Arabic Sentiment Analysis for Multi-dialect Text using Machine Learning Techniques

Aya H. Hussein[1], Ibrahim F. Moawad[2], Rasha M. Badry[3]
Department of Information Systems, Faculty of Computers and Information, Fayoum University, Fayoum 63514[1, 3]
Department of Artificial Intelligence Engineering, Faculty of Computer Science and Engineering, Galala University, New Galala City, Suez 43511[2]

***Abstract*—Social media networks facilitated the availability and accessibility of a wide range of information and data. It allows the users to share and express their opinions. In addition, it presents the appraisals of the top news and the evaluation of movies, products, and services. This headway has been controlled by a well-known field called Sentiment Analysis (SA). Compared to the research studies conducted in English Sentiment Analysis (ESA), little effort is exerted in Arabic Sentiment Analysis (ASA). The Arabic language is a morphologically rich language that poses significant challenges to Natural Language Processing (NLP) systems. The purpose of the paper is to enrich the Arabic Sentiment Analysis via proposing a sentiment analysis model for analyzing an Arabic multi-dialect text using machine learning algorithms. The proposed model is applied to two datasets: ASTD Egyptian-Dialect tweets and RES Multi-Dialect restaurant reviews. Different evaluation measures were used to evaluate the proposed model to identify the best performing classifiers. The findings of this research revealed that the developed model outperformed the other two research works in terms of accuracy, precision, and recall. In addition, the Bernoulli Naive Bayes (B-NB) classifier achieved the best results with 82% for the ASTD Egyptian-Dialect tweets dataset, while the SVM classifier scored the best accuracy result for the RES Multi-Dialect reviews dataset with 87.7%.**

Keywords*—Arabic sentiment analysis (ASA); arabic tweets; sentiment analysis (SA); natural language processing (NLP); machine learning (ML)

I. Introduction

Most Internet users tend to shift from traditional communication tools (e.g., traditional blogs or mailing lists) to Micro-blogging services. It has a free format of messages and is easy to use. Micro-blogging today has become a top-rated communication tool between Internet users, reflecting users' opinions [1]. These opinions represent any kind of information (political, sport, technology, etc.) that comes from different sources. Sentiment analysis (SA) aims to extract or predict the polarity of users' opinions in a specific area which is a challenging task [2], [3]. SA is considered an important area in Natural Language Processing and Artificial Intelligence to identify emotions and trends about a specific topic. Two main approaches are adopted for SA: machine-learning and Lexicon-based approaches [3]–[5]. The machine learning approach uses a supervised learning approach where a classifier is trained on a human-annotated dataset [6], [7]. Many sentiment analysis researches have been done, especially in the English language. However, there are a huge number of Arabic users on social media posting and sharing their opinions in the Arabic language, expressing feelings and opinions, which can affect many businesses and domains.

Simultaneously, there are many Arabic dialectal variants such as classical Arabic, the language of the Quran, and modern standard Arabic (MSA). The standardized official language is written in the news and taught in schools. In addition, dialectal Arabic (DA) is used in daily life and communications. The Arabic dialects are divided into (1) Egyptian-Dialect Arabic for Egypt and Sudan (EA), (2) Levantine Arabic for Lebanon, Syria, Palestine, and Jordan (LA), (3) Gulf Arabic for Gulf area (GA), and (4) Maghrebi Arabic for Morocco, Algeria, Tunisia, Mauritania, and Libya (MA). Furthermore, Arabic used in social media is usually a mixture of MSA and one or more Arabic dialects [8]–[10].

Arabic sentiment analysis has challenging issues based on two main vectors: Arabic-specific and general linguistic problems. Arabic morphological complexity, limited resources cause the Arabic-specific, and dialects, while the general linguistic issues include polarity fuzziness and strength, implicit sentiment, sarcasm, spam, reviews quality, and domain dependence [9], [11], [12].

The importance of this research is that a sentiment analysis model for analyzing and extracting Arabic text multi-dialect opinions is proposed based on machine learning algorithms and gets high accuracy results. The proposed model experimented using two different datasets (Egyptian and Multi-Dialects datasets). First, the Arabic text is preprocessed to enhance the classifier's performance, such as de-nosing, removing stop words, and applying the lemmatization technique. Then, feature weight and feature selection methods are used. Finally, several machine learning classifiers are applied to extract the text polarity.

The rest of the paper is organized as follows: Section 2 reviews the previous studies and related work, while Section 3 introduces the proposed model. The experimental results are presented in Section 4.

II. Related Literature

Numerous investigations on sentiment analysis approaches have been conducted. The English language has the largest number of research works, while the research efforts exerted for the other languages, including Arabic, are more restricted. This section examines the research work conducted in the field of Arabic Sentiment Analysis (ASA).

Most of the research efforts on ASA studies focused on text processing in a public domain or in news articles, while few efforts were developed in specific domains such as [1], [2], [13]–[18].

Some of the research studies achieved low accuracy results with ML classifiers, as in [1]. On the other hand, some of the research studies used two balanced classes to avoid bias and to achieve better results, such as [15] and [19].

Nabil, Aly, and Atiya [13] used an automatic approach to construct their sentiment dataset in a public domain. They collected 84000 Arabic tweets, and then they determined the most active Egyptian twitters to get the list of the top 30 users. Finally, they filtered the top recent Hashtags to get a list of 2500, and they called it ASTD; it consists of 10,006 Arabic Hashtags classified into machine learning algorithms "SVM, LR, M-NB, B-NB, KNN, SGD, Passive Aggressive, and Linear perceptron" into Subjective "positive 793, negative 1684, neutral 832" and Objective 6691 which has no opinion. Moreover, the objective class doesn't have any effect on sentiment, and its size is too big compared to subjective, positive, and negative classes. The used TF-IDF and CBOW as Text feature and accuracy results showed the best value with B-NB classifier with accuracy 74, 9%.

Abdellaoui, and Zrigui [14] used an automatic approach to construct their sentiment dataset. They collected 5,615,943 Arabic tweets, and then they determined the top 20 most used emojis on Twitter. After that, a list of the ten most used Emojis on Twitter is selected. They dealt with four different dialects, "Egyptian, Levan, Maghrebi, and Gulf," they also used various lexicons to translate dialects to modern standard Arabic MSA. After filtering, they called it TEAD; classifying it with machine learning algorithms "SVM, LR, M-NB, B-NB, DT and RF" into three classes "positive 3,122,615, negative 2,115,325, neutral 378,003 by using TF-IDF and CBOW as text feature and accuracy results showed the best value with SVM classifier with accuracy 84,8%. In this study, they translated dialects to MSA before preprocessing to facilitate the classification process. Further, the number of neutral classes is too small compared to others.

ElSahar and El-Beltagy [20] used an automatic approach to the annotated dataset. They collected four domains as follows "Hotel Reviews, Restaurant Reviews, Movie Reviews, and Product Reviews (PROD)." The dataset was divided into "15K, 8.6K, 1.5K and 15K Arabic reviews for each domain". They dealt with different dialects, "Egyptian, Gulf, and MSA." After filtering, they called each one as (HTL, RES, MOV, and PROD); it classified into two classes "positive, negative," using different machine learning algorithms as "Linear SVM, B- NB, LREG, SGD and KNN" and SVM showed the best accuracy as 82.4%. In this study, they tested the model for each domain separately, so they achieved good results.

Al Mukhaiti, Siddiqui, and Shaalany [1] utilized a new dataset by gathering data from different resources, such as Twitter, Facebook, and Instagram. Thus, overall, 58% of the reviews collected are from YouTube, 37% from Facebook, and 5% from Instagram. They manually annotated the filtered data as negative and positive and segregated them. The best result was 77.7% for accuracy. The study was in the general domain; also, the accuracy results are low despite using two classes.

El-Masri, Altrabsheh, Mansour, and Ramsay [2] utilized a new tool that applies sentiment analysis to Arabic text tweets using a combination of parameters. They tested their work in 8000 tweets with lexicon and machine learning results, and accuracy showed 66.5% with dictionary-based and 34% for SVM.

Oussous, Benjelloun, Lahcen, and Belfkih [15] decided to extract 2000 Moroccan reviews: 1000 positive and 1000 negative, and manually annotated them. They tested their system with machine learning and deep learning techniques. The best experimental results showed 80% with SVM and 95.5% for CNN.

Refaee and Rieser [8] made an Arabic dataset for conclusion investigation, which contains 2000 tweets; categorized into positive, the main half, and negative, the subsequent half. Two techniques were applied to the dataset: corpus-based "Administered Learning" and dictionary-based "Unaided Learning." Four regulated AI calculations were used, i.e., SVM, NB, D-Tree, and K-Nearest Neighbor. The SVM and NB got better outcomes, around 80%. Then again, the vocabulary-based methodology demonstrates that with a huge dictionary, the exactness results were improving. There El-Beltagy, Kalamawy, and Soliman [16] also developed an Arabic sentiment analysis task. The authors were ranked first in the SemEval 2017 task for Arabic SA. They used a set of hand-engineered and lexicon-based features, the classifier of choice was a complement NB classifier, and the accuracy result showed 77%.

Gamal, Alfonse, El-Horbaty, and Salem [17] used a dataset that included more than 151,000 different opinions in variant Arabic dialects, which are labeled into two balanced classes, namely, positive and negative. Various machine learning algorithms are applied to this dataset, including the ridge regression, which gives the highest accuracy of 99.90% with ridge Regression (RR) classifier and 98.95% with SVM. The study showed good results as they used two balanced classes.".

III. Proposed Model

The proposed model aims to extract the people's opinions in Arabic text. The opinions can be classified into three classes: positive, negative, and neutral. The proposed model is based on machine learning algorithms, where six different machine learning algorithms are exploited: Naïve Bayes (NB), Support Vector Machines (SVMs), Decision Tree (DT), Stochastic Gradient Descent (SGD), Logistic Regression (LR), and Random Forest (RF) [21], [22].

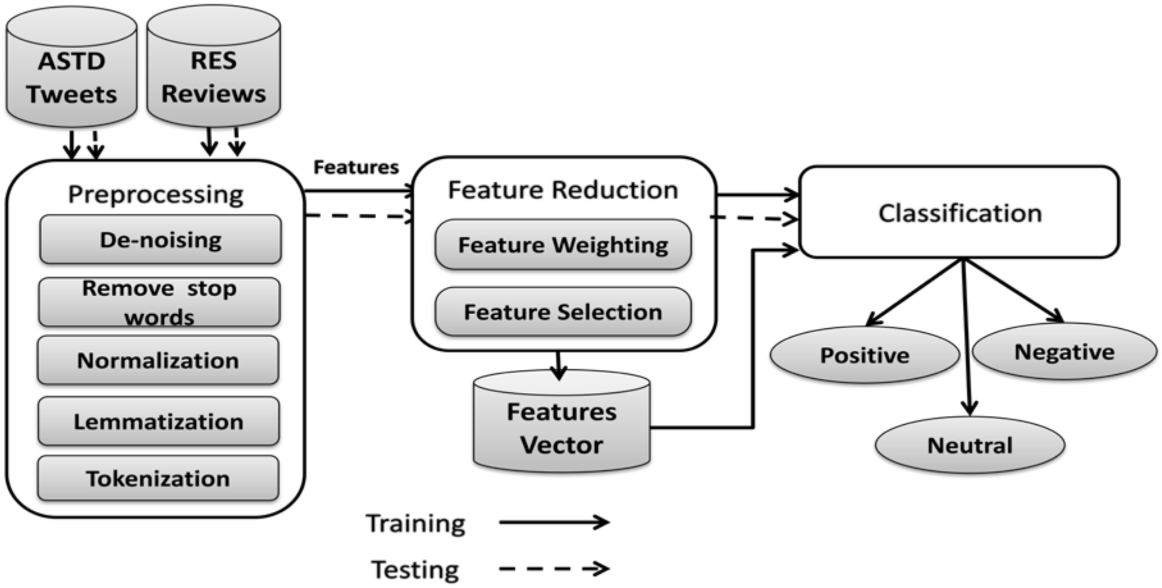

Fig. 1. The Proposed Model.

As shown in Fig. 1, the model consists of three modules, namely the preprocessing, feature reduction, and classification modules. In the preprocessing module, the data should be cleaned and transformed into a format that could be fit into the feature reduction phase. It consists of de-noising, stop-words removal, normalization, lemmatization, and tokenization steps. The feature reduction module is divided into feature weighting and feature selection, which are responsible for scoring each feature and for selecting the most effective features to build the features vector, respectively. Both Sections 3.1 and 3.2 describe the preprocessing and the feature reduction modules in detail, respectively.

Finally, the classification module is applied by using six different machine learning algorithms to classify tweets into three classes (positive, negative, and neutral), as will be explained in detail in Section 3.3. As shown in Fig. 1, the classification process is achieved in two phases: the training and the testing phases, which are represented as solid and dashed arrows, respectively.

A. The Preproccessing Module

The preprocessing process is typically conducted to convert the text into textual features that fit into the SA methods. The preprocessing was applied in a set of sequential steps on two different datasets: Egyptian-Dialect tweets and social media reviews datasets. These steps are tokenization, de-noising, normalization, stop-words removal, and lemmatization.

The preprocessing algorithm is illustrated in Algorithm (1), where the tokenization is the task of chopping a sequence of characters in a document up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation, space, and punctuation marks [19], [23]. Then, it applies the de-noising step that simply involves using a neural sequence transaction model to back translate the noised text to the original clean text [24]. The de-noising text includes removing URLs, which began by (http ://) until the following space, hash-labels subjects, mentions, punctuations, and special characters. In addition, the symbols for emotions are removed [25].

Algorithm 1 Pre-processing

```
//Tokenization
Input:        tweets_table
Output:       token_table
While (more tweets exit in tweet table) do
        { T=next tweet; P=1;
        While not end of T
                { I= next_space (T);    // determine the position of next
space in T
                 Token = tirm_substring (T,P,I,string);
                  Insert into token_table (tweet number,     token);
                 T=rest_string;}
//De-noising and normalization
Input:        tokens_table

Output:       Updated tokens_table
Arabic_punctuations = {"‘, `, ÷, ×, <,>, _, (,), *, &, ^,], [, ,+ ,¦ ,~ ,{,} ,‘ ,“ ,/ ,-
,,,,,,,, , , ,-,–,“,…,”,!,|}
Emotions = {⌨, ↪, •,          ...#...}
S=a set of Arabic vowels
While (more token exist in tokens_table) do
{T=next token;
F= the first letter of T;
While (F in Emotions or Arabic_punctuations or Emotions or S) do // remove
symbols from the beginnig of token
                        {trimleft (T.F.T):
                         F= the first letter in T:}
            E= the last letter of T;
While (E in symbols) do //remove symbols from the end of token
                        {trimright (T.F.T):
                         E= the last letter of T;}
                    Concatenation (T, E, T);
 If (T starts with # or @ or the length of T<=2) then
              Delete T from tokens_table}
//Stop word removing
While (more token exist in tokens_table) do
{T=next token;
If (T in stop_words_table then
 Delete the entry of T from token_table}
//Lematization
While (more token exist in tokens_table) do
{T=next token;
        If (T=3) then
              Lemma=T
              Root=T
        If (T>t) then
              Remove suffixes from T
              Lemma=T}

        }
```

While the normalization is responsible for transforming text into a single canonical form [19], for example:

- Removing Arabic short vowels (diacritics) [26] such as " َ ، ِ ،ًّ".
- Removing the duplicate characters that doesn't influence the significance of the word such as "جميييييييل" is converted to "جميل" "Beautiful" [26].
- Replacing the letter "ة" to the letter "ه ", replacing the letters "آ ، إ ، أ" to the letter "ا".
- Normalizing the words that can be written in different forms to one form. For example, the "emergency" word can be written in Arabic language as "الطوارى" or "الطوارئ". It should be mapped into "الطوارئ".
- Normalizing compound words by going along with them by the character "_ "; for example, "في عالم موازي", which will be standardized to "فى_عالم_موازى" ""in_aparallel_world".
- Furthermore, in removing stop words, there are some words which don't have any effect on sentiment analysis, such as "من ، في ، عن"”from, in, about”. These stop words are removed using a pre-defined list of stop words [15]. Here an example for removing stop words form a tweet “ موقعه مناسب ممكن الوصول اليه بسهولة عن طريق التاكسي” “Convenient location, easily accessible by a taxi” comes to “ موقعه مناسب ممكن الوصول بسهولة طريق التاكسي”, which minimized data size.

The last step is lemmatization, which converts inflected words to their root form; It helps to minimize the dataset size, which improved the accuracy and performance. For example, “معاك لو تعيدها الف مره” "With you if you repeat it a thousand of times" begins to “مع لو أعاد ألف مر”.

After the tweets/reviews data are preprocessed, a list of features is generated and can be represented as ([f1], [f2] … [fn]).

B. The Feature Reduction Module

The Feature reduction module aims to reduce the extracted features. It consists of two processes, which are feature weighting and feature selection. In feature weighting, the extracted features are weighted using the Term Frequency-Inverse Document Frequency (TF-IDF).

Finally, feature selection is applied to select the most candidate features. Then, the selected features will be used as an input for the classification modules.

1) Feature selection: Term frequency-inverse document frequency (TF–IDF) [27] was applied to find strongly affected words for sentiment by evaluating the weight of the feature, where each term ‘t’ is weighted by a weight ‘w’ using TF-IDF [28]. The Term frequency computes the number of instances of “t” term that appears in the tweets/reviews document, as shown in Eq. (1).

$$Tf(t,d) = (fd(t))/(\max[fd(t)]) \quad (1)$$

Where fd (t) is the frequency of the term “t” in tweet/review “d”.

The Inverse Document Frequency (IDF) is used to estimate the importance of a given term. It measures how rare a given term in the whole tweet/ review document, using Eq (2):

$$IDF(t) = \log(N/dft) \quad (2)$$

Where dft is the number of tweets /reviews with the term “t”, and “N” is the total number of tweets / reviews per document.

Finally, the TF-IDF is calculated as the result of the multiplication of TF and IDF as shown in Eq(3):

$$TF\text{-}IDF = tf(t,d)\ IDF(t) \quad (3)$$

2) Feature weighting: After feature weighting, the feature selection process is applied to select the most important and relevant features. The used techniques are classified into two categories are supervised and unsupervised techniques. The supervised technique is used for labeled data to identify the relevant features for increasing the efficiency of supervised models like classification and regression. In unsupervised technique, it can be used for unlabeled data.

In addition, threshold-based approach is applied to identify the least number of features as shown in Table I.

TABLE I. TWEETS-BASED EXPERIMENT ACCURACY RESULTS USING DIFFERENT NUMBER OF FEATURES

NUMBER OF FEATURES		SGD	DT	M-NB	B-NB	SVM	LR	RF
2000	ACCURACY	68	50	60	68	45	62	62
5000		71	57	65	75	50	70	64
10000		78	74	78	82	78	78	77.8
15000		78	74	78	82	78	78	77.8

Cross-validation is used to tune the soft margin parameter C (λ = 1/2C). Higher values of C add a higher penalty for the misclassified points rather than maximizing the separation margin. Therefore, the optimization problem will lead to a larger number of selected features to reduce the misclassified errors then the threshold value is determined for each experiment to get the results for accuracy.

C. The Classification Module

The most important advantage of ML algorithms is that they deal with a complex problem and gets closer results than humans. In Arabic text processing, complex problems exist and need efficient solutions to solve these problems [6], [29], [30].

In the proposed model, six supervised machine-learning (ML) classifiers were exploited: NB, LR, SVM, DT, SGD, and RF. The used NB algorithms are Multinomial Naive Bayes (M-NB) and Bernoulli Naive Bayes (B-NB). These classifiers have been chosen since they are the most commonly used in the literature of SA [7], [31], [32]. These algorithms are applied to the training set to build the classification model.

SVM is one of the most robust prediction methods based on statistical learning frameworks [33]–[35]. While the NB assumption of attribute independence works well for text categorization at the word feature level [35], [36]. On the other hand, DT is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility [37], [38].

Furthermore, the SGD is an iterative method for optimizing an objective function with suitable smoothness properties. RF is an ensemble learning method for classification, regression, andother tasks. It is operated by constructing many decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression) of the individual trees [37]–[39].

IV. Experimental

This section presents the experiments conducted in this research and a discussion of the results. The Two experiments have been conducted using two different data sets: Arabic Sentiment Tweets Dataset (ASTD) [13] and Restaurant Reviews dataset (RES) [20].

A. *Tweets-based Experiment*

In this experiment, the ASTD [13] dataset is used, which contains 10,006 Egyptian-Dialect tweets that are divided as follows: 793 positive sentiment, 1684 negative sentiment, 832 neutral sentiments, and 6691 objective tweets. An example for a positive labeled tweet statement is "محبين البرنامج بيزيدوا", which is equivalent in English to “fans of El-Bernameg are increasing”. Further, ASTD data were used in different ASA research because it is completely available. In the first experiment, as the objective class isnot effective in SA, only three classes were considered: positive, negative, and neutral with a total data size of 3,316 Egyptian-Dialect tweets.

Six different classifiers were applied in the tweets-based experiment: DT, SVM, RF, LR, M-NB, B-NB, and SGD. It has been applied to the ASTD Egyptian-Dialect tweets dataset with all the preprocessed steps described above, and hence a feature vector, which consists of 10000 top selected features has been created.

Table II illustrates the evaluation measures of the proposed model using the precision, recall, and accuracy measures. Since tweets in our model are divided into three classes, we have three precisions and recalls value for each class to be calculated [40], [41]by the following Eq. (4):

$$Precision_i\,Or\,Recall_i = \frac{\textit{Tweets correctly assigned to class}_i}{\textit{Tweets attributed to class}_i} \quad (4)$$

The B-NB scored the best accuracy with 82 %, followed by SVM, LR, M-NB, and SGD with 78% accuracy, as shown in Fig. 2.

Further, the tweets-based experiment results are compared with the related works that used the ASTD Egyptian-Dialect tweets dataset, as shown in Table III. The results show the different machine learning algorithms used by the proposed model and the related work with their evaluation measures.

TABLE II. Tweets-based Experiment Classifiers Results using ASTD Egyptian-Dialect Tweets Dataset

	SGD	DT	M-NB	B-NB	SVM	LR	RF
ACCURACY	78	74	78	82	78	78	77.8
PRECISION	77	72	80	82	78	77	70
RECALL	78	74	79	82	78	78	86

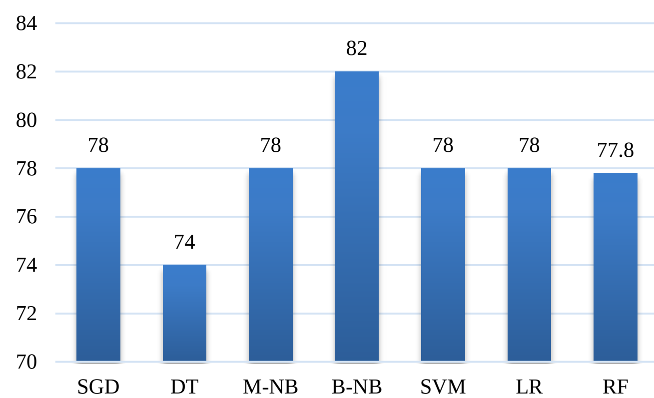

Fig. 2. Proposed Model Accuracy Results for Tweets-based Experiment.

TABLE III. Tweets-based Evaluation Results: the Proposed Model Versus the Related Work using ASTD Egyptian-Dialect Tweets

		SGD	DT	M-NB	B-NB	SVM	LR	RF
Abdellaoui&Zrigui[14]	accuracy	--	68.7	74.4	74.9	75.5	74.9	68.7
	precision	--	78	72	81	75	76	84
	Recall	--	73	72	74	76	74	73
Kaseb and Ahmed [42]	accuracy	--	--	--	--	64	--	--
	precision	--	--	--	--	58.3	--	--
	Recall	--	--	--	--	63.9	--	--
Nabil and Atiya[13]	accuracy	67.1	--	67	66.9	68.9	67.6	--
The proposed model	accuracy	78	74	78	82	78	78	77.8
	precision	77	72	80	82	78	77	70
	Recall	78	74	79	82	78	78	86

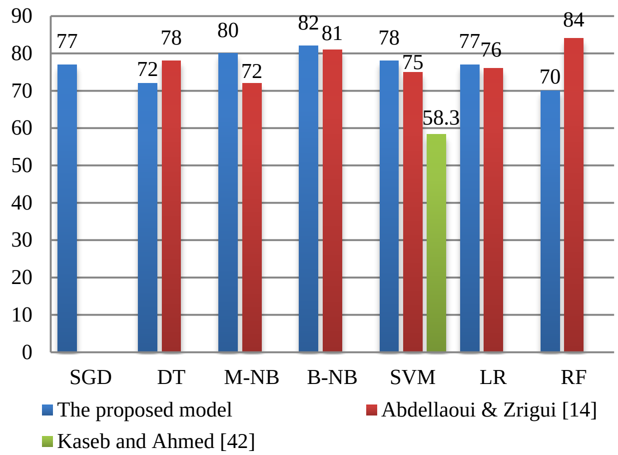

Fig. 3. Precision Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

It is noticed that the proposed model has better values with all classifiers compared with Nabil and Atiya [13] who classified the ASTD Egyptian-Dialect tweets dataset into four classes (positive, negative, neutral, and objective); while the proposed model classified the ASTD Egyptian-Dialect tweets dataset into three classes (positive, negative, and neutral), which are the most popular classes. Further, Nabil and Atiya didn't apply DT and RF classifiers; on the other hand, they are applied by the proposed model.

Both Fig. 3 and Fig. 4 show the precision and the recall for the proposed classifiers model versus two related works.

In Abdellaoui and Zrigui [14] research, SGD is not applied to the ASTD Egyptian-Dialect Tweets dataset, but the proposed model applies it. Moreover, the proposed model achieved better values with all classifiers compared with the above study.

On the other hand, Kaseb, and Ahmed [42] filtered and cleaned ASTD Egyptian-Dialect tweets to 1652 records. Unfortunately, they applied only one SVM classifier and achieved a lower accuracy of 64% compared with the proposed model with 78%.

As shown in Fig. 5, the proposed model with B-NB, SVM, LR, M-NB, and SGD achieved a better accuracy compared to the related work. B-NB achieved 82% versus 66.9% and 74.9% for the related works [13]&[14]. While SVM achieved 78 % versus 68.9%, 75.5%, and 64 % for the related works. Further, LR has higher accuracy with 78% compared to related works with 67.6% and 74.9%. M-NB achieved 78% with higher accuracy than the related works.

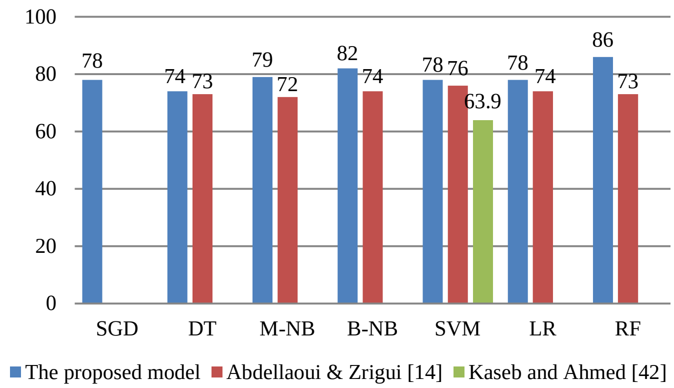

Fig. 4. Recall Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

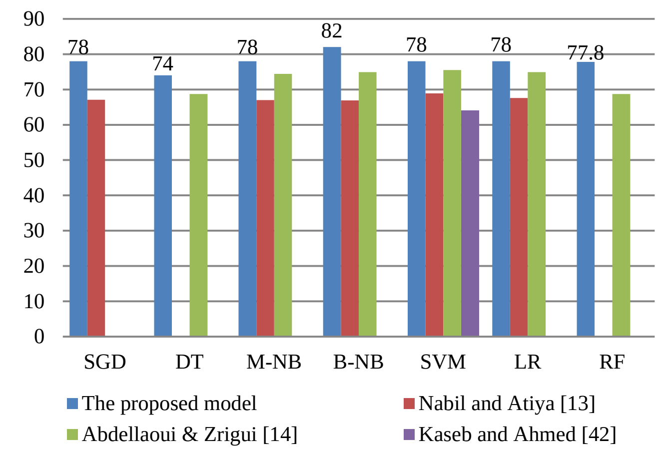

Fig. 5. Accuracy Value Results for Proposed Model Compared with Related Works for the Tweets-based Experiment.

In DT classifier, our module has higher accuracy rating than [14], but without knowing its recall, we cannot comfortably trust the results. Interestingly, recall and accuracy are often at odds with each other, as attempts to boost recall often negatively impact accuracy and vice versa.

B. Review-based Experiment

The Restaurant Review dataset RES [20] is used in the review-based experiment, which was collected from the trip advisor site with a total number of 10,871 reviews. The RES is divided as follows 8021 positive sentiments, 2625 negative sentiment, and 225 neutral reviews. An example of a positive sentiment tweet is "مطعم ممتاز و خدمة حلوى أوى و مكان متميز و معاملة راقية" which is equivalent in English to "Excellent restaurant, great service, excellent place and classy treatment" [20].

TABLE IV. COMPARING REVIEWS-BASED EVALUATION RESULTS AND THE RELATED WORK USING RES DATASET

		SGD	**DT**	**M-NB**	**B-NB**	**SVM**	**LR**	**RF**	**KNN**
ElSahar and El-Beltagy[20]	accuracy	78.4	--	--	82.1	81.4	70.4	--	49.5
The proposed model	accuracy	85.6	76.9	82	79.8	87.2	85.9	83.8	--
	precision	84	74	83	80	85	82	84	--
	Recall	86	77	82	80	87	84	86	--

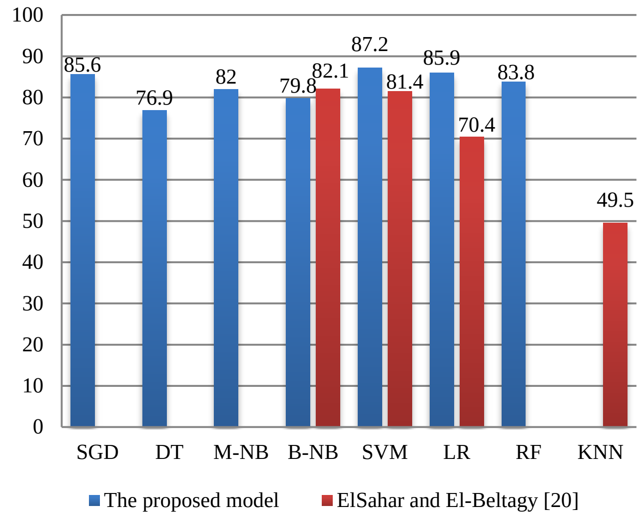

Fig. 6. Accuracy Value Results for Proposed Model Compared with Related Works for the Review-based Experiment.

The review-based experiment is applied using different classifiers DT, SVM, RF, LR, M-NB, B- NB, and SGD with a feature vector that consists of 15000 top selected features. The review-based experiment evaluation results are shown in Table III. It has been revealed that SVM scored the best accuracy with 87.2%, followed by LR with 85.9%, and SGD with 85.6% accuracy. On the other hand, the tweets-based experiment results are compared with the related works that use the RES dataset, as shown in Table IV. The table presents the different machine learning algorithms used by the proposed model and the related work with their evaluation measures. It is noticed that the proposed model has better values with most classifiers compared with ElSahar and El-Beltagy[20]. Also, ElSahar and El-Beltagy did not apply DT and M-NB classifiers, while the proposed model applies them.

Moreover, Fig. 6 Shows the accuracy of the proposed classifier model versus the previous works. The proposed model with SVM, LR, and SGD achieves better accuracy than the previous works. SVM achieved 87.2% versus 81.4% for the previous work while LR achieved 85.9 % versus 70.4% for the related work.

V. Conclusion

The paper has introduced a new model for Arabic sentiment analysis and the effect of different text preprocessing techniques on classification accuracy. The proposed model was evaluated using recall, precision, and accuracy measures. Two different types of Arabic datasets are used: (1) ASTD is an Egyptian-Dialect tweets, and (2) RES, which is Multi-Dialect reviews. Two main experiments have been conducted using machine learning algorithms (DT, SVM, RF, LR, M-NB, and B-NB). The first experiment was applied to the ASTD dataset with 3,316 Egyptian-Dialect tweets. It is noticed that B-NB scored the best accuracy with 82%, followed by SVM, LR, M-NB, and SGD with 78% accuracy. The second experiment was applied to the RES dataset with 10K Multi-Dialect Arabic reviews. In addition, SVM achieved accuracy with 87.2%, followed by LR with 85.9%, and SGD with 85.6%. These results revealed that the proposed model outperformed the related works in the two conducted experiments.

Further, the experiments showed that de-noising, stop words removal, lemmatization, and normalization slightly improved the classification's performance. The proposed model will use different techniques in future work, such as deep learning or lexicon-based approaches.

References

[1] "Dataset Built for Arabic Sentiment Analysis | SpringerLink." https://link.springer.com/chapter/10.1007/978-3-319-64861-3_38 (accessed Oct. 29, 2021).

[2] M. El-Masri, N. Altrabsheh, H. Mansour, and A. Ramsay, "A web-based tool for Arabic sentiment analysis," ProcediaComput. Sci., vol. 117, pp. 38–45, 2017.

[3] B. Liu and L. Zhang, "A survey of opinion mining and sentiment analysis," in Mining text data, Springer, 2012, pp. 415–463.

[4] R. K. Bakshi, N. Kaur, R. Kaur, and G. Kaur, "Opinion mining and sentiment analysis," in 2016 3rd international conference on computing for sustainable global development (INDIACom), 2016, pp. 452–455.

[5] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," Knowl.-Based Syst., vol. 89, pp. 14–46, 2015.

[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," ArXivPrepr. Cs0205070, 2002.

[7] M. S. Neethu and R. Rajasree, "Sentiment analysis in twitter using machine learning techniques," in 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), 2013, pp. 1–5.

[8] E. Refaee and V. Rieser, "An arabic twitter corpus for subjectivity and sentiment analysis.," in LREC, 2014, pp. 2268–2273.

[9] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Arabic language sentiment analysis on health services," in 2017 1st international workshop on arabic script analysis and recognition (asar), 2017, pp. 114–118.

[10] E. Benmamoun, The feature structure of functional categories: A comparative study of Arabic dialects. Oxford University Press, 2000.

[11] I. Moawad, W. Alromima, and R. Elgohary, "Bi-Gram Term Collocations-based Query Expansion Approach for Improving Arabic Information Retrieval.," Arab. J. Sci. Eng. Springer Sci. Bus. Media BV, vol. 43, no. 12, 2018.

[12] A. Hamdi, K. Shaban, and A. Zainal, "A review on challenging issues in arabic sentiment analysis," 2016.

[13] M. Nabil, M. Aly, and A. Atiya, "Astd: Arabic sentiment tweets dataset," in Proceedings of the 2015 conference on empirical methods in natural language processing, 2015, pp. 2515–2519.

[14] H. Abdellaoui and M. Zrigui, "Using tweets and emojis to build tead: an Arabic dataset for sentiment analysis," Comput. Sist., vol. 22, no. 3, pp. 777–786, 2018.

[15] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, and S. Belfkih, "ASA: A framework for Arabic sentiment analysis," J. Inf. Sci., vol. 46, no. 4, pp. 544–559, 2020.

[16] [16] S. R. El-Beltagy, M. E. Kalamawy, and A. B. Soliman, "Niletmrg at semeval-2017 task 4: Arabic sentiment analysis," ArXivPrepr. ArXiv171008458, 2017.

[17] D. Gamal, M. Alfonse, E.-S. M. El-Horbaty, and A.-B. M. Salem, "Twitter benchmark dataset for Arabic sentiment analysis," Int J Mod EducComputSci, vol. 11, no. 1, p. 33, 2019.

[18] L. Khreisat, "A machine learning approach for Arabic text classification using N-gram frequency statistics," J. Informetr., vol. 3, no. 1, pp. 72–77, 2009.

[19] S. Bedrick, R. Beckley, B. Roark, and R. Sproat, "Robust kaomoji detection in Twitter," in Proceedings of the Second Workshop on Language in Social Media, Montréal, Canada, Jun. 2012, pp. 56–64.

Accessed: Oct. 29, 2021. [Online]. Available: https://aclanthology.org/W12-2107.

[20] H. ElSahar and S. R. El-Beltagy, “Building large arabic multi-domain resources for sentiment analysis,” in International Conference on Intelligent Text Processing and Computational Linguistics, 2015, pp. 23–34.

[21] G. Bonaccorso, Machine learning algorithms. Packt Publishing Ltd, 2017.

[22] T. O. Ayodele, “Types of machine learning algorithms,” New Adv. Mach. Learn., vol. 3, pp. 19–48, 2010.

[23] R. M. Badry and I. F. Moawad, “A semantic text summarization model for Arabic topic-oriented,” in International Conference on Advanced Machine Learning Technologies and Applications, 2019, pp. 518–528.

[24] Z. Xie, G. Genthial, S. Xie, A. Ng, and D. Jurafsky, “Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction,” in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), New Orleans, Louisiana, Jun. 2018, pp. 619–628. doi: 10.18653/v1/N18-1057.

[25] A. Pak and P. Paroubek, “Twitter as a corpus for sentiment analysis and opinion mining.,” in LREc, 2010, vol. 10, no. 2010, pp. 1320–1326.

[26] “Key issues in conducting sentiment analysis on Arabic social media text | IEEE Conference Publication | IEEE Xplore.” https://ieeexplore.ieee.org/abstract/document/6544396/ (accessed Oct. 29, 2021).

[27] J. Ramos, “Using tf-idf to determine word relevance in document queries,” in Proceedings of the first instructional conference on machine learning, 2003, vol. 242, no. 1, pp. 29–48.

[28] E. M. Bahgat, S. Rady, W. Gad, and I. F. Moawad, “Efficient email classification approach based on semantic methods,” Ain Shams Eng. J., vol. 9, no. 4, pp. 3259–3269, 2018.

[29] “A multi-objective feature selection method based on bacterial foraging optimization | SpringerLink.” https://link.springer.com/article/10.1007/s11047-019-09754-6 (accessed Oct. 29, 2021).

[30] P. Wang, Y. Li, and C. K. Reddy, “Machine learning for survival analysis: A survey,” ACM Comput. Surv. CSUR, vol. 51, no. 6, pp. 1–36, 2019.

[31] M. Alruily, “Classification of Arabic Tweets: A Review,” Electronics, vol. 10, no. 10, p. 1143, 2021.

[32] M.-Y. Cheng, D. Kusoemo, and R. A. Gosno, “Text mining-based construction site accident classification using hybrid supervised machine learning,” Autom. Constr., vol. 118, p. 103265, 2020.

[33] R. Moraes, J. F. Valiati, and W. P. G. Neto, “Document-level sentiment classification: An empirical comparison between SVM and ANN,” Expert Syst. Appl., vol. 40, no. 2, pp. 621–633, 2013.

[34] A. Borg and M. Boldt, “Using VADER sentiment and SVM for predicting customer response sentiment,” Expert Syst. Appl., vol. 162, p. 113746, 2020.

[35] S. Rana and A. Singh, “Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques,” in 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016, pp. 106–111.

[36] M. Govindarajan, “Sentiment analysis of movie reviews using hybrid method of naive bayes and genetic algorithm,” Int. J. Adv. Comput. Res., vol. 3, no. 4, p. 139, 2013.

[37] A. Singh, N. Thakur, and A. Sharma, “A review of supervised machine learning algorithms,” in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310–1315.

[38] J. Snoek, H. Larochelle, and R. P. Adams, “Practical bayesian optimization of machine learning algorithms,” Adv. Neural Inf. Process. Syst., vol. 25, 2012.

[39] L. Bottou and O. Bousquet, “The tradeoffs of large-scale learning: Optimization for Machine Learning, 351,” 2011.

[40] K. A. Djaballah, K. Boukhalfa, and O. Boussaid, “Sentiment analysis of Twitter messages using Word2vec by weighted average,” in 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS), 2019, pp. 223–228.

[41] K. Boukhalfa and O. Boussaid, “Sentiment analysis of twitter messages using Word2vec by weighted average”.

[42] M. Farouk and G. Kaseb, “Extended-ATSD: Arabic Tweets Sentiment Dataset,” vol. 14, pp. 4780–4785, May 2019.