the training data set An n-featured record is displayed for each data entity. Each data entity displays an n-featured record. To guess a class label for an unidentified record, The kNN algorithm selects k recodes from the training data set that are the nearest to the unknown records. One of the most widely used for calculating this distance is called Euclidean measurement. In this study, the value of k is set at 5 and the used distance measure is Mixed Euclidean Distance.

3) Support vector machine (SVM): The SVM method was chosen because it is one of the most well-known classifiers in latest years. In the sentiment analysis research area, the SVM classifier outperformed other classifiers in several studies such as in [27,28]. The goal of SVM is to discover the Highest Marginal Hyperplane, which is the maximum margin between the hyperplane and the points on the hyperplane border. In general, SVM provides the benefit of overfitting protection and the ability to handle huge feature spaces. In this study, a radial basis function kernel SVM (RBF SVM) was used.

IV. Performance Evaluation Metric

To analyze the efficiency of the suggested strategy, the following standard assessment measures were used:

1) Accuracy: It's an important metric to assess the achievement of the classification task. It is represented as the percentage of correctly classified samples to the total samples. thus, it can be calculated mathematically by using the formula shown in (3) [29]:

$$Accuracy = \frac{TP+FP}{TP+FP+TN+FN} \tag{3}$$

2) Precision: It determines how strict the classification output is. It is described as the percentage of samples accurately classified as positive compared to the total number of samples classified as positive. thus, it can be calculated mathematically by using the formula shown in (4) [30]:

$$Precision = \frac{TP}{TP+FP} \tag{4}$$

3) Recall: this metric is used to determine the quality of the classifier's output. It can be calculated mathematically by using the formula shown in (5) [30]:

$$Recall = \frac{TP}{TP+FN} \tag{5}$$

Where:

(TP) is the number of Arabic tweets properly classified as belonging to the correct class by the classifier.

(TN) is the number of Arabic tweets successfully classified as not belonging to the right class by the classifier.

(FP) is the number of Arabic tweets wrongly classified as belonging to the correct class by the classifier.

(FN) is the number of Arabic tweets wrongly classified as not belonging to the correct class by the classifier.

V. Result and Discussion

In this work, three experiments were undertaken to examine the Impact of feature selection techniques and some classification algorithms on the dialectal Arabic sentiment analysis. In the first experiment, we applied classification algorithms directly to the AJGT database, while we applied two methods of feature selection (information acquisition and Chi-Square) in the second and third experiments to compare them and choose the best method that enhances the accuracy of sentiment classification. In the next part, the experiments are explained in more detail.

A. First Experiment

The goal of this experiment is to It is to find the most accurate classification algorithm based on all the features in the database. In this experiment, we compare the accuracy of the most used classifiers in sentiment analysis systems without employing feature selection methods. First, we prepared the data set using the preprocess approaches mentioned in Section B. After that, numerous classifiers were used, including SVM, DT, and k-NN for sentiment classification. To avoid dataset overfitting and improve model performance, we verified the models using the cross-validation approach, which randomly separates the dataset into a training dataset and a testing dataset depending on the k-fold value [31].

The first experiment employed a different number of K-fold that was chosen as (5, 10, 15, and 20). Whereat each time the tweets are tested with various folds and various classifiers. Fig. 3 shows the average accuracy for the various folds (5, 10, 15, and 20) using the DT, KNN, and SVM classifiers. The results are as follows: the average accuracy of the DT is 62.8%, 62.4%, 62.8%, and 62.4%, respectively. The average accuracy of the KNN is 62.6%, 62.6%, 62.7%, and 62.8%, respectively. Also, the average accuracy of the SVM is 62%, 62.6%, 62.7%, and 63.2% respectively. After comparing all of the accuracy values, we observed that the SVM RBF method achieved the greatest accuracy value of 63.2% at 20-folds.

B. *Second Experiment*

The purpose of this experiment is to study the effect of employing the Information Gain approach on the suggested system in two aspects: accuracy improvement and dimension reduction. In the first portion of the experiment, the accuracy of the classifiers (DT, KNN, and SVM) is computed after applying the Information Gain approach as a feature selection algorithm to obtain the best subset of features from all features. In the second portion of the experiment, we are interested in assessing the reduction ratio achieved by using the IG strategy, because the IG technique is primarily utilized to choose features that better match the given classes. Based on that, the IG algorithm was able to reduce the total number of features from 1170 to 713 and the results were presented in Table II. Table II demonstrates the performance of the proposed approach for Arabic sentiment analysis based on the IG feature selection algorithm at 20-folds (determined based on previous experience as in 4.1).

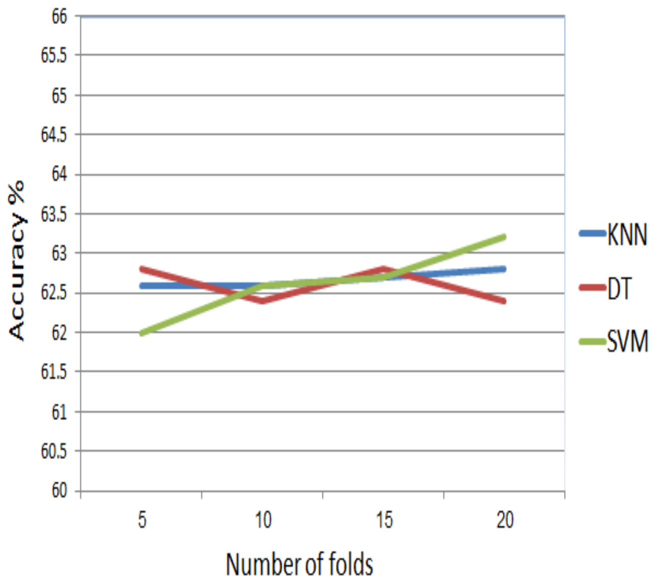

Fig. 3. Accuracy Comparison of Different Classifiers before using Feature Selection.

TABLE I. EXAMPLE OF THE CLEANING ALGORITHM OUTCOMES

Tweet text before cleaning	Tweet text after cleaning
"ما شاء الله ... ونعم التعليم في الاردن .... ابدعوا الشباب !!!"	"ما شاء الله ونعم التعليم في الاردن ابدعوا الشباب "
" الجمال : ليس فقط شيئا نراه بل هو شيء نكتشفه ، روح جميله ، وفكر جميل ، اخلاق جميله ، وادب جميل"	" الجمال ليس فقط شيئا نراه بل هو شيء نكتشفه روح جميله وفكر جميل اخلاق جميله وادب جميل"
"وانا 100% بس شكلها هاي مو مدرسه"	"وانا بس شكلها هاي مو مدرسه"

TABLE II. COMPARISON OF ACCURACY AFTER USING IG WITH SEVERAL CLASSIFIERS

Algorithms	Accuracy %	Precision %	Recall %
DT	75	74	75
KNN	72	75	70
SVM	85	85	84

As demonstrated in Table III, using the IG technique increased the accuracy of the proposed model by about 15% on average in all employed classifiers. Furthermore, utilizing the IG approach decreases the AJGT dataset's feature vector length by around 61%. This allows the classifiers to distinguish efficiently between positive and negative classes with lower computational requirements. Furthermore, applying the IG approach in the feature selection step reduces the feature vector length for the AJGT dataset by around 61%. This enables the classifiers to discriminate between positive and negative classes more effectively while using fewer computational requirements. Based on the findings in Table I, we can observe that using the IG technique not only reduces the dimension of the feature vector but also significantly improves the performance of all classifiers.

C. *Third Experiment*

In this experiment, we aim to evaluate the performance of the proposed model after using the Chi-Square approach as a feature selection method. After applying the same evaluation process described in the second experiment on the AJGT dataset, the Chi-Square Algorithm was able to reduce the total number of features from 1170 to 750 features and the results were presented in Table III. Table III shows the performance of the proposed approach for Arabic sentiment analysis based on the Chi-Square feature selection algorithm at 20-folds (determined based on previous experience as in Section 4.1).

After comparing the results of previous experiments, we conclude the following:

- The SVM classifier shows the greatest classification performance among all the classifiers tested, but its performance decreases as the number of features increases. In the second and third tests, the D-tree classifier's performance appears to remain stable, despite the low number of features. In all experiments, K-NN had the worst results.
- The proposed approach based on the Information Gain Algorithm in the feature selection step achieved a 2% improvement overall classifiers compared to the Chi-Square Algorithm.

TABLE III. COMPARISON OF ACCURACY AFTER USING CHI-SQUARE WITH SEVERAL CLASSIFIERS

Algorithms	Accuracy %	Precision %	Recall %
DT	73	73	71
KNN	70	72	70
SVM	83	84	82

- The efficiency of this proposed approach is due to two main factors: First, we utilized a preprocessing step to improve the quality of the data we used, which leads to an increase in the quality of the results. Second, using feature selection algorithms to select the best subset of features reduces the total number of features and improves the classification model's accuracy. An IG algorithm outperformed the Chi-Square Algorithm as it was able to reduce the number of features from 1170 to 731, whereas the Chi-Square Algorithm reduced it to 750.

VI. Conclusion

This study examined the impact of two feature selection approaches (IG, and chi-squire) on the performance of the (DT, KNN, and SVM) classifiers for dialectal Arabic sentiment analysis. The experiments are implemented in the RapidMiner data mining tool by using an AJGT dataset and a 20-fold cross-validation technique. The dataset consists of 1800 Arabic Jordanian tweets labeled by two different classes (negative and positive). The experimental results showed the Information Gain Algorithm outperformed the Chi-Square Algorithm in the feature selection step, as it was able to reduce the number of features by 61% and increase the accuracy of the classifiers by 10%. Moreover, the SVM classifier shows the greatest classification performance rather than DT, and KNN among all the experiments which give the highest accuracy of 85% with the IG algorithm.

References

[1] T. Singh, & M. Kumari, "Role of text pre-processing in twitter sentiment analysis. Procedia Computer Science", vol. 89, pp. 549-554, 2016.

[2] O., E. Cambria,, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language". Future Generation Computer Systems, vol. 112, pp. 408-430, 2020.

[3] M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad, "Machine learning techniques for sentiment analysis: A review," Int. J. Multidiscip. Sci. Eng, vol. 8, no.3, pp.27, 2017.

[4] A. Elnagar, S. Yagi, A. B Nassif,, I. Shahin, & S. A. Salloum, "Sentiment analysis in dialectal Arabic: a systematic review". In International Conference on Advanced Machine Learning Technologies and Applications, pp. 407-417, Springer, 2021.

[5] R. S. Jagdale, V. S. Shirsat, and V. S. Deshmukh, "Sentiment analysis on product reviews using machine learning techniques," In Cognitive Informatics and Soft Computing, pp. 639-647, 2019.

[6] R. B. Shamantha, S. M. Shetty, and P. Rai, "Sentiment analysis using machine learning classifiers: evaluation of performance". In 2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS), pp. 21-25,IEEE, February.2019 .

[7] Koncz, and J. Paralic,"An approach to feature selection for sentiment analysis," In 2011 15th IEEE International Conference on Intelligent Engineering Systems, pp. 357-362, IEEE, June.2011.

[8] R. Duwairi, and M. El-Orfali, "A study of the effects of preprocessing strategies on sentiment analysis for Arabic text,". Journal of Information Science, vol.40, no.4, pp. 501-513, 2014.

[9] Q. B. Baker, F. Shatnawi, S. Rawashdeh, M. Al-Smadi, and Y. Jararweh, "Detecting Epidemic Diseases Using Sentiment Analysis of Arabic Tweets," J. Univers. Comput. Sci., vol.26, no.1, pp.50-70, 2020.

[10] A. Altaher, "Hybrid approach for sentiment analysis of Arabic tweets based on deep learning model and features weighting," International Journal of Advanced and Applied Sciences, vol.4, no.8, pp.43-49, 2017. https://doi.org/10.21833/ijaas.2017.08.007.

[11] S. A. El Rahman, F. A. AlOtaibi, and W. A. AlShehri, "Sentiment analysis of twitter data," In 2019 International Conference on Computer and Information Sciences (ICCIS), pp. 1-4, IEEE, April. 2019.

[12] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: A systematic literature review," Applied Computational Intelligence and Soft Computing, 2020.

[13] K. M. Alomari, H. M. ElSherif, and K. Shaalan, "Arabic tweets sentimental analysis using machine learning," In International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, pp. 602-610, Springer, Cham, June. 2017.

[14] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," In 2016 7th international conference on information, intelligence, systems & applications (IISA), pp. 1-5, IEEE, July. 2016.

[15] S. Vijayarani, and R. Janani, (2016). "Text mining: open source tokenization tools-an analysis," Advanced Computational Intelligence: An International Journal (ACII), vol. 3, no.1, pp.37-47,2016.

[16] F. Aisopos, G. Papadakis, and T. Varvarigou, "Sentiment analysis of social media content using n-gram graphs," In Proceedings of the 3rd ACM SIGMM international workshop on Social media, pp. 9-14, November.2011.

[17] A. W. Pradana, and M. Hayaty, "The effect of stemming and removal of stopwords on the accuracy of sentment analysis on indonesian-language texts," Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control, pp.375-380, 2019.

[18] J. Kumar, J. K. Rout, and S. K. Jena, "Sentiment analysis using weight model based on SentiWordNet 3.0.," In Recent Findings in Intelligent Computing Techniques, pp. 131-139, Springer, Singapore, 2018.

[19] A. Mee, E. Homapour, F. Chiclana, and O. Engel, "Sentiment analysis using TF–IDF weighting of UK MPs' tweets on Brexit," Knowledge-Based Systems, vol.228, pp.107238,2021

[20] B. Agarwal, and N. Mittal, "Optimal feature selection for sentiment analysis," In International conference on intelligent text processing and computational linguistics, pp. 13-24, Springer, Berlin, Heidelberg, March.2013.

[21] B. Azhagusundari, and A. S. Thanamani, "Feature selection based on information gain," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 2, no.2, pp.18-21, 2013.

[22] A. M. Bidgoli, and M. N. Parsa," A hybrid feature selection by resampling, chi squared and consistency evaluation techniques," World Academy of Science, Engineering and Technology, vol.68, pp. 276-285, 2012.

[23] A. Sharma, and S. Dey, "A comparative study of feature selection and machine learning techniques for sentiment analysis," In Proceedings of the 2012 ACM research in applied computation symposium, pp. 1-7, October.2012.

[24] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," Emerging artificial intelligence applications in computer engineering, vol.160, no.1, pp.3-24, 2007.

[25] M. Du, S. M. Wang, and G. Gong, "Research on decision tree algorithm based on information entropy," In Advanced Materials Research, vol. 267, pp. 732-737, Trans Tech Publications Ltd, 2011.

[26] H. A. Abu Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. Eyal Salman, and V. S. Prasath, (2019). "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," Big data, vol.7, no.4, pp.221-248, 2019.

[27] M. Ahmad, S. Aftab, M. S. Bashir, N. Hameed, I. Ali, and Z. Nawaz, "SVM optimization for sentiment analysis," Int. J. Adv. Comput. Sci. Appl, vol. 9, no.4, pp.393-398, 2018.

[28] R. M. Duwairi, and I. Qarqaz, "Arabic sentiment analysis using supervised classification," In 2014 International Conference on Future Internet of Things and Cloud, pp. 579-583, IEEE, August. 2014.

[29] M. Shahrokh Esfahani, and E. R. Dougherty, (2014). "Effect of separate sampling on classification accuracy," Bioinformatics, vol.30, no.2, pp. 242-250, 2014.

[30] D. M. Powers, (2020). "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," arXiv preprint arXiv:2010.16061, 2020.

[31] M. W. Browne, "Cross-validation methods," Journal of mathematical psychology, vol.44, no.1, pp. 108-132, 2000.