Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents

Tarek Kanan[1], Amal Aldaaja[1], Bilal Hawashin[2]

[1] Department of Computer Science, Alzaytoonah University of Jordan, Jordan
[2] Department of Computer Information Systems, Alzaytoonah University of Jordan, Jordan
tarek.kanan@zuj.edu.jo, amalmohammad555@yahoo.com, b.hawashin@zuj.edu.jo

Abstract

The social media has provided users with the chance to publish their written and multimedia content and express feelings and emotions about particular subjects via the internet. However, some users have abused these platforms by performing various acts such as Cyber-Bullying and Cyber-Harassment. These phenomena are dangerous and have negative psychological, health, and social effects. Although multiple works have focused on detecting these phenomena on English text, few works studied this phenomenon on Arabic. Moreover, these works used limited number of methods and datasets. Furthermore, there is a lack in Arabic datasets that are concerned with this topic.

We propose the use of Machine Learning to detect such negative written acts. We apply various classification algorithms to the dataset, and we use various Arabic Natural Language Processing (NLP) tools.

To evaluate the performance of the classifiers, we use Recall, Precision, and F1-Measure. The results show that the Random Forest algorithm yields the highest values of F1-Measure. The same results occurred when no stemming and no stop-word removal are applied. However, when separating datasets into Facebook Posts dataset and Twitter Tweets dataset, SVM gives the highest F1-Measure value. Significant tests were conducted to support our results.

Keywords: Social media content, Cyber-Bullying, Cyber-Harrasement, Machine learning, Natural language processing

1 Introduction

Since the introduction of the first social media platform on the internet, the social media has been increasing and spreading in all societies, including Arab societies, where they have become a rich source in expressing opinions, emotions, and sentiments. Facebook and Twitter are two of the most social media widely used applications in the Arab world. Facebook and Twitter environment allow bullies to bully and harass vulnerable groups because there is no censorship from government agencies and lack of awareness through social media. Do not forget also the strong impact on the victims where it is very easy to transfer and spread bullying and harassment to reach as many users as possible, making the situation worse. To make Arab social media a safe environment for all groups of society, cyber-bullying and cyber-harassment must be detected through Twitter tweets and Facebook posts. The importance of the Arabic language lies in the increasing number of its speakers, whether native speakers or non-speakers [1]. The advantages of the Arabic language that the addition of accessories to the root of the triangular gives several derivations including (name; اسم), (verb; فعل), (adverb; حال) and (adjective; صفة. [2-3].

NLP is used in the creation and development of many modern applications [2-4]. Arabic NLP depends on Morphological analyzers/generators and Syntactic analyzers/generators [2]. And one of the difficulties faced by NLP is the richness of the Arabic language morphology [5].

Despite the importance of detecting Arabic cyber-bullying and cyber harassment on social media, this topic has attracted only few studies [6-7]. Furthermore, these studies used limited number of classifiers and Arabic NLP techniques.

In our research, we applied supervised machine learning; classification using several classifiers such as K-Nearest Neighbor [8], Support Vector Machine [9], Random Forest [10], and J48 [11], and Naive Bayes [12]. Texts are classified based on previously trained and categorized groups. We made various experiments using various ANLP techniques to study their effect on the classification process. Moreover, we collected two datasets from social media. This would provide more reliable results.

The contribution of this work is as follows.

- Providing an efficient Arabic cyber-bullying and

*Corresponding Author: Tarek Kanan; E-mail: tarek.kanan@zuj.edu.jo
DOI: 10.3966/160792642020092105016

cyber-harassment method on social media.
- Providing a comprehensive study using various classifiers, various ANLP, and multiple social media datasets.
- Insisting on the importance of Arabic offensive language detection on social media.

The rest of the study is planned as follows: Section Literature Review reviews the most related works. Section Methodology shows the methodology used in this research including: dataset collection, text preprocessing, and classification and evaluations. The experimental results including its dissuasion and the significant tests are given in section Experimental Results. The conclusion is in section Conclusions, and the final part is the references section.

2 Literature Review

In this section, we give some review of previous researches related to Arabic Social Media, Arabic Natural language processing tools, Machine Learning using Arabic Document, and Cyber-bullying and Cyber-harassment.

[13] analyzed emotions in Twitter through what is known as SemEval. Their tasks identifying the overall sentiment of the tweet, sentiment towards a topic with classification on a two-point and on a five-point ordinal scale.

[14] created a continuous flow of annotated Arabic twitter data through semi-supervised online learning. The results showed that the method gives good results for subjectivity analysis, and they showed a significant drop in performance for sentiment analysis.

[15] proposed a framework that combines sentiment analysis and subjective analysis through social media, with the aim of determining whether or not users are interested in a particular subject. The results of the experience were very encouraging for further research.

[16] studied research issues facing Arabic sentiment analysis through social media. The authors immediately suggested minimizing the challenges faced by Arab sentiment analysis, through determining the semantic orientation of Arabic Egyptian tweets.

[17] suggested using ML techniques to determine the polarity of tweets written in Arabic with the presence of dialects. The results showed that the dialect lexicon of the dialects increases the accuracy of the classifiers.

[18] developed SAMAR, a subjectivity and sentiment analysis (SSA) method for the Arabic social media. They produced annotated data for there is no gold-labeled social media SSA data for the Arabic language. The annotated data contains an assortment of DARDASHA (DAR) data.

[19] collected political articles and comments manually and made several copies of them. Then they researched the performance of different feature reduction techniques and several classifiers on data-sets. As well they compared namely the Traditional Text Categorization approach and the Stylometric Features approach as feature detection approaches.

[20] proposed a model based on correcting wrong Arabic words, they used machine learning algorithms; Support Vector Machine (SVM) to develop the model. The author's utilized 1,300,000 tweets arrive from 49,200 Twitter accounts to create the bigram-words list containing misspelled words.

[21] discussed the rule-based approach that has positively been used in emerging numerous natural language processing systems. The benefit of the rule-based approach over the corpus-based approach is clear for fewer-resourced languages, for which big corpora, may be parallel or bilingual, with representative structures and entities are neither available nor simply affordable, and for morphologically plentiful languages.

[22] compared stemming and light stemming. The results showed the light stem representation was superior in terms of classifier accuracy.

[23] designed a tailored new Arabic light stemmer called P-Stemmer. It deletes prefixes from words. By evaluation techniques commonly used in the information retrieval community, including 10-fold cross-validation and the Wilcoxon signed-rank test, they displayed that their approach stemming and classification are superior to state-of-the-art techniques.

[24] suggested creating Shami corpus, the first Levantine Dialect Corpus (SDC) covering data from the four dialects spoken in Palestine, Jordan, Lebanon, and Syria..

[25] studied the classification of the Arabic language. They suggested an approach to tackle these challenges. The proposed approach used the Frequency Ratio Accumulation Method (FRAM) as a classifier. Its features are selected using a bag of word technique and an improved.

[26] presented the second version of AlKhalil Morpho analyzer. The second version was created to correct the errors in the first version, and to strengthen the database with the missing data, the second version became more accurate in analysis and high coverage exceeding 99% for the words that have been analyzed.

[27] compared TF-IDF and syntax-based for feature selection and weighting, and class association rules vs. support vector machines for classification. The results showed the classification of lightly stemmed text gives more performance than the classification based on roots.

[28] compared three types of classifiers for Arabic text categorization. The results showed that the classification of the Arabic text using the NB outperforms the other classifiers.

[29] used Support Vector Machines (SVM) method in classifying Arabic text documents. They concluded that the Rocchio classifier gives better results when the size of the feature set is small, while the set of features

is great, the classifier is superior to other classifiers.

[30] studied text classification to extract useful information from big data. The results showed that the SMO classifier outperforms the three other classifiers as a training model and a classifier.

[31] improved classification accuracy by combining Naïve Bayes algorithm with Support vector machine by stacking.

[32] suggested a new multi-label classification algorithm named ITDGM. The algorithm is based on the interaction-based gravitational coefficient (IGC) and utilizes the algorithm to place the gravitational force to replace the mass of the particles.

[33] proposed a solution to the problem of document noise, specifically documents in Arabic, through a new Keyphrases extraction algorithm based on the Suffix Tree data structure (KpST). The results showed the approach for extracting Keyphrases improves the clustering results, and it is useful to support the research in the field any Arabic Text Mining applications.

[34] studied datasets collected from Twitter to understand the behavior of students at King Abdulaziz University. The dataset was compiled by developing a desktop application using the programming language, it is called Twitter Data Grabber, and the tweets were gathered in fifty days. These datasets inclusive 1,121 tweets. The authors used the K-Mean clustering algorithm unsupervised machine learning with different vector representation schemes; TF-IDF (term frequency-inverse document frequency) and BTO (binary-term occurrence).

[35] studied sentiment analysis by analyzing Arabic text and classifying it into previously known categories; Positive, Negative and Neutral. The author used a tree, Naïve Bayesian (NB), K Nearest Neighbor (KNN) and Support Vector Machines (SVMs) on Arabic Twitter corpus. The results showed that the SVM algorithm is superior of both NB and KNN in terms of recall, precision, and F1 measures.

[36] studied three techniques to classify Arabic datasets. The results of the experiment showed that the Support Vector Machine algorithm gives better comparison results than other algorithms.

[37] collected Arabic data-set from Twitter and applied Supervised Machine Learning algorithm includes; SVM, J48, C5.0, NNET (Neural Networks), NB, and k-NN classifiers, the better outcome was achieved by SVM classifier.

[34] proposed a model for analyzing Twitter in Modern Standard Arabic and Arabian Gulf dialect using K-Means with different vector representation schemes like TF-IDF and BTO. The authors applied clustering methods because there is no predictable category. The outcome displays that better vector representation through BTO instead of the TF-IDF scheme.

[38] suggested a methodology for extracting data from social networks. They used the power of sentiment analysis to detect cyber-bullying on Twitter. Then they used LingPipe tool to apply Naïve Bayes classifier, the result was achieved around 70%.

[39] applied three studies that have been conducted to examine the prevalence of cyber-bullying among university on the Internet. The first study showed that text messages and media social communication are the most applications through which to carry out electronic bullying. The second study showed that features of the goal of cyber-aggressive comments influenced perceptions of cyber-bullying. Previous research results showed that the impact of peer-directed online comments is more negative than those directed at unknown or unspecified individuals. The third study, using a methodology for checking electronic bullying, showed that the place, comments and forum participation.

[40] focused on increasing the prevalence of online social networking sites (SNS), especially among adolescents. Where it offers opportunities for cyber-bullying. The study indicates that the causes of the emergence of electronic bullying are psychological distress.

[6] presented predictive modeling detection of anti-social behavior, offensive language, and harassment through Arab social media. The results showed that the SVM classifier gives high accuracy and when using used the N-gram feature, the performance of the SVM classifier improves.

[41] studied the effects of the development of negative technology.

Recently, bullying has become moving from schools into social media to be now recognized as cyber-bullying. [7] suggested a solution to discovering and stopping cyber-bullying, from social media through two methods, a PHP language script for Twitter data and script in python to extract data from Facebook [7].

Table 1 shows the comparison between some previous researches. It can be concluded from these studies that only few studies concentrated on Arabic cyber-bullying and cyber-harassment such as [6-7]. Moreover, these studies used limited number of classifiers and ANLP techniques in their works. Furthermore, the used datasets were limited from one source such as YouTube reviews [6] and Twitter [7]. There is a vital need to concentrate more on the detection of Arabic offensive language and to study thoroughly using more comprehensive techniques. This was the motivation of our work.

3 Methodology

The chapter supplied an explanation of the implemented framework. The section demonstrates the overall methodology that must be followed to detection Cyber-Bullying and Cyber-Harassment by classification.

Table 1. Comparison of previous research

Ref. #	Data-set	Data-set from	Machine Learning Algorithms	Best Result	F1
[16]	3,500 Arabic Tweets	Twitter	NB and SVM	SVM	0.88
[17]	22,550 Arabic Tweets	Twitter	NB and SVM	NB	0.87
[20]	1,300,000 Arabic Tweets	Twitter	SVM	SVM	0.96
[23]	237,000 Arabic news Articles	Websites	NB, SVM, and RF	NB	0.99
[29]	1,123 documents	Websites	NB, SVM, k-NN, and Rocchio	SVM	0.88
[34]	1,121 Arabic Tweets	Twitter	K-Mean	K-Mean	N/A
[35]	3,700 Arabic Tweets	Twitter	NB, K-NN, and SVM	SVM	0.72
[37]	1,434 Arabic Tweets	Twitter	SVM, J48, C5.0, NNET, NB, and k-NN	SVM	0.93
[38]	15 million Tweets	Twitter	NB	NB	0.67
[6]	15,050 YouTube Comments	YouTube	SVM	SVM	0.91
[41]	1245 Tweets	Twitter	NB	NB	0.86
[7]	126,704 Tweet and Post	Twitter and Facebook	NB and SVM	SVM	0.93

3.1 System Architecture

In this section, the research proposed the use of both clustering algorithms and classifications to detect Cyber-Bullying and Cyber-Harassment in social media. The system is represented in Figure 1. In details, the data is firstly collected from social media. Next, the data is cleaned before being processed using ANLP techniques. Later, the data is being used by classifiers to learn patterns, and finally, the classifiers can detect testing posts as intact or containing bullying/ harassment. Aside from this, a youtube channel, a facebook page, and a twitter page were created to increase the awareness of this topic.

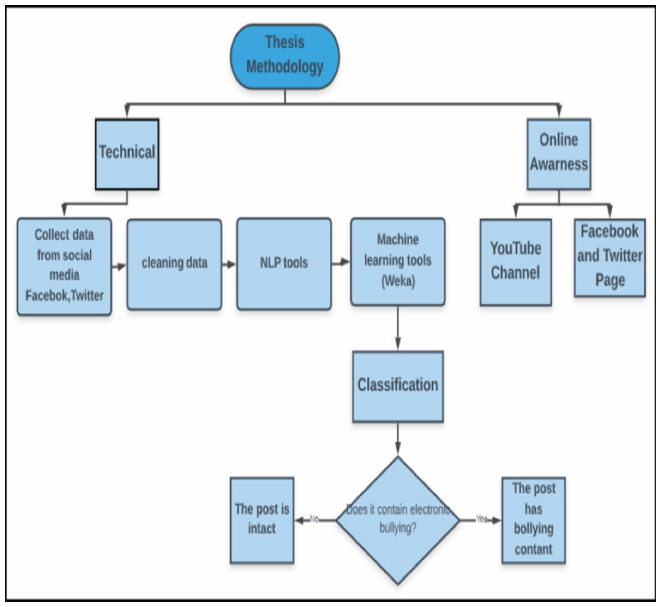


Figure 1. The proposed system architecture

In order to gather datasets for training and testing, we have used posts from the most important social media in the Arab world; Facebook and Twitter. Facebook and Twitter are very popular with young people and teenagers. We have compiled data sets written in Arabic from Twitter and Facebook, where a large number of datasets must be collected for training and testing.

To create a Twitter application, we visited the Twitter Developers Site, signed in, then went to "My Applications" and created a New Application. Create our Access Token, we chose the type of access we needed and make a note of our Oath Settings.

RStudio and Rtool are two software tools used in statistics and mathematics to extract tweets from Twitter. The two programs are interconnected with each other and are provided when you download the RStudio. You have to download Rtools first in order to install the libraries RStudio.

The total number of datasets is approximately 19,650 tweets and post, before deleting duplicate data and deleting publications that have hashtags. After deleting all the tweets and frequent posts and after deleting Hashtag, the total number of datasets reached is 6,138 tweet and post.

3.2 Pre-processing Tools

3.2.1 Normalization

Normalization is a process by the researcher to place all sentences in a common form to ensure the consistency in the text. Such conversion makes it easy to handle and is considered an earlier step for the Stemming.

To do normalization we created a code in Java programming language, where we used version 8.0.2 to construct our code. The code removes the Diacritics to demystify the words, remove the elongation, remove numbers and non-Arabic letters, remove special characters and punctuation, replacing 'Alif' (ا .... أ, إ, آ) and 'Yeh'(ى .... ي), 'Taa Marbuta' (ه .... ة) and 'Hamza' (ؤ, ئ .... ء) . Table2 Normalization Examples shows some Normalization examples.

Table 2. Normalization Examples

Normalization examples.	
Strip Diacritics المرأة / المراة	المرأة / المراة (The woman)
Strip Elongation	الم ..راة / المراة (The woman)
Normalize Hamza	ء / ىء، ؤ
Normalize Alif	أ / ا ،أ، آ
Normalize Yeh	ى / ي
Normalize Taa Marbuta	ه / ة
Normalize Repeated words	جدا جدا / جدا (Very Very)
Normalize Repeated letters	بلغووونا / بلغونا (Infoooormed Us)

3.2.2 Stopword Removal

Stop Word Removal is one of the pre-processing steps for texts before analysis. It is based on the exclusion of common words that have no value in the text and does not help the user to analyze texts and extracting the needs of the text. Stop-words are noise words, where the words of the text are compared with the words of the stop list [42]. The stop list is a list containing all the words of the sentence such as:

- Pronouns: أنا, هو, هي, أنتما, أنتم, انت, نحن, هما, هم (Translation: I, He, She, You, We, They)
- Conjunctions: عن, الى, على, في, حروف الجر: (Translation: In, on, To, …)
- Prepositions: لا, لكن, أم, حتى, ثم, الفاء, الواو, حروف العطف: (And, Then, Until, Or, But, …)
- Words that are not useful in classification: بالاشارة, بالنسبة, الاخرى, ممكن, غيرها, اولئك, بعض, كذلك, كهناك, حاليا (Translation: As for, Regarding, Could, Those, Some, Also, Currently, …)
- Directions: يسار, يمين, بعد, قبل, بجانب, خلف, امام (Translation: Front, Behind, Besides, Before, After, Left, Right)
- Any word that doesn't add any meaning to the text: غدا, اليوم, امس (Translation: Yesterday, Today, Tomorrow)

We created a java code to remove the stop words based on the Khoja Stemmer stop list. In addition to the additional words we added to the list, these words are removed from the data sets. Khoja stemmer stop list example:(, اهذا , الذى , فقط , إلي , عن , ألا , بشكل , مثل, أنفسنا غير)(Translation: Like, that, just, on, …) [43]

3.2.3 Stemming (P-Stemmer)

The main purpose of stemming is to get rid of the word forms. This means that one word may be either name, adverb, adjective, or character. By using the stem, all these shapes are discarded by returning the words to their original roots. As a result, reducing the number of words in the text.

P-Stemmer; removal for Prefixes, removing the prefix of words increases the effectiveness of document classification. Example المشاجرات.... مشاجرات

In this research, the P-Stemmer tool was used, and this tool was utilized to manipulate our dataset.

3.3 Machine Learning

To evaluate the performance of cyber-bullying and cyber-harassment detection we have used machine learning algorithms; supervised (classification). We applied five classification algorithms. In details, we used KNN, SVM, NB, RF, and J48. These are part of the classification methods, which include many more classifiers. We selected these classifiers as they are commonly used in the literature and cover various techniques. For example, RF and J48 are tree-based classifiers, NB is probabilistic-based classifier, SVM is kernel-based classifier, and KNN is an instance-based classifier. As for the implementations of these classifiers, we used WEKA toolkit [44]. In the following subsections, we provide a theoretical background for each of these classifiers.

3.3.1 K Nearest Neighbor (KNN)

This method uses the closest training records to the target testing record in order to label the target testing record. The closeness can be measured using the similarity or the distance. The maximum similarity or the smallest distance is the desired target [8].

3.3.2 Support Vector Machines (SVM)

SVM classifier uses the support vectors to create an N-dimensional hyper-plane to divide the data set into two classes. It is well known in the literature due to its highly accurate results [9].

3.3.3 Naïve Bayes

This classifier uses conditional probability, specifically, it adopts Bayes' theorem and assumes a strong independence between features. This classifier is known for its fast training and classification time [12].

3.3.4 Random Forest

This classifier is a part of ensemble learning classifiers, which use the training data to build many tree models that can be used for prediction in later phases. This type does not suffer from the overfitting problem that is common in decision trees [10].

3.3.5 Decision Trees J48

This classifier is a type of decision tree classifiers, which use the training data to build a tree model that can be used later for prediction. It is an extension to the previous ID3 decision tree classifier [11].

4 Experimental Work

The aim of this study is to detect cyber-bullying and cyber-harassment through Arabic contents on social media platforms such as Facebook and Twitter. This will be detected by applying classification algorithms. The used evaluation measurements are Recall, Precision, and F-Measure.

In this study, five classifiers were compared, namely KNN, SVM, NB, Random Forest, and J48. Three natural language processing tools were applied including Stop-Word Removal, Normalization, and Stemming. We conducted four different set of experiments as follows.

First, we trained the classifiers on datasets collected from Arabic social media, having been divided by human experts into positive and negative based on the sentiment of the writer. We used 10-fold Cross-Validation to divide training and testing articles.

Second, in a different set of experiments, we repeated the same steps without applying the Stemming (full word) on the dataset. We then compared the results with the stemming and without the stemming.

Third, the same experiments were repeated without applying the Stop-Word Removal. Then we compared the five classifiers with the Stop-Word Removal and without it.

Fourth, we compared the classifiers based on the type of the used dataset. We compared the performance of the classifiers in Facebook platform with their performance in Twitter platform.

4.1 Dataset Description

The total number of records in our dataset is composed 6,138 Twitter tweets and Facebook posts. In details, the number of Facebook records was 2,138; 1,000 of which were positive and 1,138 were negative. As for Twitter, the number of records was 4,000; 2,100 of which were positive and 1,900 were negative.

4.2 Evaluation Measurements

There are many approaches to evaluate the performance of text classification. We adopted the use of recall, precision, and F1 measurements as they are widely used in the literature in text classification.

The recall is the section of relevant documents that are retrieved and calculated as follows:

$$R = \frac{TP}{TP + FP}, \text{ if } TP+FN > 0, \quad \textbf{(1)}$$

Whereas TP is true positive and FP is false positive.

Precision is the section of retrieved documents that are relevant, and calculated as follows:

$$P = \frac{TP}{TP + FN}, \text{ if } TP+FN > 0, \quad \textbf{(2)}$$

Whereas TP is true positive and FN is false negative.

F-measure is the arithmetic mean between recall and precision. It is also used for comparisons, and calculated as follow:

$$F_1 = \frac{2 * R * P}{R + P} \quad \textbf{(3)}$$

4.3 Experimental Settings

In this research, we used WEKA 3.8 toolkit [44]. This toolkit was developed by the University of Waikato in New Zealand. This toolkit is written in Java. The default parameters of the classifiers were used. As for SVM, poly kernel was used, and K=3 was used in KNN. We used Quad Core i7 with 3.1GHZ CPU speed, 16G Ram memory.

4.4 Experimental Results

4.4.1 Comparing Classifiers Using All ANLP Tools

Table 3 showed the Recall results of the five classifiers on the dataset with all ANLP preprocessing techniques used, namely, Stop-Word Removal, Normalization, and Stemming. In these experiments, the whole dataset was used, which contains records from both Facebook and Twitter. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Recalls respectively compared to KNN.

Table 3. Recall values for all classifiers using all ANLP tools

Classifiers	Recall
KNN	0.625
SVM	0.937
Naïve Bayes	0.910
Random Forest	0.947
J48	0.873

Table 4 shows the Precision results of the five classifiers on the dataset using all ANLP preprocessing techniques. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Precisions respectively compared to KNN.

Table 4. Precision values for all classifiers using all ANLP tools

Classifiers	Precision
KNN	0.769
SVM	0.938
Naïve Bayes	0.911
Random Forest	0.947
J48	0.874

Table 5 showed the F-Measure results of the five classifiers on the dataset using all ANLP preprocessing

techniques. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest F-Measures respectively compared to KNN (0.555). The Random Force algorithm yields a higher F-measure accuracy (0.947). Figure 2 shows the F-Measure values for all classifiers. These results were expected as both RF and SVM classifiers proved their efficiency in various domains in the literature.

Table 5. F1-measure values for all classifiers using all ANLP tools

Classifiers	F-Measure
KNN	0.555
SVM	0.937
Naïve Bayes	0.910
Random Forest	0.947
J48	0.873

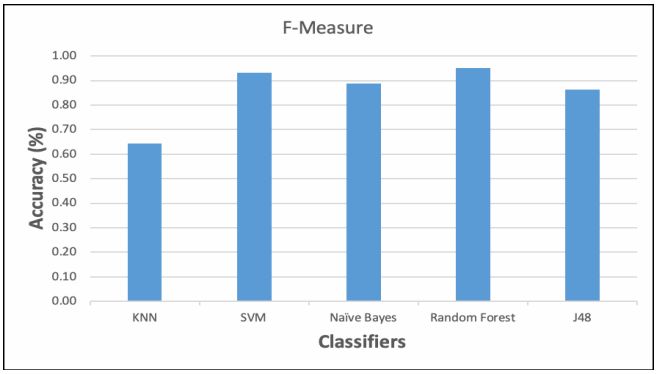


Figure 2. F-measurement of classifiers using all ANLP tools

4.4.2 Studying the Effect of Stemming on Classification Performance

Table 6 shows the Recall result of the five classifiers on the dataset using Stop-Word Removal and Normalization. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Recalls respectively compared to KNN.

Table 6. Recall values for all classifiers without stemming

Classifiers	Recall
KNN	0.644
SVM	0.931
Naïve Bayes	0.888
RF	0.950
J48	0.862

Table 7 showed the Precision result of the five classifiers on the dataset using Stop-Word Removal and Normalization. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Precision respectively compared to KNN.

Table 7. Precision values for all classifiers without stemming

Classifiers	Precision
KNN	0.779
SVM	0.932
Naïve Bayes	0.896
RF	0.950
J48	0.870

Table 8 shows the F-Measure result of the five classifiers on the dataset using Stop-Word Removal and Normalization. The results showed that the RF, SVM, NB, and J48 classifications achieved the highest F-Measures respectively compared to KNN (0.583). The Random Force algorithm yields a higher F-measure accuracy (0.949). Figure 3 shows the F1-Measure values for all classifiers. From these results, it was noted that stemming had a negative effect on the accuracy.

Table 8. F1-measure values for all classifiers without stemming

Classifiers	F-Measure
KNN	0.583
SVM	0.931
Naïve Bayes	0.887
RF	0.949
J48	0.861

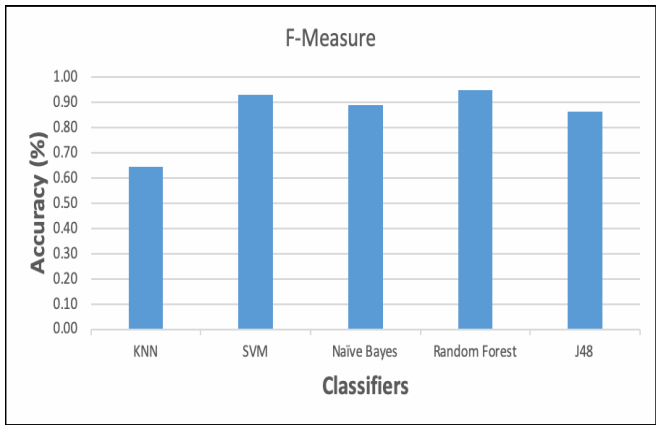


Figure 3. F measurement for all classifiers without stemming

Table 9 and Figure 4 represent the comparisons of the F1-Measure values of all classifiers with all NLP tools and with only Normalization and Stop-word removal (without stemming). Obviously, different classifiers acted differently to stemmed data. While the RF had an increase in F1 measure, this increase is insignificant. Other methods had their F1 measure decreased. SVM was less affected by stemming than J48, NB, and KNN.

Table 9. F1-measure values for all classifiers using stemmed and non-stemmed data

Classifiers	F1-Measure for all NLP	F1-Measure all NLP Except Stemming
KNN	0.697	0.583
SVM	0.937	0.931
NB	0.910	0.887
RF	0.947	0.949
J48	0.873	0.861

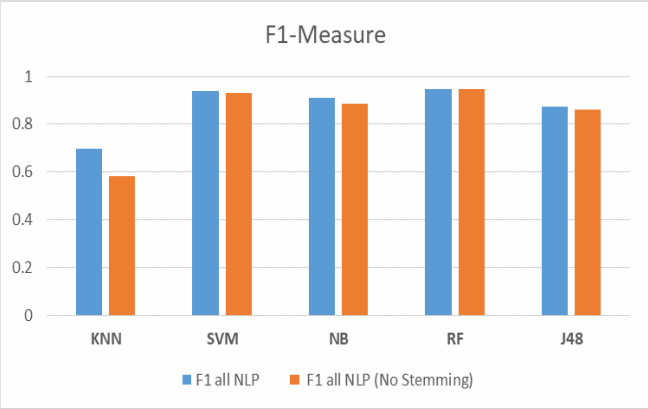


Figure 4. Comparing the F1 measurement of classifiers on Stemmed and non-stemmed data

Table 10 shows the Recall result of the five classifiers on the dataset using Stemming and Normalization (without Stop-Word Removal). The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Recall respectively compared to KNN.

Table 10. Recall values for all classifiers without stop-words

Classifiers	Recall
KNN	0.727
SVM	0.930
Naïve Bayes	0.837
RF	0.934
J48	0.830

Table 11 shows the Precision result of the five classifiers on the dataset using Stemming and Normalization (without Stop-Word Removal). The results showed that the RF, SVM, NB, and J48 classifications achieved the highest Precision respectively compared to KNN.

Table 11. Precision values for all classifiers without stop-words

Classifiers	Precision
KNN	0.776
SVM	0.930
Naïve Bayes	0.839
RF	0.935
J48	0.830

Table 12 shows the F1-measure results of the five classifiers on the dataset using Stemming and Normalization (without Stop-Word Removal). The results showed that the RF, SVM, NB, and J48 classifications achieved the highest F1-measures respectively compared to KNN (0.715). The Random Force algorithm yields a higher F1-measure accuracy (0.934). Figure 5 shows the F1-measure values for all classifiers. From these results, it was noted that the stopword removal contributed in increasing the F1 for classifiers in general, except in KNN. Obviously, stopwords are noisy terms that do not belong to any class, and therefore, they would have a negative effect on the accuracy.

Table 12. F1-measure values for all classifiers without stop-words removal

Classifiers	F1-Measure
KNN	0.715
SVM	0.930
Naïve Bayes	0.837
RF	0.934
J48	0.830

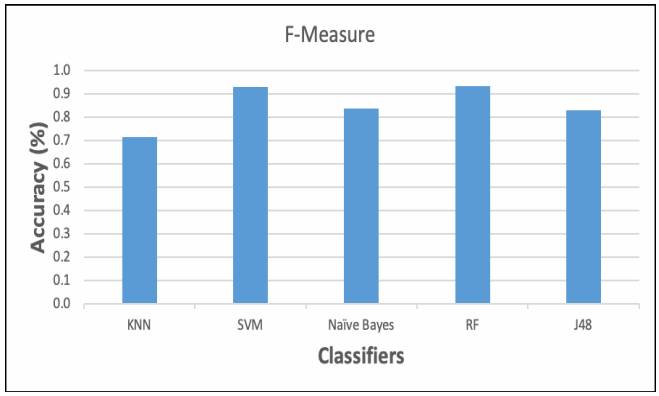


Figure 5. F measurement of all classifiers without stopwords

Table 13 and Figure 6 represent the comparisons of the F1-Measure values for all classifiers with all ANLP tools, without Stemming, and without Stop-Word Removal. It was noted that the best performing classifier was RF, whether all ANLP is used, without stemming, and without stop-words. It was noted also that the different classifiers were affected by ANLP differently. For example, stemming had a negative effect on the F1 measurement in all classifiers except RF, where the effect was negligible. The decrease in accuracy was due to the effect of the stemmer that produced non-dictionary terms after the stemming process, which in consequence decreased the F1 measurement. When removing stop-words, the F1 measurement was decreased in all the classifiers except in KNN. This could be because KNN does not use machine learning, as it is a lazy learner classifier.

Table 13. Comparison of F1-measure for all classifiers using various ANLP tools

Classifiers	F-Measure with all ANLP	F-Measure without Stemming	F-Measure without Stop-word
KNN	0.697	0.583	0.715
SVM	0.937	0.931	0.930
NB	0.910	0.887	0.837
RF	0.947	0.949	0.934
J48	0.873	0.861	0.830

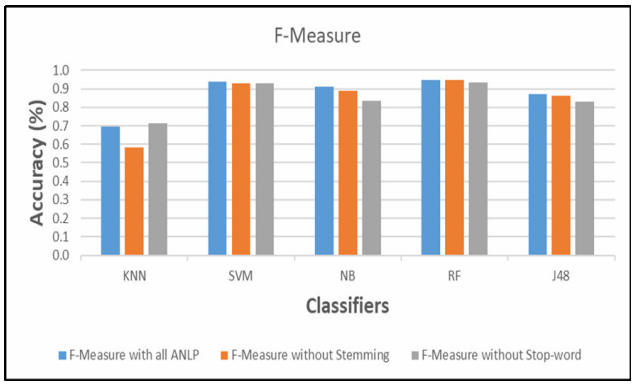


Figure 6. Comparing classifiers on various ANLP tools

4.4.3 Studying the Effect of the Dataset Type on the Classification Performance

4.4.3.1 Comparing the Classification Performance of all Classifiers using Facebook Dataset

Table 14 and Figure 7 represent the comparison of all the classifiers using all ANLP tools on Facebook dataset. It can be noted that the best performing classifier is SVM, hardly outperforming RF. NB comes next, followed by J48 and KNN respectively.

Table 14. Comparing the classification performance of all classifiers using Facebook dataset

Classifiers	Recall	Precision	F-Measure
KNN	0.789	0.837	0.782
SVM	0.918	0.918	0.917
NB	0.844	0.866	0.841
RF	0.914	0.914	0.914
J48	0.792	0.829	0.786

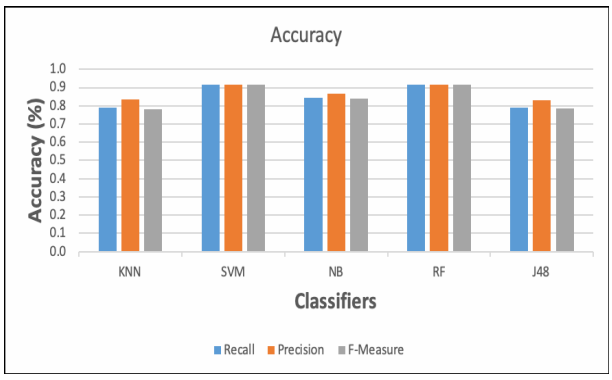


Figure 7. Comparing the classification performance of all classifiers based on the Facebook dataset

4.4.3.2 Comparing the Classification Performance of All Classifiers Using Twitter Dataset

Table 15 and Figure 8 represent the comparison of all classifiers using all ANLP tools on Twitter dataset. Again, the top performer here is SVM, hardly outperforming RF. NB comes next, followed by J48 and KNN. Here, the difference in performance between KNN and J48 is more obvious.

Table 15. Comparing the Classification Performance of all Classifiers Using Twitter Dataset

Classifiers	Recall	Precision	F-Measure
KNN	0.644	0.806	0.619
SVM	0.944	0.945	0.944
NB	0.898	0.898	0.898
RF	0.940	0.943	0.941
J48	0.839	0.861	0.840

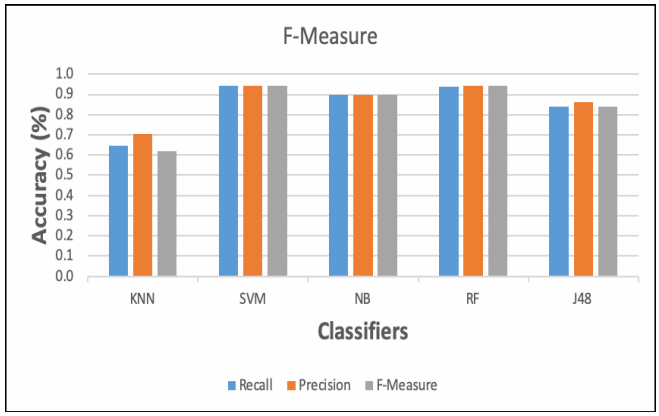


Figure 8. Comparing the classification performance of all classifiers based on Twitter dataset

4.4.3.3 Studying the Effect of the Dataset Type on Classification Performance

Table 16 and Figure 9 represent the comparison of F1-Measure values of all classifiers with all ANLP tools for both Facebook and Twitter datasets.

The results show that the accuracy of the Twitter dataset is higher than the accuracy of the Facebook dataset.

Table 16. Studying the effect of the dataset type on classification performance

Classifiers	F-Measure for Facebook	F-Measure for Twitter
KNN	0.782	0.675
SVM	0.917	0.944
NB	0.841	0.898
RF	0.914	0.941
J48	0.786	0.840

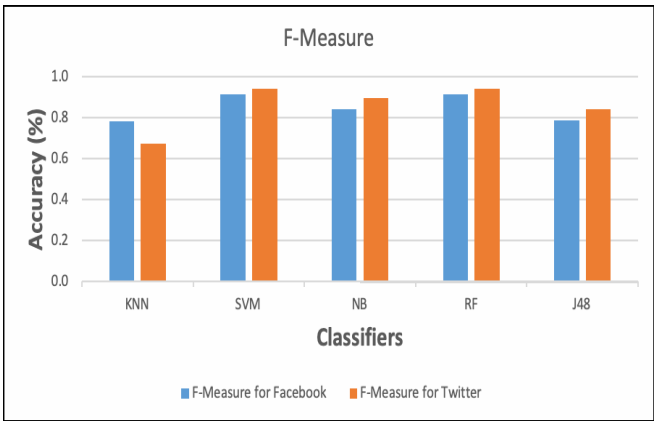


Figure 9. The effect of the dataset type on classification performance

When comparing the classifiers based on their performance on the two dataset types, namely Facebook and Twitter, we see that the F1-measure is higher on the Twitter dataset. The F1-measure of Random Forest on Twitter dataset is (0.941), while it is (0.914) on Facebook. The differences are due to the nature of the data as the dataset collected from Twitter was belonging to personal accounts, while the Facebook data was from public pages and groups.

To summarize, from the experimental results, it can be seen that the top performers are RF and SVM. These two classifiers have proved their efficiency in various domains related to text classification. As for the effect of stemming, it was noted that stemming decreased the accuracy. Further studies are needed to use more stemmers and to optimize these stemmers for Arabic language to improve the results. As for the stopword removal, it proved to improve the accuracy as this process removes noisy terms. Regarding the type of dataset, twitter had better classification accuracy than Facebook as the data belongs to personal accounts.

As can be seen from the results, the use of classification is promising in detecting cyber-bullying and cyber-harassment. These results can have significant impact both theoretically and practically. Theoretically, this study has improved the literature by concentrating on the cyber-bullying and cyber-harassment in Arabic region. It provided a comprehensive comparison that would aid future research works in this direction. Practically, machine learning can be integrated and embedded in social media and other Arabic forums so that they can be used in the real time detection and alerting many departments that are concerned with detecting such acts. This would be of a great aid to these parties and would contribute in fighting against these acts.

One drawback of this stemmer is that it considers only the prefixes of terms. It is worth mentioning that some existing stemmers already remove these parts such as snowball stemmer, which removes prefixes and suffixes. Therefore, the integration of such stemmers with our stemmer can be further analyzed to optimize the results. Another drawback of the method is its use of a static list of prefixes. Finally, it was noted that the performance of this stemmer varies according to the used ANLP methods and classifiers. This work can be further extended to include more classifiers such as deep learning classifiers, which have been used in various domains such as network intrusion [46-47], multimedia [48], and energy saving [49]. Furthermore, this work could be the basis for many domains such as recommender systems [50]. Finally, the performance of these methods need to be studied with huge data sizes [51].

4.5 Significance Test

We used the F1-Measure results to perform a statistical significance test between Random Forest and each one of the SVM and NB. Table 17 shows the F1-Measure results of RF, SVM, and NB classifiers using various pre-processing tools.

Table 17. F1-measure for RF, SVM, and NB classifiers using various pre-processing tools

	RF	SVM	NB
With all Preprocessing tools	0.947	0.937	0.910
Without Stemming	0.949	0.931	0.887
Without Stop-word	0.934	0.930	0.837

We used the Wilcoxon signed-ranked test to compare our proposed RF classifier and each of SVM and NB classifiers [45], with the P-value less than or equal to 0.05. This test is very popular for information retrieval evaluation.

We did the test three times and successfully rejected our null hypothesis. For each one of the three tests: "the median difference of the F-Measure of RF and SVM and NB Classifiers is less than or equal to zero". We concluded that, using the F-Measure for evaluation, RF is statistically significant than each of the SVM and NB. Table 18 and Table 19 consequently show the calculations and results, with the final Wilcoxon (Wcal) values for the RF classifier and SVM and NB classifiers.

Table 18. Values toward calculating wcal for the wilcoxon signed-rank for the F1 measure by absolute differences (abs) between RF classifier and SVM classifier

i	X2,i	X1,i	sgn	abs	Ri	Sgn* Ri
3	0.934	0.930	1	0.004	1	1
1	0.947	0.937	1	0.010	2	2
2	0.949	0.931	1	0.018	3	3

Table 19. Values toward calculating wcal for the wilcoxon signed-rank for the F1-measure by absolute differences (abs) between RF classifier and NB classifier

i	X2,i	X1,i	sgn	abs	Ri	Sgn* Ri
3	0.934	0.837	1	0.097	3	3
1	0.947	0.910	1	0.037	1	1
2	0.949	0.887	1	0.062	2	2

5 Conclusions

This work proposed the detection of cyber-bullying and cyber-harassment by applying the classification algorithms and clustering on a dataset collected from Facebook and Twitter. The difficulty of Arabic language was one of the challenges in this work as it has high morphology, derivation, and many other characteristics.

When applying the classification algorithms to all datasets with all ANLP tools, the results showed that the RF algorithm gives the highest value for the F-Measure scale followed by SVM, NB, J48, and KNN respectively. The same result occurs when applying the classification algorithms to all datasets without Stemming and also when applying the classification algorithms to all datasets without Stop-Word Removal. However, when applying the classification algorithms to the Facebook dataset with all ANLP tools, the results showed that the SVM algorithm gives the highest values for the F-Measure followed by the RF, NB, J48, and KNN respectively. The same result occurs when applying the classification algorithms to the Twitter dataset with all ANLP tools.

From the experimental results, it can be observed that SVM gives better results as the size of datasets increases. It can be noted also that the results are better when applying classification algorithms with the application of natural language processing tools. When datasets are separated, the results of the classification algorithms on the Twitter dataset is better than the Facebook dataset, because tweets collected from Twitter are from personal accounts while the Facebook Posts are collected from general open pages.

Future work can be conducted in many directions. First, more classifiers can be used in the detection process such as deep learning classifiers. Second, the effect of stemmers can be studied more to optimize the results. Finally, including dialects and Arabic jargons would be interesting to study in more details.

References

[1] S. Tartir, I. Abdul-Nabi, Semantic Sentiment Analysis in Arabic Socialmedia, *Journal of King Saud University: Computer and Information Sciences*, Vol. 29, No. 2, pp. 229-233, April, 2017.

[2] A. Farghaly, K. Shaalan, Arabic Natural Language Processing: Challenges and Solutions, *ACM Transactions on Asian Language Information Processing (TALIP)*, Vol. 8, No. 4, Article number: 14, December, 2009.

[3] K. Darwish, Building a Shallow Arabic Morphological Analyser in One Day, *ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, Pennsylvania, USA, 2002, pp. 1-8.

[4] A. H. Aliwy, K. S. Aljanabi, Z. A. A. AboAltaheen, Classification of Arabic Texts Using Four Classifiers, *International Journal of Computer Science and Information Security*, Vol. 15, No. 8, pp. 16-19, August, 2017.

[5] W. Salloum, N. Habash, Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation, *The First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, Edinburgh, Scotland, 2011, pp. 10-21.

[6] A. Alakrot, L. Murray, N. S. Nikolov, Towards Accurate Detection of Offensive Language in Online Communication in Arabic, *Procedia Computer Science*, Vol. 142, pp. 315-320, November, 2018.

[7] B. Haidar, M. Chamoun, A. Serhrouchni, A Multilingual System for Cyberbullying Detection: Arabic Content Detection Using Machine Learning, *Advances in Science, Technology and Engineering Systems Journal*, Vol. 2, No. 6, pp. 275-284, December, 2017.

[8] J. Wang, J. D. Zucker, Solving the Multiple-Instance Problem: A Lazy Learning Approach, *17th International Conference on Machine Learning*, Stanford, CA, USA, 2000, pp. 1119-1126.

[9] C. Cortes, V. N. Vapnik, Support-vector Networks, *Machine Learning*, Vol. 20, No. 3, pp. 273-297, September, 1995.

[10] L. Breiman, Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5-32, October, 2001.

[11] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[12] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-color Illustrations*, Springer, 2001.

[13] S. Rosenthal, N. Farra, P. Nakov, Semeval-2017 Task 4: Sentiment Analysis in Twitter, *11th International Workshop on Semantic Evaluation (SemEval-2017)*, Vancouver, Canada, 2017, pp. 502-518.

[14] E. Refaee, V. Rieser, An Arabic Twitter Corpus for Subjectivity and Sentiment Analysis, *Ninth International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, 2014, pp. 2268-2273.

[15] N. F. B. Hathlian, A. M. Hafez, Subjective Text Mining for Arabic Social Media, *International Journal on Semantic Web and Information Systems (IJSWIS)*, Vol. 13, No. 2, pp. 1-13, June, 2017.

[16] S. R. El-Beltagy, A. Ali, Open Issues in the Sentiment Analysis of Arabic Social Media: A Case Study, *IEEE 9th International Conference on Innovations in Information Technology (IIT)*, Abu Dhabi, UAE, 2013, pp. 215-220.

[17] R. M. Duwairi, Sentiment Analysis for Dialectical Arabic, *IEEE 2015 6th International Conference on Information and Communication Systems (ICICS)*, Amman, Jordan, 2015, pp. 166-170.

[18] M. Abdul-Mageed, M. Diab, S. Kubler, Subjectivity and Sentiment Analysis for Arabic Social Media, *Computer Speech & Language*, Vol. 28, No. 1, pp. 20-37, January, 2014.

[19] R. Abooraig, S. Al-Zu'bi, T. Kanan, B. Hawashin, M. Al Ayoub, I. Hmeidi, Automatic Categorization of Arabic Articles Based on Their Political Orientation, *Digital Investigation*, Vol. 25, pp. 24-41, June, 2018.

[20] E. A. Abozinadah, J. Jones, Improved Micro-blog Classification for Detecting Abusive Arabic Twitter Accounts, *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, Vol. 6, No. 6, pp. 17-28, November, 2016.

[21] K. Shaalan, Rule-based Approach in Arabic Natural Language Processing, *The International Journal on Information and Communication Technologies (IJICT)*, Vol. 3, No. 3, pp. 11-19, June, 2010.

[22] R. Duwairi, M. Al-Refai, N. Khasawneh, Stemming versus Light Stemming as Feature Selection Techniques for Arabic Text Categorization, *2007 IEEE Innovations in Information Technologies (IIT)*, Dubai, UAE, 2007, pp. 446-450.

[23] T. Kanan, E. A. Fox, Automated Arabic Text Classification with p-s Temmer, Machine Learning, and a Tailored News Article Taxonomy, *Journal of the Association for Information Science and Technology*, Vol. 67, No. 11, pp. 2667-2683, November, 2016.

[24] K. Abu Kwaik, M. K. Saad, S. Chatzikyriakidis, S. Dobnik, Shami: A Corpus of Levantine Arabic Dialects, *Eleventh International Conference on Language Resources and Evaluation (LREC18)*, Miyazaki, Japan, 2018, pp. 3645-3652.

[25] R. M. Sallam, H. M. Mousa, M. Hussein, Improving Arabic Text Categorization Using Normalization and Stemming Techniques, *International Journal of Computer Applications*, Vol. 135, No. 2, pp. 38-43, February, 2016.

[26] M. Boudchiche, A. Mazroui, M. O. A. O. Bebah, A. Lakhouaja, A. Boudlal, Alkhalil Morpho Sys 2: A Robust Arabic Morpho-syntactic Analyzer, *Journal of King Saud University-Computer and Information Sciences*, Vol. 29, No. 2, pp. 141-146, April, 2017.

[27] Y. Haralambous, Y. Elidrissi, P. Lenca, *Arabic Language Text Classification Using Dependency Syntax-based Feature Selection*, https://arxiv.org/abs/1903.11549, 2014.

[28] R. M. Duwairi, R. Al-Zubaidi, A Hierarchical k-nn Classifier for Textual Data, *International Arab Journal of Information Technology*, Vol. 8, No. 3, pp. 251-259, July, 2011.

[29] A. Wahbeh, M. Al-Kabi, Q. Al-Radaideh, E. Al Shawakfa, I. Alsmadi, The Effect of Stemming on Arabic Text Classification: An Empirical Study, *International Journal of Information Retrieval Research (IJIRR)*, Vol. 1, No. 3, pp. 54-70, July-September, 2011.

[30] R. Elhassan, M. Ahmed, Arabic Text Classification on Full Word, *International Journal of Computer Science and Software Engineering (IJCSSE)*, Vol. 4, No. 5, pp. 114-120, May, 2015.

[31] A. O. Christiana, O. S. Oladeji, A. T. Oladele, Binary Text Classification Using an Ensemble of Naïve Bayes and Support Vector Machines, *Computer Science & Telecommunications*, Vol. 52, No. 2, pp. 37-45, October, 2017.

[32] L. Peng, Y. Liu, Gravitation Theory Based Model for Multi-label Classification, *International Journal of Computers, Communications & Control*, Vol. 12, No. 5, pp. 689-703, October, 2017.

[33] H. Froud, I. Sahmoudi, A. Lachkar, An Efficient Approach to Improve Arabic Documents Clustering Based on a New Keyphrases Extraction Algorithm, *Second International Conference on Advanced Information Technologies and Applications (ICAITA-2013)*, Dubai, UAE, 2013, pp. 243-256.

[34] H. Al-Rubaiee, K. Alomar, Clustering Students' Arabic Tweets Using Different Schemes, *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 4, pp. 276-280, April, 2017.

[35] W. Hadi, Classification of Arabic Social Media Data, *Advances in Computational Sciences and Technology*, Vol. 8, No. 1, pp. 29-34, June, 2015.

[36] A. H. Mohammad, T. Alwada'n, O. Al-Momani, Arabic Text Categorization Using Support Vector Machine, Naïve Bayes and Neural Network, *GSTF Journal on Computing (JoC)*, Vol. 5, No. 1, pp. 108-115, August, 2016.

[37] W. Alabbas, H. M. al-Khateeb, A. Mansour, G. Epiphaniou, I. Frommholz, Classification of Colloquial Arabic Tweets in Real-time to Detect High-risk Floods, *2017 IEEE International Conference on Social Media, Wearable and Web Analytics (Social Media)*, London, UK, 2017, pp. 1-8.

[38] H. Sanchez, S. Kumar, *Twitter Bullying Detection*, https://users.soe.ucsc.edu/~shreyask/ism245-rpt.pdf, 2011.

[39] E. Whittaker, R. M. Kowalski, Cyberbullying via Social Media, *Journal of School Violence*, Vol. 14, No. 1, pp. 11-29, 2015.

[40] B. O'Dea, A. Campbell, Online Social Networking and the Experience of Cyber-bullying, in: B. Wiederhold, G. Riva (Eds.), *Annual Review of Cybertherapy and Telemedicine 2012: Advanced Technologies in the Behavioral, Social and Neurosciences*, Vol. 181 of Studies in Health Technology and Informatics, IOS Press, 2012, pp. 212-217.

[41] Hariani, I. Riadi, Detection of Cyberbullying on Social Media Using Data Mining Techniques, *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 15, No. 3, pp. 244-250, March, 2017.

[42] A. Benabdallah, M. A. Abderrahim, M. E.-A. Abderrahim, Extraction of Terms and Semantic Relationships from Arabic Texts for Automatic Construction of an Ontology, *International Journal of Speech Technology*, Vol. 20, No. 2, pp. 289-296, June, 2017.

[43] M. Saad, *Stopwords List* (koja stemmer), https://github.com/motazsaad/khoja-stemmer-command-line/blob/master/stopwords.txt, 2015.

[44] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update, *SIGKDD Explorations*, Vol. 11, No. 1, pp. 10-18, June, 2009.

[45] E. A. Gehan, A generalized Wilcoxon Test for Comparing Arbitrarily Singly-censored Samples, *Biometrika*, Vol. 52, No. 1/2, pp. 203-224, June, 1965.

[46] V. Punitha, C. Mala, N. Rajagopalan, A Novel Deep Learning Model for Detection of Denial of Service Attacks in HTTP Traffic over Internet, *International Journal of Ad Hoc and Ubiquitous Computing*, Vol. 33, No. 4, pp. 240-256, 2020.

[47] S. Bhardwaj, R. R. Ginanjar, D. S. Kim, Deep Q-learning Based Resource Allocation in Industrial Wireless Networks for URLLC, *IET Communications*, Vol. 14, No. 6, pp. 1022-1027, April, 2020.

[48] Z. Wang, S. Mao, W. Yang, Deep Learning Approach to Multimedia Traffic Classification Based on QoS Characteristics, *IET Networks*, Vol. 8, No. 3, pp. 145-154, May, 2019.

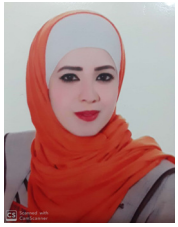
[49] L. X. Wu, S. J. Lee, A Deep Learning-Based Strategy to the Energy Management-Advice for Time-of-Use Rate of Household Electricity Consumption, *Journal of Internet Technology*, Vol. 21, No. 1, pp. 305-311, January, 2020.

[50] C. Chootong, T. K. Shih, A. Ochirbat, W. Sommool, W. K. T. M. Gunarathne, C. K. Chang, LCRec: Learning Content Recommendation (Wiki-based Skill Book), *Journal of Internet Technology*, Vol. 20, No. 6, pp. 1753-1766, November, 2019.

[51] J. Liu, J. Wu, L. Guo, M. Li, M. Zhang, Research on Intelligent Scheduling Strategy of Elevator Group under the Big Data Platform, *International Journal of Internet Protocol Technology*, Vol. 13, No. 2, pp. 85-93, 2020.

Biographies

Tarek Kanan is an assistant professor in the Department of Computer Science/Artificial Intelligence at Al-Zaytoonah University of Jordan. He obtained his Ph.D. in 2015 from Virginia Tech. His research interests are in the Artificial Intelligence domains. He had several prestigious Journal/Conference publications and was in various conferences' committees.



Amal Aldaaja obtained her Master degree in 2019 from AlZaytoonah University of Jordan. She obtained her Bachelor's degree in Computer Engineering in 2015 from AlBalqa Applied University. Her research interest is in artificial intelligence. She had several Journal/ Conferences publications.

Bilal Hawashin is at the Department of Computer Information Systems at Alzaytoonah University of Jordan. He received his Ph.D. in Computer Science, College of Engineering, from Wayne State University in 2011. His research interest is in artificial intelligence fields. He has various publications in referred journals and conference proceedings.