Improving Arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques

Ahmed Al-Azawei
University of Babylon

Available online at: https://ijact.in

Date of Submission	11/05/2019
Date of Acceptance	15/06/2019
Date of Publication	30/06/2019
Page numbers	3150-3157(8 Pages)

Cite This Paper: EK Al-Yasiri, Ahmed Al-Azawei.(2019) Improving Arabic sentiment analysis on social media: A comparative study on applying different pre-processing techniques, 8(6), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3150-3157.

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

IMPROVING ARABIC SENTIMENT ANALYSIS ON SOCIAL MEDIA: A COMPARATIVE STUDY ON APPLYING DIFFERENT PRE-PROCESSING TECHNIQUES

Essam Kazem Al-Yasiri[1], Ahmed Al-Azawei[2]

[1,2]Department of Software, College of Information Technology
[1,2]University of Babylon, Iraq

Abstract: Regardless of the clear growth of Arabic texts on social networking sites (SNSs), it is still difficult to understand or summarize users' opinions or perspectives on a specific topic. Accordingly, Arabic text classification is one of the most challenging topics. This is because of several issues related to the nature of the Arabic language and words that have different variation in meaning. In this paper, after tokenizing the Arabic words, we investigate the role of several pre-processing techniques before classifying Arabic text into different categories. Arabic words were converted into vectors using the term frequency-inverse document frequency (TF-IDF) technique. The findings show that applying Linear Support Vector Machine (LSVC) with stop words and without stemming techniques can outperform the application of Decision Tree (DT) and Random Forest (RF) methods. It was found that the effectiveness of the proposed LSVC is 99.37%. These outcomes are significant to identify users' opinions on SNSs and can have many implications on political, social, economic, and business sectors.

Keywords: Social Networking Sites; Arabic sentiment analysis; Pre-processing techniques; Classifying Arabic text; Data mining algorithms.

I. INTRODUCTION

In recent years, users of the Internet and social networking sites (SNSs) have widely expressed their opinions and comments online. This is due to the interactive features of Web 2.0 technology. It allows for the Internet users sharing their thoughts and views through several sites such as Twitter, WhatsApp, Facebook, forums, blogs, etc. These technologies resulted in a huge quantity of raw data of which creative data mining methods are needed to extract knowledge[1].

Currently, SNSs have played a critical position in modern life. The Internet offers an opportunity for users to express and share their ideas on particular topics [2]. The impact of SNSs such as Twitter and Facebook has increased in the last few years. SNSs also represent a cornerstone of information and data sharing for the great number of people.

Despite different SNSs developed in contemporary life, Twitter has been widely used which offers a more fertile site for sentiment analysis. This is because tweets can be fetched using particular keywords to obtain data on a specific subject. As such, positive, negative, and neutral opinions can be extracted from the Twitter application program interface (API). Hence, Twitter is an efficient source for researchers and organizations as they can extract and analyze people's perspectives and feelings [3]. However, obtaining accurate sentiment analysis using Twitter or other similar micro-blogs still faces several challenges.

The typical short length and irregular structure of such content are some issues that need to be addressed. Such problems might be appearing more in Arabic tweets due to the complex structure of the Arabic language. According to Jamal, Keohane, Romney, and Tingley[4],approximately 40 percent of the Arab public had Internet access and approximately 30 percent of them had a Twitter account. This means that Arab Twitter users represent a large part of its overall participants.

To elaborate on the above, the pre-processing of textual data is a key step in sentiment analysis. It is shown in previous literature that the analysis accuracy can be enhanced if a suitable approach is chosen in the pre-processing stage[5].Considerable attention should be given for pre-processing because it could have a positive or negative influence on the analysis of any text document. In this way, it is expected that applying different pre-processing techniques might improve the accuracy of classifying Arabic tweets. This paper integrates different pre-processing techniques such as normalization, stopping words, and stemming beforehand using data mining classification methods. Three well-known machine learning algorithms are implemented for text classification, namely Linear Support Vector Classifier (LSVC), Decision Tree (DT), and Random Forest Classifier (RFC).

II. RELATED WORK

In spite of previous research that has utilized data from SNSs to elicit new information on a specific topic, it is still difficult to obtain accurate results in Arabic sentiment analysis. This is because of the noise with the original Arabic text where a few researchers have conducted studies on the categorization of Arabic texts. Here, some of the related works which focus on applying pre-processing techniques are reviewed.

In [6], stemming was used as a pre-processing stage before classifying Arabic text. Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (J48), and Sequential Minimal Optimization (SMO) were applied. The obtained accuracy of SVM and NBwas87.79% and 88.54% respectively without stemming. The findings suggest that stemming had a negative impact on the algorithm's accuracy where the results with stemming were 84.49% and 86.35% respectively.

Ayedh, Tan, Alwesabi, and Rajeh[7]investigated the influence of pre-processing techniques on the accuracy of Arabic document classification (DC). Three pre-processing techniques were used which are normalization, stop words, and stemming. The classification model was built using k-Nearest Neighbor (kNN), Naive Bayes (NB), and SVM classifiers. It was found that SVM classifier outperformed the other classifiers when stemming and normalization were combined. The accuracy achieved was 96.74%.

In [8], a classification model for sentiment analysis in micro-blogging posts of Twitter was developed. Various pre-processing techniques and feature selection methods were implemented in which the Naïve Bayes (NB) classifier was used. The proposed model achieved reasonably good performance for the training set. The model classified more accurately the positive class in comparison to the negative one. The results showed that the Naïve Bayes algorithm with a minimum of three in the selection of high information features using Chi-Square led to achieving 89% of accuracy.

Comparing the application of different stemmer techniques was carried out in [9]. Naïve Bayes (NB) classifier was used to evaluate the impact of a Khoja stemmer, Root extractor, and Light stemmer approaches on Arabic text documents. The findings show that root extractor was the best stemmer where the algorithm performance reached to 75.4% in comparison to other approaches.

In the research study conducted by Alhaj, Xiang, Zhao, Al-Qantas, Elaziz, Dahou[10],several stemming techniques were implemented, namely Information Science Research Institute (ISRI), Tashaphyne, and ARLStem on Arabic document classification (DC). Naive Bayesian (NB), Support Vector Machine (SVM), and k-Nearest Neighbors (KNN) were used. It was shown that ARLStem outperformed other stemmer techniques. SVM also achieved best classification accuracy after applying the ARLStem stemmer (94.64%).

Sallam, Mousa, and Hussein[11]examined the effect of applying normalization and stemming techniques such as ISRI stemmer, and Tashaphyne stemmer on Arabic document classification. In order to extract features, the Bag of Words was used. The outcomes indicate that the use of normalization and stemming can improve the accuracy to97.50%, whereas it was96.66% without stemming and normalization. Moreover, stemming techniques such as Tashaphyne and ISRI achieved less accuracy (95.83% and 93.06% respectively).

Al-Shargabi, Al-Romimah, and Olayah[12]examined three classifiers in analyzing Arabic texts which are Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO), Naïve Bayesian (NB), and J48. The three algorithms were compared based on their accuracy in identifying Arabic text. The removing of stop words was applied as a pre-processing step. The results for all classifiers were improved after removing stop words. Moreover, SVM with SMO achieved the best classification results (96.08%).

In another study performed by Haddi, Liu, and Shi [5], the role of text pre-processing in sentiment analysis was investigated. The experimental results highlight that the accuracy of sentiment analysis can be improved with Support Vector Machines (SVM) when using an appropriate feature selection and representation.

Based on this discussion, it could be clear that text classification is a challenging area of research. However, the implementation of pre-processing techniques can directly affect the findings. The classification accuracy can also be enhanced if appropriate pre-processing approaches are used. Hence, this present research sought to improve the accuracy of Arabic text analysis by using several pre-processing techniques and classification algorithms.

III. THEORETICAL BACKGROUND

A. *Sentiment Analysis*

Sentiment analysis is also known as opinion analysis/mining (OM). It is one of the natural language processing (NLP) tasks that deals with unstructured text to be classified into a positive, negative, or neutral category [13].Generally, there are two types of sentiment analysis which are machine learning and lexicon-based approaches. The key aim of text analysis is to extract beneficial information for particular goals and to make it easier for researchers to handle unstructured data[14]. Organizations can also utilize sentiment analysis in identifying customers' attitudes towards particular products and/or services[15]. Thus, Twitter could be a significant source of sentiment analysis.

B. *Arabic Language Structure*

Arabic language is a semantic grammar which has a significant variation from other common languages (e.g., English, Spanish, Chinese, and French). Arabic language is the Arab nations native language and the optional language in a number of other countries [16]. As pointed out in[17], over 422 million individuals speak this language. Therefore, it is the fifth most used language in the world. It consists of 28 letters.

The Arabic language is divided into three structures: classical Arabic (CA), colloquial Arabic dialects (CAD), and modern standard Arabic (MSA)[7]. The direction of writing in the Arabic language is from right to left.

C. *Arabic stemming*

In general, the stemming method is one of the most efficient techniques utilized in several applications such as machine learning (ML), natural language processing (NLP), information retrieval (IR), document classification (DC), and machine translation[10].The purpose of stemming is to eliminating the affixes such as prefixes and suffixes from words. This can help reduce the number of words in the feature space, precisely matching stems, and saving memory space and time. For example, the root of the word (منظمات,organizations)is (organize,نظم).

D. *Feature extraction*

Feature extraction is the process of extracting important features or attributes from initial textual data for feeding it in machine learning algorithms[18].In text classification, term frequency-inverse document frequency (TF-IDF) weighting scheme is a standard approach in building a feature vector. TF-IDF is a numerical statistic that reflects how a particular word is significant to a document in the corpus[19].

E. *Text mining techniques*

In the field of data classification, several different techniques were proposed such as Bias and decision tree methods [20]. In spite of many other data mining techniques that have been used in previous literature, here we are presenting the most related methods for this research.

1. Support Vector Machine (SVM)

For a linearly separable of a two-class learning task, Support Vector Classifier (SVC) can be used. It finds a hyperplane which can separate two classes of a given sample with a maximum margin. This margin is capable of providing the best generalization ability. Generalization refers to the fact that not only a classifier has a good classification performance (e.g., accuracy), but it also guarantees high predation[21].Figure1 depicts the optimum SVC hyperplane for a linear case.

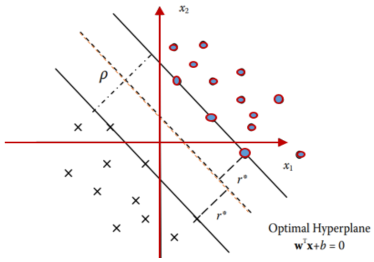

Figure1:The optimum SVC hyperplane for a linear case

2. Decision Tree Classifier (DTC)

One of the foremost well-known techniques of machine learning is the decision tree. It is created based on the decision and leaf nodes [22]. A typical decision tree technique is an effective approach in models' generation [23].The decision tree rule is shown in Equation 1.

$$Entropy(t) = -\sum_{i=0}^{c-1} p(i|t)\, log_2\, p(i|t)(1)$$

Where c is the number of classes, p(i | t) denote the fraction of records belonging to class i at a given node t.

3. Random Forest Classifier (RFC)

Random Forest is a regression and classification technique which can fit many sub-samples of a dataset with a large number of decision tree classifiers[24]. To predict future observations, these large numbers of trees are created from a training set and then to be validated. This method can result in either a categorical or a continuous value output. Random Forest is a strong and reliable technique due to many pros such as: (i) it can be applied for classification and regression tasks, (ii) it can handle missing values and maintain missing data accurately, and (iii) it can handle a large dataset of high dimensionality[25][26]. Figure 2 illustrates the main notion behind this algorithm.

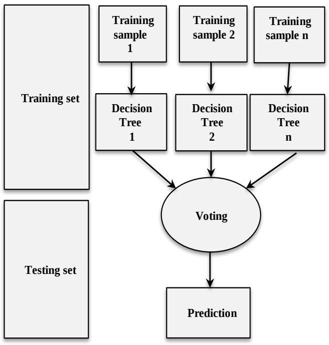

Figure2:TheRandom Forest Algorithm.

IV. RESEARCH METHODOLOGY

The methodology proposed in this research is depicted in Figure 3. In this study, Python Language was used because it provides many libraries and packages compatible with data mining algorithms such as nltk, sklearn, and pandas. Many steps have been carried out to achieve the goals of this study:

- Tweets were retrieved based on a wide variety of keywords. This was applied using Twitter API.
- Sentiment analysis was conducted to classify some texts manually.
- Pre-processing techniques were implemented in the preparation of input data before using the classifier.
- Another stage included building a classifier to categorizing all tweets into either positive or negative.

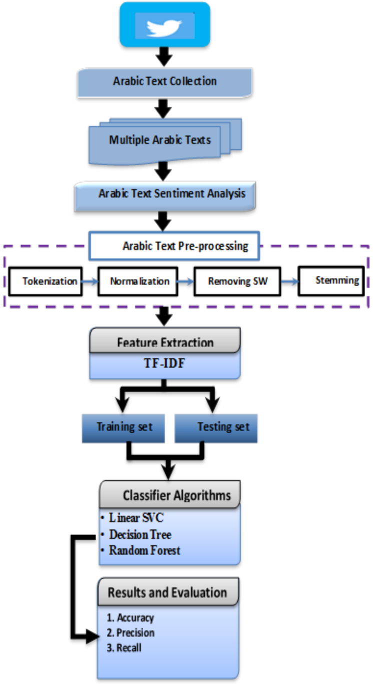

Figure3: The proposed research model

A. *Arabic Text Collection*

The first step in this research was the collection of Arabic tweets from Twitter because there is no available dataset that can be used to achieve the goals of this study. In general, over the period from 23 of November 2018 to 15of April2019, we collected more than 3 million Arabic tweets. This data is the base of our tweet-centered analysis. Then, by using sentiment analysis, some tweets were classified into either positive or negative. This was conducted for approximately 2439 positive and 2323 negative texts. The positive and negative texts dataset comprised of 4762tweets. After shuffling the positive and negative texts, they were divided into a training set (70 %) and testing set (30 %). Figure 4 illustrates the average number of the total positive and negative texts in which negative texts were labeled with 0, whereas positive tweets were labeled with 1.

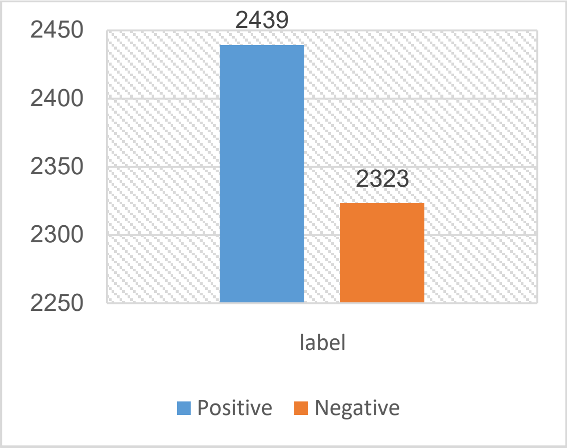

Figure4: The positive and negative texts

B. *Arabic Sentiment Analysis*

Sentiment analysis has been utilized in this study in classifying texts as a positive and negative attitude. This manual data classification was used in the construction of the research model classifier. It is expected that increasing the number of training data can lead to an improvement in the classification accuracy. To analyze the sentiment, the label and the text (texts) were placed in a proper tabular format.

C. *Arabic Text Pre-processing*

After collecting Arabic text, different pre-processing techniques were used before applying the data mining techniques. Various methods of text pre-processing were applied to convert all Arabic text into a form that is suitable for the classification algorithms. This research adopts four steps of text pre-processing which are tokenization, normalization, removing stop words, and text stemming. These steps are described as follows:

1. **Tokenization:** it is a very important step in natural language processing (NLP). Tokenization is the process of splitting the text into smaller

meaningful parts called tokens. In this study, tokenization was performed by dividing the text document into words based on spaces between them. Typically, the python library provides different interfaces for word tokenization.

2. **Normalization:** this involves eliminating irrelevant information such as punctuations(e.g., '!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~'), non - Arabic letters, symbols, numbers, multiple spaces, and special characters. Implementation of such pre-processing techniques resulted in obtaining fitting texts to be used in the proposed classifier.
3. **Removing stop words:** stop words refer to all frequently occurring words that bring no meaning or indication about the content (e.g., conjunctions, pronouns, prepositions,“لذا -so, "من – from, on-على, in- في”).These words are not important in distinguishing various classes. Stop words removal involved the elimination of unimportant words. The list of stop words was prepared and removed from all texts. There are more than 273 Arabic stop words that have been removed from the text.

4. **Arabic Stemming:** words in the Arabic language are derived from root sets that illustrate the fundamental meaning of a word with the added affixes that changing the pronunciation of the word. Accordingly, a root extraction stemmer (Information Science Research Institute (ISRI)) was used in this study. Algorithm 1 summarizes the pre-processing phases.

Algorithm 1: The data pre-processing phase

Input: Multiple Arabic Texts	
Output: A set of tokens for each Arabic text	
step	**Begin**
Step 1	**Read Arabic texts from the collected data**
Step 2	**Applying the tokenization process on Arabic texts to divide them into tokens based on the space.**
Step 3	**Eliminate all non-Arabic letters such as English letters.**
Step 4	**Eliminating all punctuation from the Arabic text, symbols, and special characters including ['!', '"', '#', '$', '%', '&', "'", '(', ')', '*', '+', ',', '-', '.', '/', ':', ';', '<', '=', '>', '?', '@', '[', '\\', ']', '^', '_', '`', '{', '|', '}', '~'].**
Step 4	**Eliminating all elongated and repeated letters from the Arabic text.**
Step 5	**Eliminating all Arabic stop words.**
Step 6	**Eliminating words with a length of fewer than three letters.**
Step 7	**If the word length equal to 3, then return the word without stemming.**
Step 8	**Applying the ISRI stemming algorithm for the Arabic texts to obtain Arabic stemmed text without prefixes and suffixes.**

D. *Feature Extraction*

In text pre-processing, the task of transforming a particular Arabic text into a matrix based on space is important where rows represent the texts and columns represent the features/words. Methods of feature extraction are used to extract features because text data includes formats that are not suitable to be directly used by machine learning techniques. Therefore, the text features are extracted in a specific format that is appropriate for these algorithms. The matrix values are created based on different schemas which are Term Frequency and Term occurrence. One of the most common techniques is TF-IDF. This technique takes into account the number of times that the word appears in all documents and the document set. Once the TF and IDF values are obtained, it will be possible to calculate the TF-IDF. Equation 2 presents how TF-IDF can be calculated.

$$TF - IDF_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right) \qquad (2)$$

Where$tf_{i,j}$ = number of times, the term i appears in a document j, N refers to the total number of documents, anddf_i =number of documents that contain term i.

V. Results and discussion

This study aims at improving the Arabic sentiment analysis on SNSs using different pre-processing techniques. This was achieved by implementing different steps such as data acquisition, Arabic sentiment analysis, data pre-processing, features' extraction, building a classifier, and evaluating the research findings. Generally, this research supports the outcomes of previous work that SNSs represent an important data source for identifying users' opinions and perspectives on a particular topic[27].

The sentiment classification was performed first by applying three data mining techniques which are Linear Support Vector Classifier (LSVC), Decision Tree (DT), and Random Forest Classifier (RFC). The aim of using these different methods is to compare their findings and highlight the best classifier with different pre-processing techniques. The pre-processing and classification results are showed and discussed here.

1. *Results of the pre-processing phase*

Table 1 illustrates the findings of the pre-processing steps. It is noteworthy that all pre-processing steps are presented using the same sentence. The results include removing punctuations, link URLs, re-texts, non-Arabic letters, symbols, numbers, multiple spaces, and special characters. Findings of the pre-processing steps suggest that the implemented steps were effective as these techniques resulted in a noise-free dataset. Such steps are important to achieve better output in the classification process.

Table 1: The results of the pre-processing stage

Pre-processing	Resulted texts
Cleaning Texts	"text": "RT @Lionel__FCB: - تطوير البنية التحتية\n- رفع مستوى التعليم\n- تخفيض نسبة البطالة\n- تحسين مستوى الدخل\n و أمثلة بسيطة توضح معنى التطور و مواكبة الدول المتقدمة…", RT @Lionel__FCB: - Improving infrastructure - Raising the level of education - Reducing the unemployment rate - Improving the level of income are simple examples that explain the meaning of development and keeping pace with the developed countries…
Removing usernames	RT - تطوير البنية التحتية - رفع مستوى التعليم - تخفيض نسبة البطالة - تحسين مستوى الدخل أمثلة بسيطة توضح معنى التطور و مواكبة الدول المتقدمة… RT - Improving infrastructure - Raising the level of education - Reducing the unemployment rate - Improving the level of income are simple examples that explain the meaning of development and keeping pace with the developed countries…
Removing re-tweet	- تطوير البنية التحتية - رفع مستوى التعليم - تخفيض نسبة البطالة - تحسين مستوى الدخل أمثلة بسيطة توضح معنى التطور و مواكبة الدول المتقدمة… - Improving infrastructure - Raising the level of education - Reducing the unemployment rate - Improving the level of income are simple examples that explain the meaning of development and keeping pace with the developed countries…
Removing Punctuation	تطوير البنية التحتية رفع مستوى التعليم تخفيض نسبة البطالة تحسين مستوى الدخل أمثلة بسيطة توضح معنى التطور و مواكبة الدول المتقدمة Improving infrastructure Raising the level of education Reducing the unemployment rate Improving the level of income are simple examples that explain the meaning of development and keeping pace with the developed countries
Removing stop words	تطوير البنية التحتية رفع مستوى التعليم تخفيض نسبة البطالة تحسين مستوى الدخل أمثلة بسيطة توضح معنى التطور مواكبة الدول المتقدمة Improving infrastructure Raising the level of education Reducing the unemployment rate Improving the level of income are simple examples that explain the meaning of development keeping pace with the developed countries
ISRI Stemmer	طور بنة حتي رفع ستى علم خفض نسب بطل تحس ستى دخل مثل بسط وضح عنى تطر وكب دول تقدم

Note: The stemming step cannot be translated into English because this can change the meaning.

2. *Results of the classification methods*

Three methods were used in classifying the texts after extracting the features from the pre-processed dataset. The classification metric of the three implemented algorithm with neither stemming nor stop words are shown in Table 2 and Figure 5. Table 3 and Figure 6, on the other hand, present the research findings with ISRI stemmer and stop words. In addition, the pre-processing of text using stop words and without stemmer is depicted in Table 4 and Figure 7.

Table 2: A comparison of different research findings without ISRI Stemmer and stop words removing

METRIC	MACHINE LEARNING ALGORITHM		
	LINER SVC	DT	RF
ACCURACY	99.23	98.46	98.25
PRECISION	0.99	0.98	0.98
RECALL	0.99	0.98	0.98
F1-SCORE	0.99	0.98	0.98

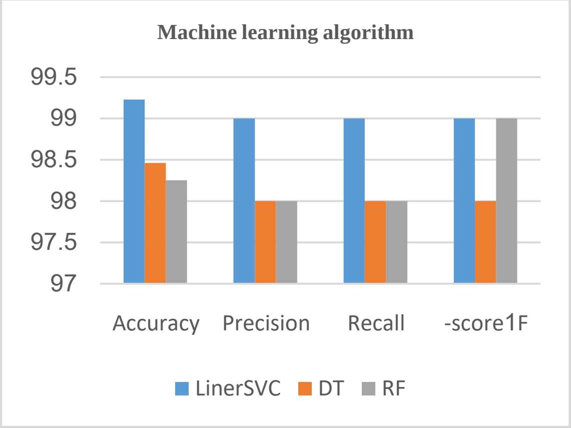

Figure5: A comparison of different findings without ISRI Stemmer and stop words removing

Table 3: A comparison of different research findings with ISRI Stemmer and stop words removing

METRIC	MACHINE LEARNING ALGORITHM		
	LINER SVC	DT	RF
ACCURACY	99.30	98.32	98.74
PRECISION	0.99	0.98	0.99
RECALL	0.99	0.98	0.99
F1-SCORE	0.99	0.98	0.99

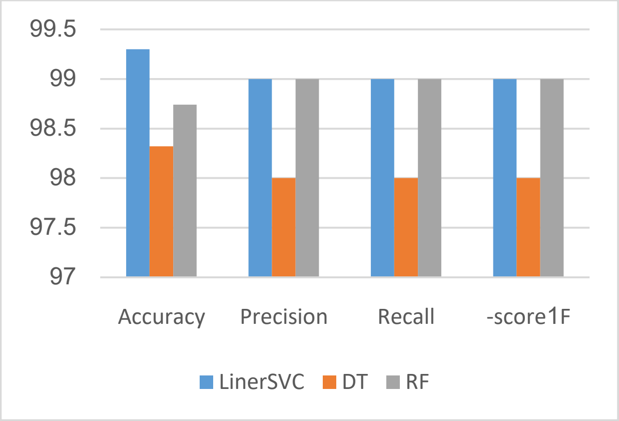

Figure6: A comparison of different findings with ISRI Stemmer and stop words removing

Table 4: A comparison of different research findings with stop words removing and without ISRI Stemmer

Metric	Machine learning algorithm		
	Liner SVC	DT	RF
Accuracy	99.37	97.83	98.53
precision	0.99	0.98	0.99
Recall	0.99	0.98	0.99
F1-score	0.99	0.98	0.99

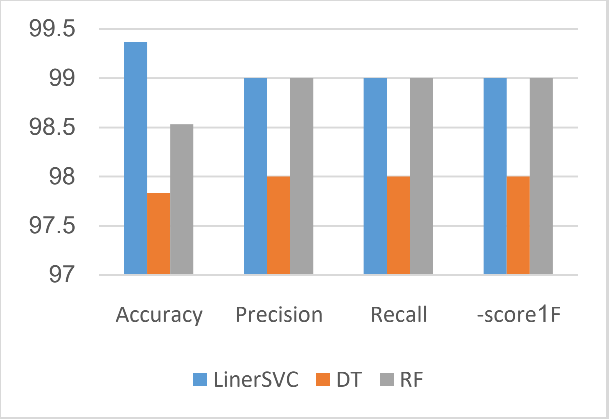

Figure 7: A comparison of different findings with stop wordsremoving and without ISRI Stemmer

The results show that the highest accuracy was achieved by Linear SVC when stemmer is not used(99.37%). This might be because our dataset includes two-classes, namely text, and labels. Based on its key notion, LSVC can accurately separate two sample classes with a maximum margin which, in turn, could lead to high predictive accuracy. In agreement with our research findings, Wahbeh et al. [6] also showed that applying stemmer can negatively affect classification accuracy. Moreover, Al-Shargabi et al.[12] found that using stop words can lead to better results.

This study adds a significant contribution to earlier literature where it bridges a research gap regarding the shortage of empirical work on analyzing users' opinions on SNSs in the Arab world. This can assist in designing efficient recommender systems where such applications can respond to users' individual preferences and needs. This research also sheds some light on the role of pre-processing techniques in enhancing the accuracy of different classification methods.

VI. Conclusion and Future Work

In this research, many pre-processing techniques were implemented on a dataset including Arabic text. It aimed at improving the Arabic sentiment analysis on social networking sites(SNSs) using three data mining algorithms which are Linear Support Vector Classifier (LSVC), Decision Tree (DT), and Random Forest Classifier (RFC). The pre-processing steps were carried out to clean the texts from noise. All algorithms were performed with an accuracy of more than 97%. LSVC outperformed other algorithms with 99.37% accuracy. This result was achieved after removing stop words and without using ISRI Stemmer. For future work, we will attempt to gather texts in different languages. This can highlight the efficiency of our implemented steps on different datasets. Thus, the findings can be generalized. Furthermore, data from other social networks can also be gathered to investigate the Arabic sentiment analysis and highlight users' attitudes towards a particular area of interest.

VII. References

[1] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *2013 IEEE Jordan conference on applied electrical engineering and computing technologies (AEECT)*, 2013, pp. 1–6.

[2] A. Al-Azawei, "Predicting the adoption of social media: An integrated model and empirical study on Facebook usage," *Interdiscip. J. Information, Knowledge, Manag.*, vol. 13, pp. 233–258, 2018.

[3] "The Statistics Portal." [Online]. Available: https://www.statista.com/statistics/303681/twitter-users-worldwide/. [Accessed: 30-Nov-2018].

[4] A. A. Jamal, R. O. Keohane, D. Romney, and D. Tingley, "Anti-Americanism and anti-interventionism in Arabic twitter discourses," *Perspect. Polit.*, vol. 13, no. 1, pp. 55–73, 2015.

[5] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, pp. 26–32, 2013.

[6] A. Wahbeh, M. Al-Kabi, Q. Al-Radaideh, E. Al-Shawakfa, and I. Alsmadi, "The effect of stemming on Arabic text classification: an empirical study," *Int. J. Inf. Retr. Res.*, vol. 1, no. 3, pp. 54–70, 2011.

[7] A. Ayedh, G. Tan, K. Alwesabi, and H. Rajeh, "The effect of preprocessing on Arabic document categorization," *Algorithms*, vol. 9, no. 2, p. 27, 2016.

[8] S. F. Sayeedunnissa, A. R. Hussain, and M. A. Hameed, "Supervised opinion mining of social network data using a bag-of-words approach on the

cloud," in *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)*, 2013, pp. 299–309.

[9] S. A. Yousif, V. W. Samawi, I. Elkabani, and R. Zantout, "Enhancement of Arabic Text Classification Using Semantic Relations with Part of Speech Tagger," *W Trans. Adv. Electr. Comput. Eng.*, pp. 195–201, 2015.

[10] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. A. Elaziz, and A. Dahou, "A Study of the Effects of Stemming Strategies on Arabic Document Classification," *IEEE Access*, 2019.

[11] R. M. Sallam, H. M. Mousa, and M. Hussein, "Improving Arabic text categorization using normalization and stemming techniques," *Int. J. Comput. Appl.*, vol. 135, no. 2, pp. 38–43, 2016.

[12] B. Al-Shargabi, W. Al-Romimah, and F. Olayah, "A comparative study for Arabic text classification algorithms based on stop words elimination," in *Proceedings of the 2011 International Conference on Intelligent Semantic Web-Services and Applications*, 2011, p. 11.

[13] A. M. Alayba, V. Palade, M. England, and R. Iqbal, "Improving sentiment analysis in Arabic using word representation," in *2018 IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition (ASAR)*, 2018, pp. 13–18.

[14] F. S. Gharehchopogh and Z. A. Khalifelu, "Analysis and evaluation of unstructured data: Text mining versus natural language processing," *Int. J. Acad. Res. Comput. Eng.*, no. November, 2011.

[15] C. Aroran and Dr.Rachna, "Sentiment Analysis on Twitter Data," *Int. Res. J. Eng. Technol.*, vol. 14, no. 2, pp. 831–837, 2017.

[16] F. Thabtah, O. Gharaibeh, and R. Al-Zubaidy, "Arabic text mining using rule based classification," *J. Inf. Knowl. Manag.*, vol. 11, no. 01, p. 1250006, 2012.

[17] T. Kanan and E. A. Fox, "Automated Arabic text classification with P-S temmer, machine learning, and a tailored news article taxonomy," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 11, pp. 2667–2683, 2016.

[18] D. Sarkar, *Text analytics with Python: A practical real-world approach to gaining actionable insights from your data*. Apress, New York, 2016.

[19] 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)*, 2016, pp. 1–5.

[20] Z. Li, "A data classification algorithm of internet of things based on neural network," *Int. J. Online Eng.*, vol. 13, no. 09, pp. 28–37, 2017.

[21] X. Wu, and V. Kumar, "The Top Ten Algorithms in Data Mining,"Data Mining and Knowledge Discovery Series, CRC Press, United Statesof America, 2009.

[22] G. Stein, B. Chen, A. S. Wu, and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," in *Proceedings of the 43rd annual Southeast regional conference-Volume 2*, 2005, pp. 136–141.

[23] S. Agarwal, G. N. Pandey, and M. D. Tiwari, "Data mining in education: data classification and decision tree approach," *Int. J. e-Education, e-Business, e-Management e-Learning*, vol. 2, no. 2, p. 140, 2012.

[24] T. A. Wotaifi and E. S. Al-Shamery, "Fuzzy-Filter Feature Selection for Envisioning the Earnings of Higher Education Graduates," *Compusoft*, vol. 7, no. 12, pp. 2969–2975, 2018.

[25] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 2, no. 6, pp. 493–507, 2012.

[26] B. Luo, Q. Zhang, and S. D. Mohanty, "Data-Driven Exploration of Factors Affecting Federal Student Loan Repayment," *arXiv Prepr. arXiv1805.01586*, 2018.

[27] H. M. Habeeb, A. Al-Azawei, and N. Al-A'araji, "Developing a Healthcare Recommender System Using an Enhanced Symptoms-Based Collaborative Filtering Technique," J. Comput. Theor. Nanosci., vol. 16, no. 3, pp. 925–931, 2019.