Indonesian Journal of Electrical Engineering and Computer Science
Vol. 19, No. 2, August 2020, pp. 1010~1020
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v19.i2.pp1010-1020

Exploration of the best performance method of emotions classification for arabic tweets

Mohammed Abdullah Al-Hagery, Manar Abdullah Al-assaf, Faiza Mohammad Al-kharboush
Department of Computer Science, College of Computer, Qassim University, Kingdom of Saudi Arabia

Article Info

Article history:

Received Nov 13, 2019
Revised Jan 14, 2020
Accepted Mar 9, 2020

Keywords:

Arabic tweets
Emotion analysis classification
Machine learning
Feature extraction
N-gram

ABSTRACT

Arab users of social media have significantly increased, thus increasing the opportunities for extracting knowledge from various areas of life such as trade, education, psychological health services, etc. The active Arab presence on Twitter motivates many researchers to classify and analysis Arabic tweets from numerous aspects. This study aimed to explore the best performance scenarios in the classification of emotions conveyed through Arabic tweets. Hence, various experiments were conducted to investigate the effects of feature extraction techniques and the N-gram model on the performance of three supervised machine learning algorithms, which are Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression (LR). The general method of the experiments was based on five steps; data collection, preprocessing, feature extraction, emotion classification, and evaluation of results. To implement these experiments, a real-world Twitter dataset was gathered. The best result achieved by the SVM classifier when using a bag of words (BoW) weighting schema (with unigrams and bigrams or with unigrams, bigrams, and trigrams) exceeded the best performance results of other algorithms.

Corresponding Author:

Mohammed Abdullah Al-Hagery,
Department of Computer Science,
College of Computer,
Qassim University,
Al-Mulida, Qassim Region, Kingdom of Saudi Arabia.
Email: hajry@qu.edu.sa

1. INTRODUCTION

Social media processing in the real world includes analysis of real problems, events, and a wide range of applications [1-3], as well as analysis of tweets associated with the cybersecurity problems [4, 5], opinions mining, analysis of tweets associated with areas like automated business, education [6, 4]or other social issues. Usually, the concentration of these analyses is on the contents given as a text segment, such as tweets, emails, messages, etc. The expression of emotions is a particularly integral part of text segments in social media because emotions represent a universal language that all people can understand. Emotions represent a key factor in human nature and behaviour and are a means for individuals to express their perspectives and opinions, analyse events, provide assessments, and communicate with each another via social media messages [7]. Therefore, social media networks provide a host of information revealing users' opinions and insights into current affairs, ongoing events, and human interests [8-10].

Twitter is a massive repository of text segments [8] that are constantly being written by users. It is a rapidly growing micro-blogging social media platform where individuals post their emotions and opinions in simple expressions. Twitter publishes more than 400 million tweets daily [11], with a maximum of 280 characters in each tweet. In January 2019, Twitter users in Saudi Arabia ranked fourth in the world, indicating high Arabic interaction on Twitter. The statistics are graphically illustrated in Figure 1.

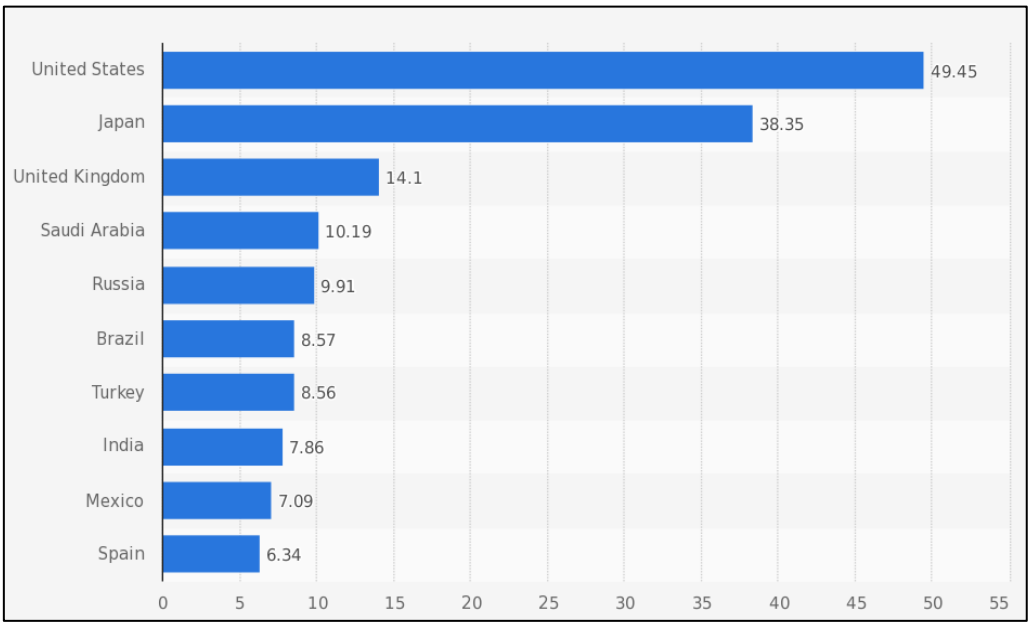

Figure 1. Leading countries, with the number of twitter users in millions [12]

Analysing tweets faces many difficulties due to spelling mistakes, emoticons, and slang (colloquial wording) [13] presented in the tweets, in addition to the type of language used and its complexity during processing and analysis. Therefore, these difficulties affect the classification of human emotion in tweets [14]. Many researchers have studied emotions in English tweets, yet few of them have focused on emotions in Arabic tweets [15].

Due to the nature of the Arabic language, analysing emotions is considered a difficult challenge, especially during the preprocessing phase. Arabic is a complicated language for various reasons; it has an exceptionally complex morphology compared to other languages. It also has complex sentences and many vocabularies that have multiple synonyms. This causes a higher difficulty in analysing emotions written in Arabic than in other languages.

Many scientists have studied the field of sentiment analysis (SA); however, few of them have analysed or detected emotions in tweets. Emotion analysis (EA) and characterisation are not like SA [16]. For example, SA aims to classify a text as positive, negative, or neutral; EA is more open to elucidate emotions conveyed in the text, such as sadness, optimism, joy, and so on. Although just six or eight emotions are viewed as fundamental emotions, the number of emotions considered by EA can be much greater [17].

The EA can introduce services for psychological health that improve the follow-up of patients with depression by using applications based on machine learning. Machine learning techniques are used to recognise, analyse, or classify human feelings, opinions, assessments, demeanours, and emotions toward entities such as products, administrations, people, issues, etc. [6, 18]. Therefore, the ability of machines to classify users' emotions correctly should be exploited to follow up on patients' psychological states. When patients answer their doctor's question, "How has your condition been in the last month?", the answer does not accurately reflect the psychological state of the patient. Since these answers do not provide further details about the patient's condition daily, they are general and lack a deep analytical vision. Consequently, a machine's ability should be utilized to give the most accurate answer by classifying the patient's tweets based on his or her emotions. Indeed, EA services are not limited to psychological health but can also contribute to the detection of chronic psychiatric illnesses such as depression [19-21]. Therefore, this research focused on employing a valuable approach for analysing the emotion of tweets written in Arabic to be utilised effectively for individuals' psychological health. Additionally, this study reduced the challenges posed by analysing the Arabic language, especially in the preprocessing phase, by using Python libraries.

Therefore, the objective of this study is to explore the best method for classifying emotions in Arabic tweets to understand people's impressions of provided services or products. The study used the following methods: tweets collection, preprocessing, feature extraction, emotion classification, and evaluation of results. This paper is organized as follows: Section 2 elucidates the literature review. The methodology is described in Section 3, as well as Section 4 presents the experimental results whereas Section 5 shows the discussion and evaluation of the results. Finally, Section 6 provides conclusions and future work.

2. LITERATURE REVIEW

EA is mostly based upon empirical studies that investigate how to detect emotions in texts. Therefore, this section provides brief reviews of related work on EA and presents recent efforts made in this field as well. For example, a suggested model for classifying emotions in Arabic tweets accomplished [14] and Waikato Environment for Knowledge Analysis (WEKA) was utilized for building this model, which categorised Arabic tweets into four principle feelings: sadness, joy, disgust, and anger. The accuracy of this model reached 80%. In another study, a rules-based approach and knowledge base were used to classify vast amounts of tweets into four classes of emotion based on the circumplex model. For feature extraction, part of speech (POS) tagging was employed to implement rules to detect emotion conveyed in tweets correctly. The overall accuracy of this method was 85%, which is considered a satisfactory result [22]. Furthermore, the first emotion intensity dataset for tweets built in [23]. The researchers utilized best-worst scaling to increase the consistency of annotations and to obtain fine-grained scores. They found that the emotional intensity of the tweets was expressed by emotion-word hashtags. Also, Badarneh et al. considered EA as a fine-grained approach, tackling an EA problem as a multi-label problem. They created a dataset of Arabic tweets that was annotated by two native Arabic speakers. Cohen's kappa was used to measure the agreement between annotators.

The annotation task was applied from reader and writer perspectives; the highest agreement in the writer dataset was about joy, while the highest agreement in the reader dataset was about fear [17]. Additionally, Jain et al. proposed a computational model of emotion switching for an intelligent agent [24]. Sangam, Shinde combined two classifiers SVM and ANN for sentiment classification [6], it is a general model, the experiments were performed on movie reviews dataset for any language, without consideration of complex languages such as Arabic language that has been taken into consideration in our research.

Hasan et al. improved a system based on supervised machine learning that automatically classified emotion in tweets. Their method involved two phases; the first phase was an offline training task, while the second phase related to classifying the texts online. In the first task, the model classified emotions correctly in 90% of English tweets. The second phase contained two stages; the first stage was a binary classification of tweets with or without emotion. Then, a fine-grained emotion classification was conducted on emotional tweets [25]. In the other hand, the researchers in [26] collected a Twitter dataset and classified the data into nine emotional categories: anger, fear, disgust, guilt, joy, interest, sadness, shame, and surprise. These classifications were made using supervised machine learning classifiers. To find the effective classifier for emotion extraction of the dataset, they performed a comparative study on the performance of Artificial Neural Network (ANN), SVM, and NB classifiers. The researchers also separately investigated the performance of these classifiers with the bag of words (BoW) and bigram features. According to their results, the bigram feature provided better performance than the BoW feature. Furthermore, SVM performed better than the other two classifiers.

Moreover, many attempts have been made for using the lexical approach to detect the strength of relaxation and stress expressed in messages available on social media, for example, TensiStrength system. This system was able to extract aberrant and direct expressions of relaxation and stress. The results indicated that TensiStrength worked well for some intelligent applications[27]. As well, in [28], the researchers applied a number of supervised algorithms for irony discovery in Arabic tweets. They used a binary classifier, which had high accuracy; precision reached 72.76%.

Abdelaal et al. [29] classified Arabic tweets into five classes—sports, politics, culture, general topics, and technology—using ensemble methods (boosting, bagging, and stacking). These classifications were based on the tweets' contents and morphological characteristics. The results confirmed that ensemble methods achieved better performance than single classifiers such as Naïve Bayes (NB), decision tree, and sequential minimal optimisation classifiers [29].

As well, other studies compared many classification models using an English corpus [30-39, 40]. Also, Xu et al. proposed different sampling methods to improve the classification performance of English text by reducing the imbalance ratio between training classes [41]. Our study, however, explored the best performance scenarios in the classification of emotions conveyed through Arabic tweets.

3. METHODOLOGY

The methodology consisted of tweets collection, preprocessing, features engineering, cross-validation sampling, tweets classification (based on four emotions and using three machine learning algorithms), and evaluation of results. Figure 2 shows the main phases of the methodology.

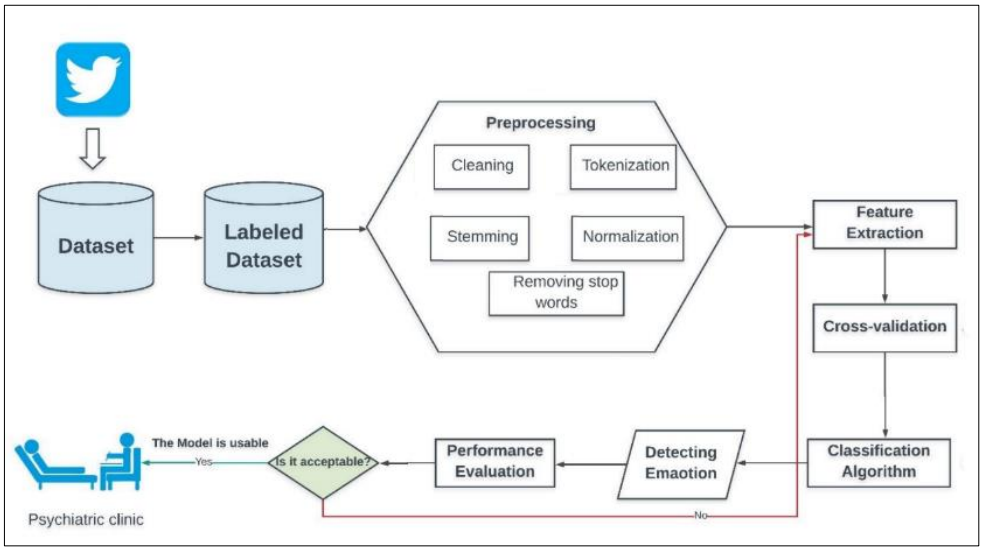

Figure 2. The overall framework

3.1. Tweets collection and labelling

The tweets were collected from Twitter using Netlytic [42], which offers a user-friendly interface and fast way to collect and visualize public data from various social media sources. With Netlytic, tweets matching a specific hashtag were collected. To collect and label tweets depending on emotion status, the Circumplex model, which was introduced and utilised by [43]. According to this model, all emotional states of humans are described in two-dimensional space. The horizontal axis represents the happiness or sadness of a person, while the vertical axis represents the activation of a person's emotion. In other words, the model divides human emotions into four main classes: Happy-Active, Happy-Inactive, Unhappy-Active, and Unhappy-Inactive, as shown in Figure 3. In the present study, a list of 28 keywords representing the four emotion classifications of the Circumplex model was created, and these keywords were then translated to the Arabic language. These keywords were subsequently used to find emotional tweets containing the keywords as hashtags.

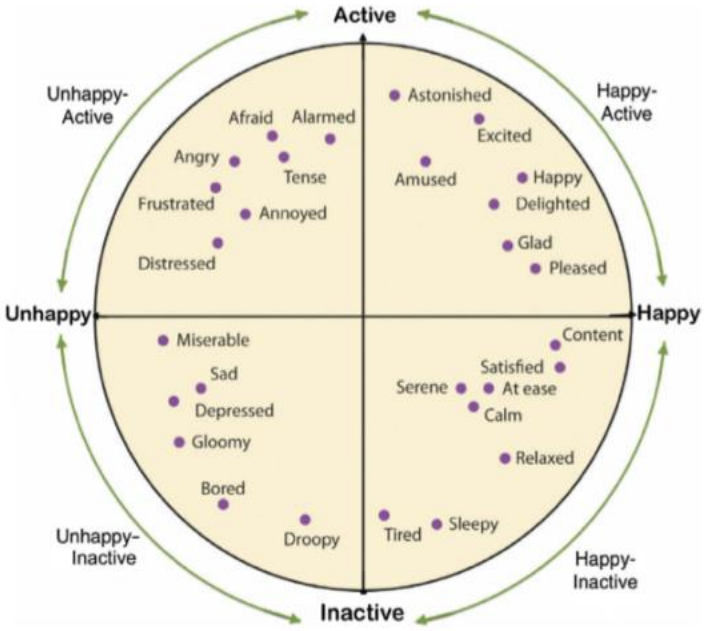

Figure 3. Circumplex model [43]

3.2. Preprocessing

The Preprocessing of the Arabic content is a critical task; Arabic has an extremely complex morphology compared to English. Therefore, the accuracy of the classification Arabic text was affected by preprocessing. The main steps of task included data cleaning, tokenization, stemming, normalisation, and stop-word removal.

3.2.1. Data cleaning

Data cleaning is a process that can improve the quality of the data, thereby improving the efficiency of machine learning algorithms in EA. Data cleaning can remove noisy words and unmeaningful content, resulting in reduced data size and increased data quality. The data cleaning steps of Arabic tweets included:

a) Remove Twitter's shortcuts, such as @username, retweet and replay.
b) Delete special characters like @, %, #, &, ~, etc.
c) Remove punctuation marks, emoticons, and numbers.
d) Delete URL links.
e) Delete words containing only one character, such as "ص"

PyArabic library, in which a specific Arabic language library in Python provides basic functions to manipulate Arabic letters and text, was used.

3.2.2. Tokenization

Tokenization is the method of dividing the given content into small pieces called tokens. In this step, the content of the tweets was divided into a sequence of tokens, where each token represented one or more words. The NLTK Python library was used to convert the tweets to tokens.

3.2.3. Normalisation

In natural language processing, normalisation standardises the shape of the text, placing all words on the same footing to be processed uniformly. This task depends on the nature of letters in the language that will be normalised. In the Arabic language, normalisation involves the following steps:

a) Replace some Arabic characters such as (أ, إ, آ) with (ا), (ى) with (ي), (ة) with (ه), and (ؤ) with (و).
b) Delete duplicate characters. For example, (خياااااااال) was changed to (خيال).
c) Delete the (ـــ) that some Arabic text contains, such as (تفـــاؤل), which was changed to (تفاؤل).

3.2.4. Stemming

Stemming is a linguistic normalisation process in which all derived words are converted to their base or stem forms. In a non-Arabic language, the stem form of the words can be obtained by removing either prefixes or post-fixes of the derived word. To obtain the stem form of an Arabic word, the root letters of the word must be extracted [44]. For example, the words "سلام، سالمات" come from the root "سلم." Therefore, the stemming of the Arabic language is viewed as a challenging task.

3.2.5. Stop-word removal

Removing stop words is a typical step in preprocessing. Stop words are usually the most common words in languages. These words do not provide important meaning; for example, conjunctions, articles, and relational words are stop words. Removing stop words helps in recognizing the most important words. The Arabic stop words used in this study were defined in a list available in [45-47]. This was with the exception of negations, which were deleted.

3.3. Feature extraction

Features engineering refers to generating metrics for the analysis process based on the dataset. Most feature engineering techniques create a large number of features that represent the data. However, some of these features are irrelevant and result in degrading the performance of text classifiers. Feature selection techniques choose a subset of a total number of features to eliminate redundant features [48]. After features are selected, they must be extracted to numerical form for the analysis. These features can then be input to the machine learning algorithms. Using a convenient feature extraction technique can improve the performance of text classifiers [49, 50].

Many models have been used for feature extraction, such as BoW, term frequency-inverse document frequency (TF-IDF), and N-gram models. The BoW was found to be the most common model used in the literature. In the BoW model, tweets are represented as vectors containing words. In the BoW model, the order of the words in the sentence is ignored and words frequency is counted. As another model, TF-IDF represents a normalised count of the words in which the count of each word is divided by the number

of tweets in which the word appears. Finally, N-gram models aim to break text into a sequence of words depending on a specified range. For instance, N-gram models spelt each word as a unique gram to form Unigram feature.

In this study, TF-IDF and BoW were used with six ranges of the N-gram model to investigate the best scenarios for the collected dataset. The used ranges are shown in Table 1. A single word was considered a unigram; bigrams represented two consecutive words, and trigrams were three successive words. Consequently, the words of texts were separated according to the selected ranges of N-gram features.

3.4. Algorithm selection for emotion classification

Since EA is a type of text classification, the most common text classification algorithms used were the NB, SVM, and LR classifiers. These algorithms were trained using a different range of N-gram features, depending on BoW and TF-IDF. The NB and LR are probabilistic classifiers that provide a probability distribution over output categories. On the other hand, SVM does not provide probabilistic values. Instead, it provides return decision scores, which are proportional to the distance from the separating hyperplane. The results generated according to the following steps:

a) Use the N-gram model to produce six combinations of N-gram features, as illustrated in Table 1.
b) Create BoW features with six ranges of N-gram features.
c) Construct a TF-IDF feature with six ranges of N-gram features.
d) Generate a test set and training set using cross-validation.
e) Fit NB, SVM, and LR algorithms with all previous features' forms.
f) Evaluate the algorithm's accuracy using various performance measures.

Table 1. Ranges of N-gram features in the experiments

#	N-grams
1	Unigram
2	Bigrams
3	Trigrams
4	Unigrams, bigrams
5	Bigrams, trigrams
6	Unigram, bigrams, trigrams

4. THE EXPERIMENTAL RESULTS

Before the data analysis, the collected dataset contained 4000 tweets. After data cleaning and normalisation, 3171 tweets were included. The distribution of classes in the collected dataset is shown in Figure 4. Stemming reduced the features by up to 14%. In addition, stop-word removal reduced features as much as stemming did.

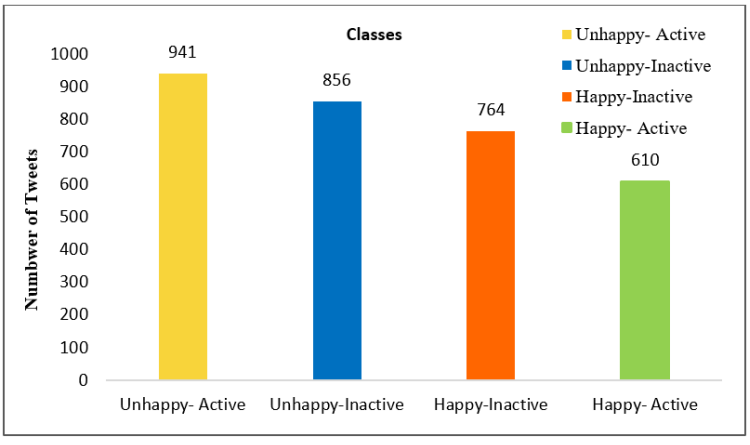

Figure 4. The distribution of emotional classes in the dataset

To analyse the impact of the feature forms on the performance of the machine learning algorithms, approximately 37 experiments conducted in different scenarios to determine the best situation of the algorithms' performance, as shown in Table 2. The results of the experiments were subjected to evaluation for a comparative analysis of the performance of the classification algorithms. The accuracy,

precision, recall, and *F1*-score were used to define the best algorithm and feature construction method. Accuracy was a ratio of correctly classified tweets according to all tweets. Precision represented the percentage of the relevant instances according to actual classes, and recall was the rate of the total relevant results correctly predicted. The *F1*-score revealed the harmonic mean of precision and recall.

Table 2. The experimental results

#	Classifier	The Experiments N-gram	Feature extraction technique	Results Validation Accuracy (%)	Precision (%)	Recall (%)	*F1*-score (%)
1	SVM	Unigram	BoW	82.27	83.47	82.27	82.72
2		Bigrams		34.31	53.65	34.31	53.65
3		Trigrams		30.55	59.97	30.55	15.40
4		Unigrams, bigrams		82.43	83.59	82.43	82.87
5		Bigrams, trigrams		34.31	53.65	34.31	22.91
6		Unigram, bigrams, trigrams		82.43	83.59	82.43	82.87
7	NB	Unigram		74.07	73.05	74.07	73.21
8		Bigrams		34.72	51.29	34.72	25.77
9		Trigrams		30.77	59.99	30.77	15.85
10		Unigrams, bigrams		73.79	72.87	73.79	73.10
11		Bigrams, trigrams		34.56	51.03	34.56	25.52
12		Unigram, bigrams, trigrams		73.60	72.82	73.60	73.07
13	LR	Unigram		82.08	82.93	82.08	82.44
14		Bigrams		34.81	50.86	34.81	24.80
15		Trigrams		30.77	59.99	30.77	15.85
16		Unigrams, bigrams		82.21	83.05	82.21	82.57
17		Bigrams, trigrams		34.81	50.86	34.81	24.80
18		Unigram, bigrams, trigrams		82.21	83.06	82.21	82.57
19	SVM	Unigram	TF-IDF	77.41	78.70	77.42	77.33
20		Bigrams		34.19	52.50	34.18	22.83
21		Trigrams		30.56	59.97	30.56	15.40
22		Unigrams, bigrams		77.14	78.42	77.14	77.02
23		Bigrams, trigrams		34.22	52.69	34.22	22.84
24		Unigram, bigrams, trigrams		77.17	78.48	77.17	77.05
25	NB	Unigram		70.55	69.75	70.55	68.65
26		Bigrams		34.94	51.43	34.94	25.79
27		Trigrams		30.78	59.99	30.79	15.85
28		Unigrams, bigrams		70.99	70.16	70.98	69.16
29		Bigrams, trigrams		34.88	51.32	34.88	25.64
30		Unigram, bigrams, trigrams		70.89	70.11	70.89	69.08
31	LR	Unigram		80.60	81.08	80.61	80.82
32		Bigrams		34.75	50.32	34.52	24.78
33		Trigrams		30.78	59.99	30.78	15.85
34		Unigrams, bigrams		80.89	81.36	80.89	81.1
35		Bigrams, trigrams		34.79	50.84	34.78	24.84
36		Unigram, bigrams, trigrams		80.86	81.33	80.86	81.06

5. RESULTS DISCUSSION

Based on the experiments, the algorithms were compared and evaluated vertically, with different feature forms, and horizontally, with each other. It was observed that when a unigram was present in N-gram combinations, this produced better results compared to groups that did not include a unigram as a feature. Generally, in terms of algorithms, the results of SVM somewhat surpassed the results of other algorithms in both accuracy and *F1*-scores. On the other hand, the NB algorithm produced the worst results for any circumstance of features. Further analysis showed that each algorithm required appropriate features in order to perform its best performance.

Figure 5 shows that SVM and LR reported their best performance with (Unigrams, bigrams) or (Unigrams, bigrams, trigrams) and BoW features. However, NB showed the greatest performance with unigram and BoW features. In the case of SVM, the best accuracy and *F1*-scores were produced with any N-gram combination that contained unigrams. The highest accuracy and *F1*-score of SVM were 82.43 and 82.72, respectively; these were obtained with BoW and (Unigrams, bigrams) or (Unigram, bigrams, trigrams) as features combinations. On the other hand, trigrams negatively affected the performance of SVM in all feature extraction techniques used. The NB classifier achieved the best results with BoW and unigram features, for which the accuracy and *F1*-score were 74.07 and 73.21, respectively. As in the performance of SVM, the accuracy and *F1*-scores significantly dropped with trigrams in both BoW and TF-IDF. The impact of N-gram features and feature extraction on LR was similar to SVM. The best accuracy for this method was

73.79 with (Unigrams, bigrams) in BoW, and the highest *F1*-score achieved was 73.07 in BoW with a (Unigrams, bigrams, trigrams) feature combination. In comparing the performance and the algorithms in features extraction angle, the best performance in BoW was achieved by SVM, which reached 82.43 in accuracy and 82.87 in *F1*-score. The LR algorithm was second, followed by NB, which reached an accuracy and *F1*-score of 30.77 and 15.85, respectively. Based on the features extraction techniques applied in the experiments, the results of the three algorithms working with BoW were better than the results with TF-IDF. In contrast, the LR algorithm provided a distinct performance with TF-IDF compared to other algorithms. For TF-IDF, the accuracy reached a peak of 80.98 for LR, and the *F1*-score reached 81.6, which was considered the best performance. The worst performance in TF-IDF was caused by NB, where the accuracy was 30.78 and *F1*-score was 15.85.

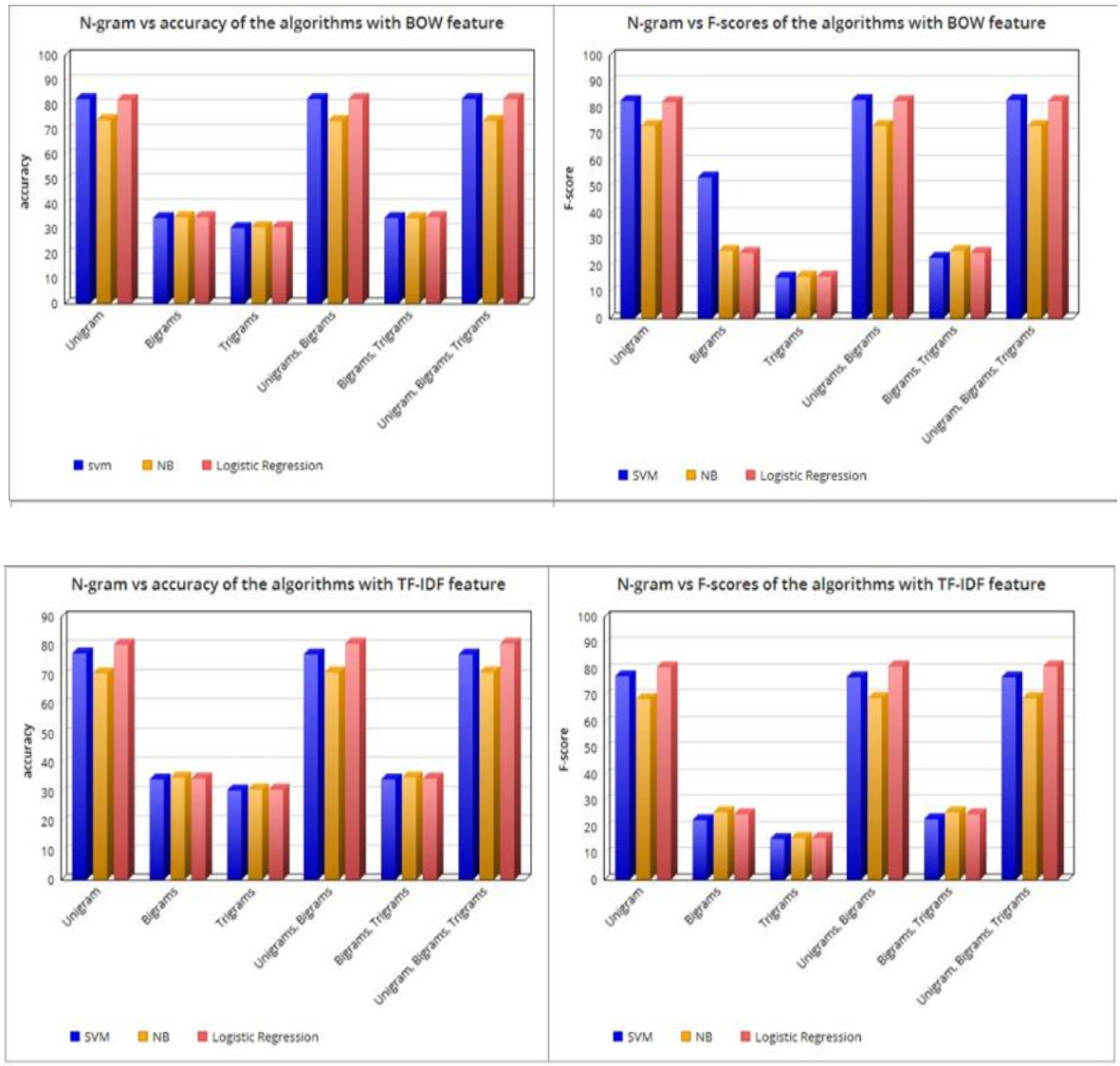

Figure 5. The accuracy and *F1*-score for each N-gram by feature extraction techniques vs. classifiers

As seen in Table 2, feature extraction methods affect the performance of the classifiers. Thus, using good classifiers with non-convenient feature methods produces poor results. For instance, the accuracy of SVM with BoW and bigrams reached 30.55%. The impacts of feature methods on the performance of the classifiers have been examined by different researchers. Azim et al. [26] compared the performance of SVM, NB, and ANN with BoW and bigrams, separately. Their best result was found with SVM with bigrams, for which accuracy was 77%. Similarly, the present study's experiments showed that SVM could reach an accuracy of 82% using BoW and (Unigrams, bigrams). Moreover, Abdullah et al. [14] used only bigrams to

detect emotions in Arabic tweets. In their work, the SVM classifier with TF-IDF performed better than other methods; its accuracy reached 80.6%. The result of the SVM classifier in the present study was better than in the study accomplished by [14] when using BoW and any combination of N-gram that contained unigrams. Therefore, researchers should conduct multiple experiments with BoW, because it was more useful than TF-IDF for detecting emotions in Arabic tweets. The best performance results are shown in Figure 6.

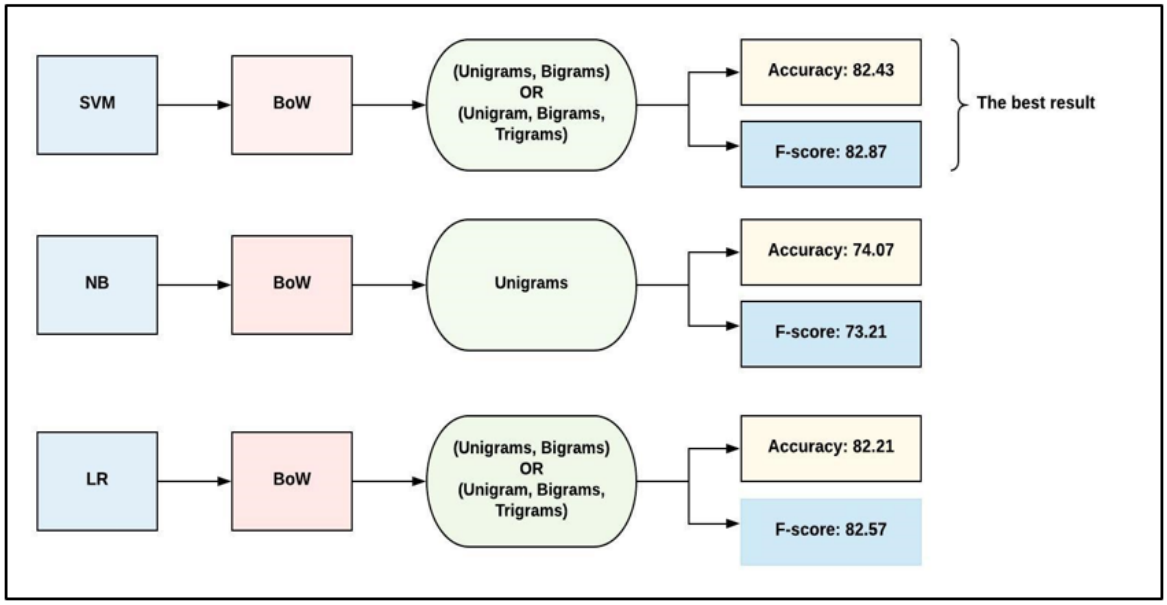

Figure 6. The best performance scenarios

6. CONCLUSIONS AND FUTURE WORK

EA is a text classification system that aims to identify human feeling conveyed through texts. In recent years, Arab users have expressed their emotions and attitudes on many of the issues raised through the Twitter platform. Therefore, this study focused on the most popular classification algorithms such as SVM, NB, and LR and applied them to a tweet's dataset as short text content. During the experiments, the impact of BoW and TF-IDF, as well as N-gram features, on these algorithms were investigated to determine the best method. The results showed that BoW performed better than TF-IDF in all cases. Moreover, the unigram feature from the N-gram model outperformed any combination of N-gram features that excluded the unigram feature. Among the different combinations of features and algorithms, SVM and LR achieved the best performance with (Unigrams, bigrams) or (Unigrams, bigrams, trigrams) and BoW features. NB achieved the lowest performance in all conducted experiments.

The best result was achieved by the SVM classifier when using BoW with unigrams and bigrams or unigrams, bigrams, and trigrams for classifying the tweets written in Arabic as a complex language. According to these results, Arabic psychiatric clinics can explore the emotional states of their patients automatically by using the best model, as shown in Figure 6. Additionally, the analysis results could provide valuable knowledge for many applications in different areas. For example, the findings could provide valuable knowledge for the economic, education, security sectors and other sectors through knowledge extraction to support decision-making. Furthermore, the results could provide a method to explore the opinions and impressions of people written in Arabic about services provided or products offered, there by increasing sales and profits by improving the quality of the products or of the services provided to customers. Many methods could be used to extend this research in the future. One of them could be capturing emoticons, which are usually used to convey the writer's emotions or intended tone. Furthermore, a study with more focus on the emotion classes in a specific field would provide a clearer picture of people's emotional states, opinions, and suggestions based on the best classifier identified by this present study. Advanced research should be done to illuminate why some features work well with some algorithms and poorly with others.

REFERENCES

[1] D. Spina, A. Zubiaga, A. Sheth, and M. Strohmaier, "Processing social media in real-time," *Information Processing & Management*, vol. 56, no. 3, pp. 1081–1083, 2019

[2] J. D. G. Paule, Y. Sun, and Y. Moshfeghi, "On fine-grained geolocalisation of tweets and real-time traffic incident detection," *Information Processing & Management*, vol. 56, no. 3, pp. 1119–1132, 2019.

[3] M. Hasan, M. A. Orgun, and R. Schwitter, "Real-time event detection from the Twitter data stream using the TwitterNews+ Framework," *Information Processing and Management*, vol. 56, no. 3, pp. 1146–1165, 2019.

[4] A. Javed, P. Burnap, and O. Rana, "Prediction of drive-by download attacks on Twitter," *Information Processing and Management*, vol. 56, no. 3, pp. 1133–1145, 2019.

[5] M. Dragoni, M. Federici, and A. Rexha, "An unsupervised aspect extraction strategy for monitoring real-time reviews stream," *Information Processing and Management*, vol. 56, no. 3, pp. 1103–1118, 2019.

[6] S. Sangam and S. Shinde, "Sentiment classification of social media reviews using an ensemble classifier," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 1, pp. 355–363, 2019.

[7] I. Perikos and I. Hatzilygeroudis, "Recognizing emotions in text using ensemble of classifiers," *Engineering Applications of Artificial Intelligence*, vol. 51, pp. 191–201, 2016.

[8] S. Wilson and R. Sivakumar, "Twitter data analysis using hadoop ecosystems and apache zeppelin," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 16, no. 3, pp. 1490-1498, 2019.

[9] H. Becker, D. Iter, M. Naaman, and L. Gravano, "Identifying content for planned events across social media sites," Proc. fifth ACM Int. Conf. Web search data Min. - WSDM '12, no. 533, p. 533, 2012.

[10] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?," *Proc. 19th Int. Conf. World wide web - WWW '10*, p. 591, 2010.

[11] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani "Evaluation datasets for Twitter sentiment analysis : A survey and a new dataset, the STS-Gold," *Conference: Workshop: Emotion and Sentiment in Social and Expressive Media: approaches and perspectives from AI (ESSEM) at AI*IA Conference*, At Turin, Italy, 2013.

[12] Statista, "Countries with most Instagram users 2019," Statista, 2019. [Online]. Available: https://www.statista.com/statistics/578364/countries-with-most-instagram-users/.

[13] L. Wikarsa and S. N. Thahir, "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method," *2015 1st International Conference on Wireless and Telematics (ICWT)*, 2016.

[14] M. Abdulllah, M. O. Almasawa, I. S. Makki, M. J. Alsolmi, and S. S. Mahrous, "Emotions classification for Arabic tweets," *International Journal of Computers and Applications*, pp. 1–15, 2018.

[15] M. N., I. M., A. H., and H. A., "Opinion mining and analysis for Arabic language," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181–195, 2014.

[16] M. Hasan, E. Rundensteiner, and E. Agu, "EMOTEX: Detecting emotions in twitter messages," *2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*, 2014.

[17] O. Badarneh, M. Al-Ayyoub, N. Alhindawi, L. A. Tawalbeh, and Y. Jararweh, "Fine-grained emotion analysis of Arabic tweets: A multi-target multi-label approach," *2018 IEEE 12th International Conference on Semantic Computing (ICSC)*, pp. 340–345, 2018.

[18] A. Shukla, S. Shukla, "A survey on sentiment classification and analysis using data mining," *International Journal of Advanced Research in Computer Science*, vol. 6, no. 7, pp. 20-25, 2015.

[19] D. Mowery, C. Bryan, and M. Conway, "Feature Studies to Inform the Classification of Depressive Symptoms from Twitter Data for Population Health," *arXiv:1701.08229v1 [cs.IR]*, pp. 0–4, 2017.

[20] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.

[21] W. Yang and L. Mu, "GIS analysis of depression among Twitter users," *Applied Geography*, vol. 60, pp. 217–223, 2015.

[22] S. Badugu and M. Suhasini, "Emotion Detection on Twitter Data using Knowledge Base Approach," *International Journal of Computer Applications*, vol. 162, no. 10, pp. 28-33, 2017.

[23] S. M. Mohammad and F. Bravo-Marquez, "Emotion intensities in Tweets," *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pp. 65-77, 2017.

[24] S. Jain and K. Asawa, "EMIA: Emotion model for intelligent agent," *Journal of Intelligent Systems*, vol. 24, no. 4, pp. 449–465, 2015.

[25] M. Hasan, E. Rundensteiner, and E. Agu, "Automatic emotion detection in text streams by analyzing Twitter data," *International Journal of Data Science and Analytics*, 2018.

[26] M. A. Azim and M. H. Bhuiyan, "Text to emotion extraction using supervised machine learning techniques," *TELKOMNIKA (Telecommunication Computuper Electronic Control)*, vol. 16, no. 3, p. 1394-1401, 2018.

[27] M. Thelwall, "TensiStrength: Stress and relaxation magnitude detection for social media texts," *Information Processing & Management*, vol. 53, no. 1, pp. 106–121, 2017.

[28] J. Karoui, F. B. Zitoune, and V. Moriceau, "SOUKHRIA: Towards an irony detection system for Arabic in social media," *Procedia Computer Science*, vol. 117, pp. 161–168, 2017.

[29] H. M. Abdelaal, A. N. Elmahdy, A. A. Halawa, and H. A. Youness, "Improve the automatic classification accuracy for Arabic tweets using ensemble methods," *J. Electr. Syst. Inf. Technol.*, no. 2017, pp. 1–8, 2018.

[30] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.

[31] H. Wang, L. Wang, and L. Yi, "Maximum entropy framework used in text classification," in *Proceedings - 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems*, pp. 828–833, 2010.

[32] W. J. Long, J. L. Griffith, H. P. Selker, and R. B. D'agostino, "A comparison of logistic regression to decision-tree induction in a medical domain," *Comput. Biomed. Res.*, vol. 26, no. 1, pp. 74–97, 1993.
[33] S. Dumais, J. Platt, D. Heckerman, and M. Sahami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pp. 148–155, 2004.
[34] A. Kehagias, V. Petridis, V. G. Kaburlasos, and P. Fragkou, "A comparison of word- and sense-based text categorization using several classification algorithms," *Journal of Intelligent Information Systems*, vol. 21, no. 3, pp. 227–247, Nov. 2003.
[35] F. Colas and P. Brazdil, "On the Behavior of SVM and Some Older Algorithms in Binary Text Classification Tasks," in *International Conference on Text, Speech and Dialogue*, pp. 45–52, 2006.
[36] E.-H. Han, G. Karypis, and V. Kumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," in *PAKDD '01: Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 53–65, 2001.
[37] A. M. Kibriya, E. Frank, B. Pfahringer, and G. Holmes, "Multinomial Naive Bayes for Text Categorization Revisited," in *AI 2004: Advances in Artificial Intelligence*, pp. 488–499, 2004.
[38] C. C. Aggarwal and C. X. Zhai, "A survey of text classification algorithms," in *Mining Text Data*, vol. 9781461432, C. C. Aggarwal and C. Zhai, Eds. Boston, MA: Springer US, pp. 163–222, 2012.
[39] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *ECML 1998: Machine Learning*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, vol. 1398, pp. 137–142), 1998.
[40] M. A. H. Madhfar and M. A. H. Al-Hagery, "Arabic text classification: A comparative approach using a big dataset," in *2019 International Conference on Computer and Information Sciences (ICCIS)*, 2019.
[41] R. Xu, T. Chen, Y. Xia, Q. Lu, B. Liu, and X. Wang, "Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification," *Cognitive Computation*, vol. 7, no. 2, pp. 226–240, 2015.
[42] "About – Netlytic.org."
[43] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
[44] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Computer Speech & Language*, vol. 28, no. 1, pp. 20–37, 2014.
[45] K. Bouzoubaa, H. Baidouri, T. Loukili, and T. El Yazidi, "Arabic Stop Words : Towards a Generalisation and Standardisation," *Proc of the 13th International Business Information Management Association Conference IBIMA 2009*, 2009.
[46] I. A. El-khair, "Effects of stop words elimination for arabic information retrieval : A comparative study," *International Journal of Computing & Information Sciences*, vol. 4, no. 3, pp. 119–133, 2006.
[47] A. Alajmi, E. M. Saad, and R. R. Darwish, "Toward an Arabic stop-words list generation," *International Journal of Computer Applications*, vol. 46, no. 8, pp. 8–13, 2012.
[48] J. D. Prusa, T. M. Khoshgoftaar, and D. J. Dittman, "Impact of feature selection techniques for tweet sentiment classification," *Twenty-Eighth Int. Flairs Conf.*, pp. 299–304, 2015.
[49] Y. Wang, Z. Zhou, S. Jin, D. Liu, and M. Lu, "Comparisons and selections of features and classifiers for short text classification," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 261, pp. 1-7, 2017.
[50] J. A. Banados and K. J. Espinosa, "Optimizing support vector machine in classifying sentiments on product brands from twitter," *in IISA 2014 - 5th International Conference on Information, Intelligence, Systems and Applications*, pp. 75–80, 2014.