SENTIMENT ANALYSIS FOR ARABIC TWEETS DATASETS: LEXICON-BASED AND MACHINE LEARNING APPROACHES

AHMAD ALOQAILY[1], MALAK AL-HASSAN[2], KAMAL SALAH[3], BASIMA ELSHQEIRAT[4], MONTAHA ALMASHAGBAH[5]

[1, 5] Prince Al Hussein Bin Abdullah II faculty for Information Technology, Hashemite University,
P.O. Box 150459, Zarqa 13115, Jordan.

[2, 4] King Abdullah II School of Information Technology, The University of Jordan,
P.O Box 11942, Amman, Jordan.

[3] Deanship of preparatory year and supporting studies, Imam Abdulrahman Bin Faisal University,
P.O Box 1982, Dammam, Saudi Arabia

E-mail: [1]aloqaily@hu.edu.jo, [2]m_alhassan@ju.edu.jo, [2]kisalah@iau.edu.sa , [4]b.shoqurat@ju.edu.jo,
[5]montaha.mashagbah@yahoo.com

ABSTRACT

Recently, Sentiment Analysis applied to social media data has gradually become one of the significant research interest in the data mining domain due to the large volume of data available on social media networks. Sentiment Analysis is concerned with analyzing text to identify opinions or emotions and categorizing them as positive, negative or neutral. Applying sentiment analysis to short texts such as Twitter messages is a challenging task because tweets might contain a combination of formal and informal language, special characters, emojis and symbols. Therefore, it is often difficult to understand the semantics of the text and it is complex to extract the proper emotions expressed by users.

In this paper, sentiment analysis approaches, namely: lexicon-based and machine learning approaches, are applied and evaluated on an Arabic tweets dataset (short texts) regarding the Syrian civil war and crises. The experimental results revealed that machine learning approaches outperformed the lexicon-based in the context of predicting the subjectivity of tweets. In terms of machine learning, five popular machine learning algorithms were applied and evaluated. According to the experimental results, the Logistic Model Trees (LMT) algorithm achieved the highest performance results, followed by the simple logistic and the SVM algorithms, respectively. The results also showed that there are enhancements in performance when utilizing feature selection approaches. Based on all performance evaluation measures, the LMT algorithms reported the best performance results (*Acc= 85.55, F1= 0.92 and AUC= 0.86*).

Keywords: *Machine Learning; Lexicon-Based Approach; Sentiment Analysis; Opinion Mining; Social Media; Twitter Datasets.*

1 INTRODUCTION

Nowadays, the Internet has become a valuable and useful source of information, events, news and opinions available on social media websites, such as Twitter and Facebook. Currently, Twitter has more than 330 million monthly active users [1]. Through Twitter, people can express their opinions and feelings, companies can get their clients' feedbacks and politicians can be in touch with their constituents and increase the number of their supporters [2]. With the availability of such abundant data, the ability to investigate people's views and opinions have become more accessible and feasible.

Consequently, there is a desperate need to process, analyze and eventually extract knowledge from data as opinions concerning significant issues, entities or topics. Analyzing Twitter data, for example, is not a trivial task and depends on the semantics of tweets, which are short concise texts (maximum 140 characters). This type of analysis is called Sentiment Analysis (SA) or Opinion Mining (OM). As stated in [3] sentiment is defined as "an attitude, thought, or judgment prompted by feeling".

While in [4], Sentiment Analysis, which falls under the data mining domain, focuses on extracting opinions or feelings from text and categorizing them as positive, negative or neutral. The process of SA begins with preprocessing (preparation) of the available dataset, then data analysis and ends up with data visualization [5].

In Sentiment Analysis, extracting the polarity of a given text, text mining techniques are usually utilized [6]. The part in which attention is taken within text mining is opinion mining, where the goal of opinion mining is to evaluate whether a given text is objective or subjective and whether it is positive or negative [7]. The objective text contains facts, whereas subjective text contains opinions, sentiments or emotions about topics or entities [8]. The opinions are usually expressed and published by users in different forms, such as document, reviews, short comments and tweets. The language used for expressing such opinions is also varied. Therefore, extracting the semantic or opinions in text is a demanding task. The subjective analysis is the way to measure the opinion or polarity of a given text.

Currently, the Arabic language has become widely spread on social networking sites and more attention must be paid to such language. Recently, considerable attention has been given to mining opinions from Arabic texts, and as a result, researchers have an interest in developing an Arabic lexicon for a word-level sentiment evaluation [6]. The number of works that have addressed SA in the Arabic language is limited [2, 9, 10]. One of the earliest research works concerning Arabic SA is presented and discussed in [2], where the goal was to mine Arabic business reviews. Mohammad, et al. [10] and Al-Kabi, et al. [11], further, proposed research works that evaluate translating Arabic text to the English language and then employ English sentiment analysis. The problem with their works is that special features for the Arabic language were not utilized.

This research paper provides an overview of the Arabic sentiment analysis. Specifically, this paper aims to provide a comprehensive review of the literature concerned with Arabic political tweets. Furthermore, the challenges that face short Arabic text sentiment analysis need to be considered and clarified. To accomplish this task, we have applied the concept of SA on the Arabic texts for Twitter datasets. The utilized dataset is based on tweets posted concerning the Syrian crises.

In this paper, sentiment analysis approaches, namely: lexicon-based and machine learning approaches, are employed and evaluated on the utilized dataset. An Arabic Sentiment Analysis methodology is proposed to accomplish this task. The proposed methodology employs machine learning techniques and a lexicon-based approach, using the Arabic Emoticon Lexicon, to explore the performance of the proposed methodology. Then, the results are analyzed in terms of accuracy, F-measure and Area Under the Curve performance measures. These measures are used to evaluate the efficiency of machine learning algorithms empirically and the best SA models are reported.

The remaining of this paper is structured as follows: Section 2 describes previous studies related to the intended research topic, including the motivation of this research work. Section 3 presents the research methodology. The experimental and performance evaluation results are discussed in Section 4. Finally, Section 5 concludes and points out the main findings of this paper and presents some possible future works.

2 LITERATURE REVIEW

This section summarizes related works in the domain of SA related to analyzing short Arabic text using machine learning and Lexicon-based approaches.

Several studies focused on analyzing short texts written in English and Arabic languages, such as textual datasets that are available on social media networks. Most studies related to sentiment analysis have been conducted for the English language and less attention has been paid to other languages such as the Arabic language. The Arab Social Media Report (ASMR) reported in 2017 that Twitter had more than 11 million active users in the Arab region. Also, the ASMR said that the total number of tweets had exceeded 849 million by March 2016. Another report stated by ASMR showed that Saudi Arabia had the highest number of active users, more than 2.6 million users (around 29% of all Twitter users in the region), with on average five tweets a day [13].

In this work, studies that are related to analyzing short text Arabic language are mainly reviewed and discussed. Al-Azani and El-Alfy [14] compared the performance of different compositions to determine polarization in short, highly balanced textual data sets of dialectical Arabic tweets. Based on their experimental results, two techniques, namely: word embedding with the ensemble and SMOTE were applied, an improvement, in terms of the F1 measure can be achieved (15% more over the baseline). Three main cases of SA were considered and addressed in their work: dealing with micro-blogging data, dealing with unequal class distribution in opinion mining, and treating colloquial Arabic.

Zhao, et al. [15] discussed the challenges of Short text classification, which are: sparse nature, noise words, syntactical structure and colloquial terminologies used. Furthermore, the authors discussed the effectiveness of the algorithms used to solve short text challenges using standard analytical measures. Some challenges regarding short text sentiment analysis were also mentioned in [16], which are: limited in length, usually spanning one sentence or less. Texts tend to have many spelling errors, colloquial and abbreviated words and symbols such as hashtags that ease the search task and often point out a topic or opinion. Examples of short text appear in several contexts, such as online reviews, chat messages and twitter feeds [17]. Siddiqui and Aalam [17] discussed various challenges in short text clustering, which are: Sparse Feature Vector, Polysemy, Synonymy (Two or more words having the same meaning), and they discussed the other possible solutions which may further improve clustering results. They proposed a framework that can resolve the issues related to Short Text Clustering. This problem was also solved in [18] by introducing a new method for measuring the similarity between short texts. Li and Qu [19] proved in their work that the classical TF-IDF is not adequate for classifying short text. Also, they argued that the improvement version of the TF-IDF algorithm proposed in [10] has noticeable deficiencies and relies heavily on the quality of training data.

Other related studies, discussed the short text sentiment classification [20], [9], [21] and [22]. These studies differ according to the type and nature of the method used in the analysis. Most studies focused on Twitter in addition to other social media datasets such as Facebook.

A study of Badaro, et al. [23] constructed a comprehensive Arabic sentiment lexicon called ArSenL. This lexicon was built based on the following: SAMA (Standard Arabic Morphological Analyzer), English WordNet (EWN), Arabic WordNet (AWN), and English sentiWordnet (ESWN). Eskander and Rambow [24] used ArSenL to construct a new lexicon called SLSA (Sentiment Lexicon for Standard Arabic). It was structured by linking sentiWordnet with the lexicon of an Arabic morphological analyzer Aramorph. Both SLSA and ArSenL rely on sentiWordNet. They are similar in terms of calculating the score of words. The SLSA is based on the principle of giving each Arabic entry associated with English gloss an SI score. The scores are assigned using a linking algorithm that links the English gloss of each Arabic entry to a synset from sentiWordnet. In fact, the research work proposed by Eskander and Rambow [24] which focuses on Arabic sentiWordnet is utilized in this study.

Shoukry and Rafea [25] proposed another approach for sentiment lexicon. They analyzed how twitter (dataset) improves a user's experience with the outside world by exposing personal information and their opinions in every aspect of life. Further, they expanded the sentiments of Twitter text by classifying into four classes which are: Dominance, Influence, Submission, and Compliance (DISC). There was also a different study of El-Beltagy and Ali [2]. The two studies are different, El-Beltagy and Ali [2] generated three categories of Arabic lexicons from Twitter, while Shoukry and Rafea [25] have four categories. On the other hand, the work that was done by Mohammad, et al. [26] has different categorizations. Their study was attentive to emoticons, while the study conducted by Shoukry and Rafea [25] was attentive to clean the extracted text by removing punctuations, numbers, common words. Mouthami, et al. [4] proposed another Arabic lexicon derived from a massive amount of Arabic tweets (about two million tweets). They used PMI (Pointwise-Mutual Information) to create a new lexicon.

In terms of twitter datasets, there are several studies reported in the literature that deals with analyzing tweets datasets of political topics [4, 27, 28]. Mouthami, et al. [4] presented a survey made with journalists and media professionals by reporting on a qualitative empirical study based on an online survey with 50 participants. Their study attempted to analyze tweets from the second round of the Brazilian presidential election in 2014. The results of their work indicated that visualization and data analysis tools were still not easily accessible by those professionals.

While Agrawal and Hamling [27] investigated and analyzed tweets to determine how people reacted to the two US presidential candidates, Donald Trump and Hillary Clinton. Even though a total of 4,044,162 tweets mentioned only Donald Trump, and a total of 2,810,051 tweets mentioned only Hillary Clinton (both tweets originated from the US), their study predicted that Clinton winning the election. However, Trump ended up with the victory. A similar study also conducted by Kharpal [28] showed that, based on tweets collected for five days, Donald Trump will not win the election.

Öztürk and Ayvaz [29] conducted a political study related to twitter dataset concerning the Syrian civil war and the refugee crisis. The datasets were written in Turkish and English languages. When comparing Turkish tweets with English tweets, the experimental results revealed that Turkish tweets were conveying more positive sentiments than English tweets. Since, the war near the Turkey border was attracted more to the Turkish speaking community, whereas, the English speaking community appealed the legitimacy of immigration.

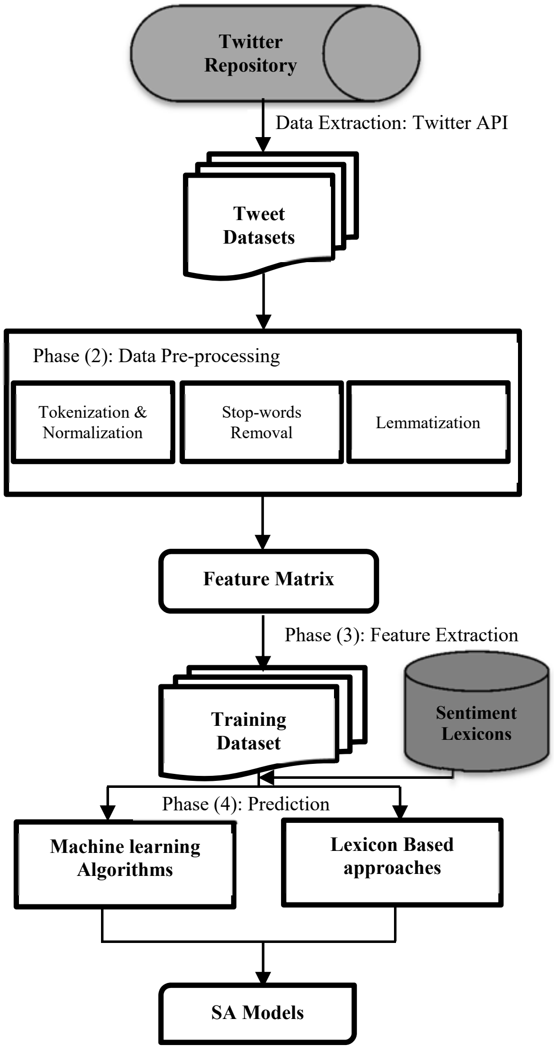

Figure 1: The Proposed Sentiment Analysis Methodology.

Accordingly, the reviewed research has examined the effects of features and methods used for twitter opinion mining, and the impact of the newly proposed methods on Arabic sentiment lexicons. Some of them analyzed a set of combined datasets from Twitter and analyzed positive and negative words. The studies in the literature applied and utilized on several topics on Twitter datasets. These studies are mainly related to the subject chosen for this study, method of analysis and the way to view and solve the problem. The effect of the newly proposed methods on Arabic sentiment lexicons [4].

In this study, we have considered a topic related to the Syrian civil war and consequent crises. This topic contains a dataset of thousands of Arabic tweets, which have not covered before. The present work is designed to be the first to consider what are the problems facing the short texts on Twitter and compare it with the difficulties faced by long texts. In this paper, a machine learning approach (corpus-based approach) and lexicon-based approaches are utilized and compared. Finally, both approaches are compared and evaluated in terms of performance based on different feature selection approaches.

3 RESEARCH METHODOLOGY

The proposed methodology of this research work composed of collecting tweets datasets and performing sentiment analysis. The dataset was compiled by collecting tweets from Twitter concerning the Syrian civil war and consequent crises. In this Section, machine learning and lexicon-based approaches are utilized and the results are compared. The proposed sentiment analysis methodology consists of four main stages, as shown in Figure 1. The methodology starts with data preprocessing and ends with generating SA models. A detailed description of each stage is explained in the following sections.

A. Data Extraction and Pre-processing

The utilized dataset in this study is based on tweets posted concerning the Syrian crises. The dataset is a short text that contains only a few words because the length of tweets is 140 words or less. The dataset contains 2000 randomly selected and labeled tweets originating from Syria (a country where Arabic is the native language). The tweets dataset was collected in 2014 by polling the Twitter API [30]. Salameh, et al. [30] manually annotated the datasets for sentiments (positive, negative or neutral) and were also provided with the confidence of the annotation calculated by CrowdFlower [26, 30]. Figure 2 shows examples of tweets including an opinion as seen in the data set. Table 1 illustrates the statistical summary of the utilized datasets.

Table 1: Statistical Summary for the Utilized Dataset.

Details	Positive tweets	Negative tweets
Total tweets	1000	1000
Total distinct Words	2431	2565
Average words in each tweet	17.21	18.54

Positive	اللهم إني أستودعك أطفال ونساء سوريا فأحفظهم بحفظك ياأرحم الراحمين Oh God, I entrust you with the children and women of Syria
Neutral	سوريا المعادل الراهن للملهاة الحزينة بكامل فصولها Syria, The current equation of futility with all of its episodes
Negative	حدث في مثل هذا اليوم #مجزره_الساعه #حمص happened on this day, the massacre of Homs occurred #Homs

Figure 2: Samples of Tweets In The utilized Dataset.

The first stage in the proposed methodology is to pre-process the Arabic tweets dataset. Text pre-processing is a crucial phase for sentiment analysis and must be applied to the raw text data. Text pre-processing includes Tokenization and Normalization, filtering and stop words removal and finally, lemmatization. Then the data is converted into a vector of features where each feature represents a word in the utilized dataset.

In this step, removing white spaces and punctuations are performed. Then, the tokenization procedure is done by splitting text into a collection of words where each word represents a single term in the corpus. After that, stop-words removals are applied by removing pronouns, adverbs, conjunctions, prepositions, and other constructive terms. These terms usually do not hold any semantic meaning and are not important to classify text data. Terms like “ال” (the), “على” (on), “هناك” (there) and many others are considered as stop-words and are not important for sentiment analysis. Thus, all stop-words are discarded based on a list of Arabic stop-words obtained from the Khoja stemmer tool [31]. As the Arabic language has different morphology, the normalization process is applied to unify words typed differently. The normalization is done by converting a set of characters to their standard form. This step includes transforming the letters (ى) to (ي), the letters (أ, آ, and إ) to (ا) and (ة) to (ه). Lastly, Lemmatization is applied, which is the process of removing inflectional forms of a word so they can be analyzed as a single term.

After applying the aforementioned pre-processing steps, all processed terms (words) are called bag-of-words (BOW) and mainly utilized to build SA models as illustrated in the next section.

B. Sentiment Prediction Using Sentiment Lexicons

People usually express their opinions through their native language and the expression carries informative words that can be extracted. The information in an opinion (expression) can be mined through a target dictionary. Online dictionaries have become widely available for almost all languages. For example, SentiWordNet, a publicly accessible lexical tool, is extensively used in sentiment analysis and opinion mining [32, 33]. This tool is based on the English lexical database called WordNet [34]. SentiWordNet is constructed through assigning to each word (synset) of WordNet three sentiment scores: positivity, negativity, objectivity. They can easily extract opinions from texts and it is a publicly available and accessible sentiment measuring tool used in sentiment classification and opinion mining [33]. There are different versions of sentiWordnet. The latest version is sentiWordnet 3.0 [32]. SentiWordnet expands on WordNet, which is a database of words organized according to their semantic relation to each other.

In this work, we have used Arabic Emoticon Lexicon [26] for SA. The goal is to classify each Arabic tweet (short text), expressed as bag-of-words (BOW), into two distinct sentiment labels; positive or negative. The core step to predict the sentiment label for a given Arabic tweet is the lookup process, where each term included in the tweet is looked up in the Arabic Lexicon. If a word is not included in the lexicon it will be considered as a neutral word with zero sentiment intensity. The overall sentiment score for a tweet is then computed as described in Algorithm 1 below. The Sentiment Label of a tweet text is categorized according to the overall sentiment score. Accordingly, the Sentiment Label Set δ is categorized as positive or negative; where: positive indicates a positive polarity and negative indicates negative polarity.

C. Feature Selection

The performance of machine learning algorithms relies heavily on the size of the utilized feature set in a dataset. Consequently, it is essential to select a smaller subset of features from the original dataset to increase the performance of utilized machine learning algorithms. This step is very significant to select a list of features that have the most discriminative classification power using appropriate feature selection methods.

Feature selection methods have been widely used and utilized in the literature, such as Mutual Information (MI), Information Gain (IG), Document Frequency (DF) and Term Frequency (TF). Based on the literature, using different feature selection methods has not much difference in the performance of constructed models, if the case is to select sets of features with different sizes [35, 36]. For experimental purposes, Information Gain measure is utilized as a feature selection method. To build sentiment analysis models, the set of feature after the pre-processing stage are ranked and the top 'N' scoring features are selected based on ranking results.

Algorithm 1: Tweet Sentiment Identification Using Sentiment Lexicon

```
INPUT: Arabic_Emoticon_Lexicon dictionary
           {T, BOW}
OUTPUT: Set of Sentiment Labels $\delta$ = {Pos, Ng}
PosCount=  Number of words having Positive
              Sentiment Intensity
NegCount= Number of words having Negative
              Sentiment Intensity
PosScore =  The accumulated Positive
              Sentiment Intensities for each tweet
NegScore =  The accumulated Positive Sentiment
              Intensities for each tweet

for all $\tau i \in T$  do
    if PosScore > NegScore then
        $\tau i$= Positive
    else if NegScore > PosScore then
        $\tau i$ = Negative
    else [PosScore = NegScore]
        if PosCount > NegCount then
            $\tau i$ = Positive
        else if NegCount > PosCount then
            $\tau i$ = Negative
    end if
end for
```

D. Machine Learning Algorithms

For the task of sentiment analysis, different machine learning algorithms for predicting the subjectivity of tweets are utilized. The utilized algorithms are selected to handle such as textual data [37, 38]. In this research work, machine learning algorithms, including Decision Tree (DT), Support Vector Machine (SVM), Simple Logistic, Voting-based and k-NN algorithms are used to classify the subjectivity of Arabic tweets. The employed DT algorithms, in this work, are Logistic Model Trees (LMT) [39] and Radom Forest (RF) [40]. Further, an implementation of the SVM algorithm called Sequential Minimal Optimization (SMO) has been utilized in this study [41]. To sum up, the machine learning algorithms, specifically: SVM, RF, LMT, Simple logistic, k-NN and vote-based are mainly employed and evaluated on the utilized Arabic tweets dataset. The experimental evaluations of the employed machine learning algorithms are carried out using the Weka software [42]. The evaluation is performed by employing different parameters setting and the best parameter settings are selected based on best-reported performance evaluation results.

E. Performance validation

In order to validate the performance of the employed machine learning algorithms, the evaluation measures must be assessed on a separate dataset called the test set. The *k*-fold cross-validation approach is mainly used. This approach is used to generate a different test set using a resampling procedure and prevent the over-fitting problem. The procedure of the *k*-fold cross-validation approach is performed through generating an independent test from the original dataset without the need for a separate test dataset. It splits the original dataset into *k* groups, each single group is used once as a test data set (validation) and the remaining groups are used as a training dataset. The SA model is built on the training set and evaluated on the test dataset. Then, the evaluation result for that fold is retained. This procedure is repeated to all generated folds and the evaluation results are finally averaged. In this work, the 10-fold cross-validation is mainly used to validate the performance of generated SA models.

In terms of performance evaluation measures, the accuracy, F-measure and Area Under the Curve (AUC) measures are mainly utilized in this study. Each measure provides a different evaluation perspective and a broader set of performance results to compare.

The accuracy metric measures how accurately a SA model predicts sentiment class. It is computed as the percentage of true positive and true negative rates to the number of all instances. While the F1 metric consolidates the precision and the recall measures into a single measure. F1 is calculated as follows:

$$\text{F1} = 2 * \frac{precision \, * \, recall}{(precision \, + \, recall)}$$

The value of the F1 metric ranges between 1 (perfect precision and recall) and 0 (worst value). The higher the F1 value, the more efficient the SA model is.

Another metric to assess the effectiveness of SA models is the Receiver Operating Characteristic (ROC). This metric compares the true positive rate with a false positive rate as a drawn curve. The ROC is generally summarized as a statistical value representing the area under the ROC curve known as

Area Under Curve (AUC). The AUC represents the possibility that the output of the SA model is better than induction using a random model, where a random model has an AUC value of 0.5, while a perfect model has an AUC of 1.

4 EXPERIMENTAL RESULTS AND DISCUSSION

The goal of the proposed Sentiment Analysis methodology is to compare and evaluate the performance of two SA approaches, namely: machine learning approaches and lexicon-based approach. To validate the performance of the proposed methodology, the performance of the employed machine learning algorithms is examined on the utilized tweets dataset. These algorithms are SVM, RF, LMT, Simple logistic, k-NN and vote-based. The lexicon-based approach using the Arabic Emoticon Lexicon is also employed as a baseline approach to be compared with machine learning algorithms. The results of employed machine learning algorithms are then analyzed in terms of utilized performance evaluation measures. Subsequently, the generated models were empirically examined and the best SA models were selected based on best performance results.

The IG feature selection method was employed to the obtained datasets after the pre-processing step. Then, the feature set is ranked based on ranking results. The top 25, 50, 75, 100, 125, 150, 200 and 300 selected features are used to generate a set of eight new datasets. The generated datasets are then utilized and analyzed in terms of selected performance measures and best SA models were reported. All the machine learning experiments with different parameters setting were carried out using the open-source Weka software [42]. Finally, the best models were reported based on the best parameter setting.

4.1 Machine learning approaches results

Table 2 shows the results of different machine learning algorithms with regards to the tweets dataset on the generated feature set after the feature extraction step. As shown in Table 2, the reported accuracy of different machine learning algorithms varies in the range of 69.12% and 84.19%. However, the reported results are biased toward the class of a large number of instances, as the utilized datasets are imbalanced. Nevertheless, the reported performance results based on the F-measure and AUC are quite promising as these measures take into consideration the class with a small number of instances. As shown in Table 2, the F-measure performance results of different algorithms vary in the range of 0.80%-0.91%and AUC results between 0.65-0.84. These measures are not biased towards the class label with a large number of instances. Based on the F-measure performance results, the LMT and RF algorithms reported the best performance results, followed by SVM, Simple logistic and Vote. While the KNN algorithm showed the lowest performance.

Table 2: Performance Results Of Machine learning Algorithms on Syrian Tweets Datasets

Classifier	Accuracy	F-measure	AUC
KNN-3	*69.12*	*0.8*	*0.66*
KNN-5	*69.68*	*0.8*	*0.67*
DT - RF	*84.19*	*0.91*	*0.84*
DT:LMT	*81.98*	*0.9*	*0.71*
Vote	*75.62*	*0.84*	*0.69*
SVM	*84.11*	*0.9*	*0.65*
Simple-logistics	*82.09*	*0.86*	*0.69*

In regards to the accuracy, the performance results of various machine learning algorithms based on the employed feature selection method are reported in Figure 3. The performance results are stated for the newly generated datasets with a different number of selected terms based on the IG method (top 25 to top 300 terms). As shown in Figure 3, when applying the IG feature selection method, the performance results of SA models of newly generated datasets, are improved as compared to the results of the SA models based on the original dataset without employing the IG feature selection. Thus, the newly generated datasets, according to the IG feature selection, can outstandingly distinguish the subjectivity of tweets.

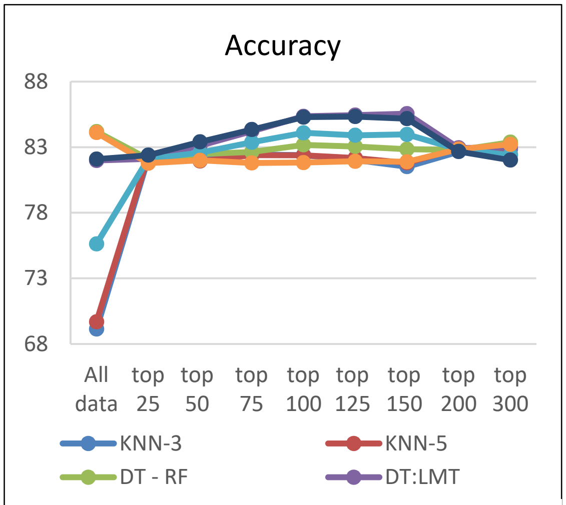

Figure 3: Accuracy of Different Machine Learning Algorithms.

As shown in Figure 3, the accuracy of the generated SA model reaches 85.55% when the number of features chosen is 150. When the IG feature selection is not applied, the reported accuracy reaches 84.19%. As shown in Figure 5, the best AUC reported result is 86% when the number of selected terms is 150 features, as compared to 84% on the original dataset. Furthermore, as shown in Figure 4, the F1 performance results reached 92% when the number of selected terms is 150, as compared to 91% (the best F-measure results obtained where IG feature selection is not employed).

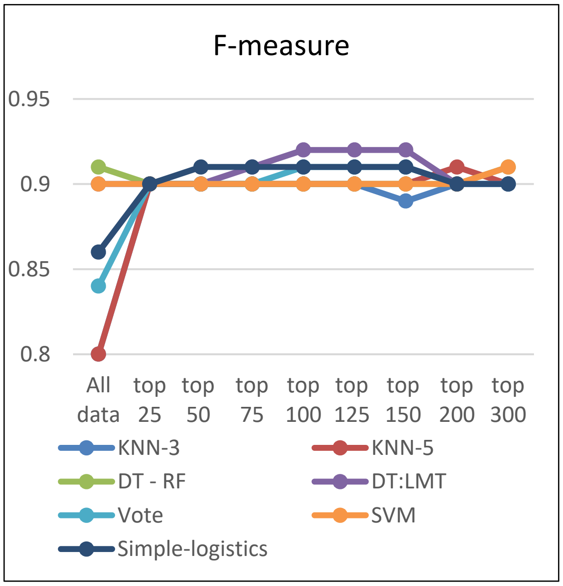

Figure 4: The F-Measure Performance Results of employed machine learning algorithms.

Furthermore, as shown in Figure 3, the performance results exposed that the accuracy results are stabilized as the number of selected terms is 100 or more. The results of F-Measure and AUC of different machine learning algorithms are shown in Figures 4 and Figure 5, respectively. As both figures show, the performance results of employed machine learning algorithms based on the original dataset, with no feature selection, are consistently lower than the results when the feature selection method is utilized. The performance results provide comparable and enhanced results depending on the number of terms considered to build the SA models. The performance results of employing different machine learning algorithms confirm that the best performance results are reported when the number of selected terms is 100. The employed feature selection method demonstrates that the subjectivity of tweets can be predicted with magnificent results, as there are enhancements in performance results than the results generated based on the original tweets dataset. Similar results are also observed with the AUC measure, as shown in Figure 5. As Table 3 and Figure 5 show, the best performance results were obtained using SVM and LMT algorithms. The LMT algorithm achieves stable and best F-measure results (0.92). However, other machine learning algorithms report varied and similar performances.

Furthermore, Table 3 shows similar performance results. When the number of selected terms is more than 100, both F1 measure and AUC achieve better results as compared to the performance results when the number of selected terms are less than 100. As a result, the generated SA models perform the best in terms of predicting the subjectivity of tweets, when the number of terms chosen is between 100 and 150.

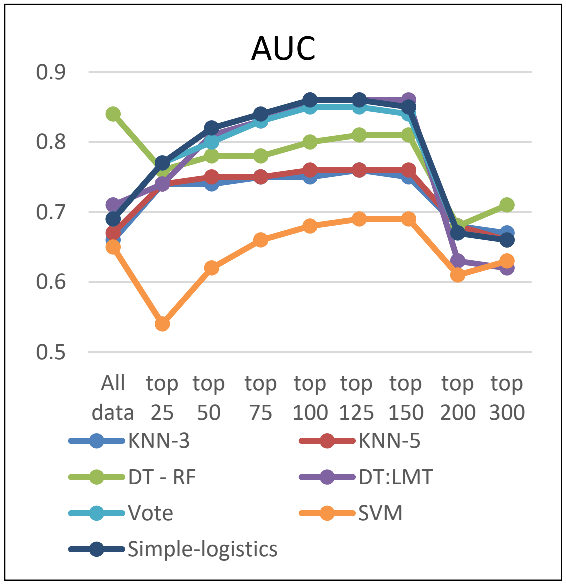

Figure 5: The AUC Performance Results of Employed Machine Learning Algorithms.

Overall, as shown in Table 3, the SA model generated by the LMT algorithm showed the best performance results in terms of accuracy, F-measure and AUC where the maximum values of F-measure and AUC are 0.86 and 0.92, respectively. It is further found that the simple logistic algorithm achieved similar results to the LMT algorithm in terms of AUC and F-measure. The minimum values of AUC and F-measure for the SA model generated by the simple logistic algorithm are 0.66 and 0.86, respectively, whereas, the maximum values of AUC and F-measure are 0.86 and 0.91, respectively. Also, it can be shown in Table 3 that the Vote model has the same behavior, but it is, in certain respects, less than the results showed for the SA models generated by the LMT and simple logistic algorithms.

Table 3: Performance Results of employed ML Algorithms on Datasets With a different number of selected Terms Based on Information Gain method.

	Top 25			*Top 50*			*Top 75*		
	Acc	*F1*	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*	*AUC*
KNN-3	82.17	0.9	0.74	82.07	0.9	0.74	82.45	0.9	0.75
KNN-5	82.34	0.9	0.74	81.94	0.9	0.75	82.38	0.9	0.75
DT: RF	82.1	0.9	0.76	82.47	0.9	0.78	82.63	0.9	0.78
DT: LMT	82.1	0.9	0.74	83.12	0.9	0.81	84.22	0.91	0.83
Vote	82.1	0.9	0.77	82.58	0.9	0.8	83.37	0.9	0.83
SVM	81.77	0.9	0.54	81.99	0.9	0.62	81.79	0.9	0.66
Simple-logistics	82.39	0.9	0.77	83.4	0.91	0.82	84.34	0.91	0.84

	Top 100			*Top 125*			*Top 150*		
	Acc	*F1*	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*	*AUC*
KNN-3	82.39	0.9	0.75	82.02	0.9	0.76	81.51	0.89	0.75
KNN-5	82.37	0.9	0.76	82.17	0.9	0.76	81.8	0.9	0.76
DT: RF	83.17	0.9	0.8	83.05	0.9	0.81	82.84	0.9	0.81
DT: LMT	85.35	0.92	0.86	85.44	0.92	0.86	85.55	0.92	0.86
Vote	84.09	0.91	0.85	83.9	0.91	0.85	83.97	0.91	0.84
SVM	81.82	0.9	0.68	81.92	0.9	0.69	81.88	0.9	0.69
Simple-logistics	85.29	0.91	0.86	85.33	0.91	0.86	85.17	0.91	0.85

	Top 200			*Top 300*			*All feature set*		
	Acc	*F1*	*AUC*	*Acc*	*F1*	*AUC*	*Acc*	*F1*	*AUC*
KNN-3	82.66	0.9	0.68	82.84	0.9	0.67	69.12	0.8	0.66
KNN-5	82.94	0.91	0.68	82.44	0.9	0.66	69.68	0.8	0.67
DT: RF	82.79	0.9	0.68	83.38	0.91	0.71	84.19	0.91	0.84
DT: LMT	82.94	0.9	0.63	82.96	0.9	0.62	81.98	0.9	0.71
Vote	82.83	0.9	0.67	82.41	0.9	0.66	75.62	0.84	0.69
SVM	82.85	0.9	0.61	83.21	0.91	0.63	84.11	0.9	0.65
Simple-logistics	82.67	0.9	0.67	82.01	0.9	0.66	82.09	0.86	0.69

Table 3 also shows the reported result of the SA model generated by the SVM. The maximum AUC and F-measure values of the SVM are 0.85 and 0.91, respectively. Nevertheless, other algorithms report divergent accuracy performance results. According to the experimental results, it can be concluded that the SA model generated by the LMT has the best result among other SA models for predicting the subjectivity of tweets under the all reported experimental setups.

The outstanding performance results reported by the LMT algorithm need further scrutiny. However, the results can be associated with the importance of utilizing the feature selection approaches, where the unimportant features are discarding. This has led the LMT algorithm to generate an optimal SA model that is significantly compact and have precise performance results.

4.2 Lexicon-based approach results

Table 4 shows the reported confusion matrix that is generated when employing the Arabic Emoticon Lexicon dictionary. Noted that we compared the predicted Sentiment Labels (Sentiment Polarity) using the lexicon with their corresponding Actual Sentiment class Labels. Table 5 shows the accuracy, precision, recall, F-Measure, sensitivity and specificity performance results of employing the lexicon-based approach (obtained from the confusion matrix presented in Table 4).

Inspection of the results presented in Table 5 indicates that the prediction of Negative tweets outperformed the prediction of positive tweets (Specificity =0.77 while Sensitivity=0.26). The Accuracy was (0.68), which is comparable with other research work described in the literature. By comparing the results obtained by machine learning approaches with the lexicon-based approach, it is clear that the machine learning approach outperformed the lexicon-based one.

Table 4: The Confusion Matrix for the Predicted Sentiment Labels vs. Actual Labels.

		Predicted		
		Positive	Negative	**Total**
Actual	Positive	74	310	384
	Negative	211	1013	1224
	Total	285	1323	1608

The highest accuracy achieved by classifiers was (0.86) while the accuracy achieved by the lexicon-based approach was (0.68). The results also found that using other performance measures including F1 measure is also superior.

Table 5: The Performance Results of the Lexicon-Based Approach.

Performance measure	Result
Precision	0.19
Recall	0.26
F-Measure	0.22
Sensitivity	0.26
Specificity	0.77
Accuracy	0.68
Error Rate	0.32

5 CONCLUSION AND FUTURE WORK

Arabic sentiment analysis research has increased noticeably in the last decade. In this paper, we analyzed Arabic short text tweets datasets using both sentiment analysis approaches, namely lexicon-based and Machine learning approaches. The differences between short and long Arabic text sentiment analysis were also outlined. One of the objectives of this paper is to compare the performance of lexicon-based and machine learning algorithms in classifying the subjectivity of tweets. Consequently, a methodology to SA was proposed and evaluated by comparing the performance of both SA approaches. Five popular machine learning algorithms were mainly utilized, namely, Logistic Model Trees, Random Forest, Support Vector Machine, Simple logistic, Voting-based and k-NN. The performances of the employed machine learning algorithms were evaluated in terms of utilized performance evaluation measures. Furthermore, another goal of this paper is to utilize feature selection approaches to reduce the number of feature sets, which will eventually enhance the prediction performance of the generated SA models. The comparison of two main approaches (i.e. Lexicon and machine learning approaches) was useful in predicting the sentiment labels for short text tweets and showed different performance results. Machine learning results show that the Logistic Model Trees (LMT) classifier achieves the highest performance accuracy, followed by the Simple logistic algorithm and the SVM algorithms, respectively. The results also show that there are enhancements in performance results when utilizing feature selection approaches.

The performance results, when utilizing feature selection, showed that the best performance results were achieved by the LMT algorithm (*Acc=85.55, AUC=0.86, F-measure= 0.92*). The ML approaches outperform the lexicon-based approach in the context of predicting tweets. The results showed that the lexicon-based approach was less effective at predicting positive attitudes (Specificity =0.77 while Sensitivity=0.26) for unclear reasons, thus providing an interesting avenue for further investigations.

As future work, we aim to compare the results obtained from utilized machine learning algorithms with other algorithms to enhance the performance of predicting subjectivity of Arabic tweets. Further, we plan to apply a more in-depth linguistic analysis for the Arabic language and compare it with pre-processing steps utilized in this research work.

6 REFERENCES

[1] T. Hamshere, *Getting Started with Twitter Flight*: Packt Publishing Ltd, 2013.

[2] S. R. El-Beltagy and A. Ali, "Open issues in the sentiment analysis of Arabic social media: A case study," in *2013 9th International Conference on Innovations in Information Technology (IIT)*, 2013, pp. 215-220.

[3] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Computer Speech & Language*, vol. 28, pp. 20-37, 2014.

[4] K. Mouthami, K. N. Devi, and V. M. Bhaskaran, "Sentiment analysis and classification based on textual reviews," in *2013 international conference on Information communication and embedded systems (ICICES)*, 2013, pp. 271-276.

[5] N. Ahmad and J. Siddique, "Personality assessment using Twitter tweets," *Procedia computer science*, vol. 112, pp. 1964-1973, 2017.

[6] Z. Salah, A.-R. F. Al-Ghuwairi, A. Baarah, A. Aloqaily, B. a. Qadoumi, M. Alhayek, *et al.*, "A systematic review on opinion mining and sentiment analysis in social media," *International Journal of Business Information Systems*, vol. 31, pp. 530-554, 2019.

[7] N. Al-Twairesh, H. Al-Khalifa, and A. Al-Salman, "Subjectivity and sentiment analysis of Arabic: trends and challenges," in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*, 2014, pp. 148-155.

[8] A. Assiri, A. Emam, and H. Al-Dossari, "Towards enhancement of a lexicon-based approach for Saudi dialect sentiment analysis," *Journal of Information Science*, vol. 44, pp. 184-202, 2018.

[9] H. Saif, Y. He, and H. Alani, "Semantic sentiment analysis of twitter," in *International semantic web conference*, 2012, pp. 508-524.

[10] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," *arXiv preprint arXiv:1308.6242*, 2013.

[11] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar, "Opinion mining and analysis for Arabic language," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 5, pp. 181-195, 2014.

[12] M. Rushdi-Saleh, M. T. Martín-Valdivia, L. A. Ureña-López, and J. M. Perea-Ortega, "Bilingual experiments with an arabic-english corpus for opinion mining," in *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 2011, pp. 740-745.

[13] F. Salem, "Social media and the internet of things towards data-driven policymaking in the Arab world: potential, limits and concerns," *The Arab Social Media Report, Dubai: MBR School of Government*, vol. 7, 2017.

[14] S. Al-Azani and E.-S. M. El-Alfy, "Using word embedding and ensemble learning for highly imbalanced data sentiment analysis in short arabic text," *Procedia Computer Science*, vol. 109, pp. 359-366, 2017.

[15] D. Zhao, N. Du, and L. Qin, "Study on Short Text Classification with Integrated Algorithm," in *2016 13th Web Information Systems and Applications Conference (WISA)*, 2016, pp. 121-124.

[16] C. Dos Santos and M. Gatti, "Deep convolutional neural networks for sentiment analysis of short texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69-78.

[17] T. Siddiqui and P. Aalam, "Short Text Clustering; Challenges & Solutions: A Literature Review," *International Journal Of Mathematics And Computer Research*, vol. 3, pp. 1025-1031, 2015.

[18] M. Sahami and T. D. Heilman, "A web-based kernel function for measuring the similarity of short text snippets," in *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 377-386.

[19] L. Li and S. Qu, "Short text classification based on improved ITC," *Journal of Computer and Communications*, vol. 1, p. 22, 2013.

[20] A. Sarlan, C. Nadam, and S. Basri, "Twitter sentiment analysis," in *Proceedings of the 6th International Conference on Information Technology and Multimedia*, 2014, pp. 212-216.

[21] E. Kouloumpis, T. Wilson, and J. Moore, "Twitter sentiment analysis: The good the bad and the omg!," in *Fifth International AAAI conference on weblogs and social media*, 2011.

[22] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 1031-1040.

[23] G. Badaro, R. Baly, H. Hajj, N. Habash, and W. El-Hajj, "A large scale Arabic sentiment lexicon for Arabic opinion mining," in *Proceedings of the EMNLP 2014 workshop on arabic natural language processing (ANLP)*, 2014, pp. 165-173.

[24] R. Eskander and O. Rambow, "SLSA: A sentiment lexicon for Standard Arabic," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2545-2550.

[25] A. Shoukry and A. Rafea, "Sentence-level Arabic sentiment analysis," in *2012 International Conference on Collaboration Technologies and Systems (CTS)*, 2012, pp. 546-550.

[26] S. M. Mohammad, M. Salameh, and S. Kiritchenko, "How translation alters sentiment," *Journal of Artificial Intelligence Research*, vol. 55, pp. 95-130, 2016.

[27] A. Agrawal and T. Hamling, "Sentiment analysis of tweets to gain insights into the 2016 US election," *Columbia Undergraduate Science Journal*, vol. 11, 2019.

[28] A. Kharpal, "Trump will win the election and is more popular than Obama in 2008, AI system finds," *CNBC, October,* vol. 28, 2016.

[29] N. Öztürk and S. Ayvaz, "Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis," *Telematics and Informatics,* vol. 35, pp. 136-147, 2018.

[30] M. Salameh, S. Mohammad, and S. Kiritchenko, "Sentiment after translation: A case-study on arabic social media posts," in *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies*, 2015, pp. 767-777.

[31] S. Khoja and S. Garside, "Stemming Arabic Text. Technical report," *Computing department, Lancaster University, UK, http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps,* 1999.

[32] S. Baccianella, A. Esuli, and F. Sebastiani, "Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining," in *Lrec*, 2010, pp. 2200-2204.

[33] N. Al-Twairesh, H. Al-Khalifa, A. Al-Salman, and Y. Al-Ohali, "Arasenti-tweet: A corpus for Arabic sentiment analysis of saudi tweets," *Procedia Computer Science,* vol. 117, pp. 63-72, 2017.

[34] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM,* vol. 38, pp. 39-41, 1995.

[35] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 488-495.

[36] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *Icml*, 1997, p. 35.

[37] S. M. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text mining: predictive methods for analyzing unstructured information*: Springer Science & Business Media, 2010.

[38] A. Baarah, A. Aloqaily, Z. Salah, M. Zamzeer, and M. Sallam, "Machine Learning Approaches for Predicting the Severity Level of Software Bug Reports in Closed Source Projects," *Machine Learning,* vol. 10, 2019.

[39] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," *Machine learning,* vol. 59, pp. 161-205, 2005.

[40] L. Breiman, "Random forests," *Machine learning,* vol. 45, pp. 5-32, 2001.

[41] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization, advances in kernel methods," *Support Vector Learning,* pp. 185-208, 1999.

[42] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann, 2016.