A two-level deep learning approach for emotion recognition in Arabic news headlines

Hoda Ahmed Galal Elsayed , Soumaya Chaffar , Samir Brahim Belhaouari & Hafsa Raissouli

Published online: 25 Nov 2020.

Submit your article to this journal ⬈

Article views: 395

View related articles ⬈

View Crossmark data ⬈

Taylor & Francis
Taylor & Francis Group

A two-level deep learning approach for emotion recognition in Arabic news headlines

Hoda Ahmed Galal Elsayed[a], Soumaya Chaffar[b], Samir Brahim Belhaouari[c] and Hafsa Raissouli[d]

[a]Alfaisal University, Riyadh, Saudi Arabia; [b]Department of Science and Technology, TELUQ University, Montreal, Canada; [c]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar; [d]College of Computer Science and Information Technology, University Putra Malaysia, Malaysia

ABSTRACT

Going online has created more opportunities for newspapers to present breaking news in a timely manner. Concentrating in spreading more bad news increases the feeling of danger and depression in the society. Some authors believe on tendency of some media to be focused on sharing the bad events in life rather than the good ones because of the impact and the attraction over the audience is more significant. Sentiment analysis work has not been recognized, proposed, or documented on Arabic news because of the challenges that Arabic raises as a language including the different Arabs' dialects and its complex grammatical structure. With the emerging flow of news that cause a global panic and anxiety worldwide, this study focuses on the serious need to identify what effective role could machine learning classifiers have in the early detection process of the psychology impact on the readers by the daily news headlines. In this work, a dataset of news headlines were gathered from top popular online Arabic news sites and were annotated to seven emotional categories: anger, disgust, fear, happiness, neutral, sadness, and surprise. A convolutional neural network-based two-level approach was proposed for sentiment classification. The performance of the proposed approach was compared to six machine learning classifiers (zeroR, k-nearest neighbor, decision trees, naïve Bayes, random forest and Support vector machine) and showed better accuracy, precision, and recall.

1. Introduction

News represent an essential domain in developing countries as they help in knowledge spreading and thus increase the level of public awareness of the surrounding world. In the past few decades, the impact of news has reached a wide range of the public in both positive and negative manner. Many studies [1–3] suggested that the negative news can develop negative impacts on readers. Media concentrates on sharing the bad events in life rather than the good ones. These impacts can easily provoke marked changes in the personality at the subjective, physical, and physiological levels and thus can surpass ones' productivity and creativity [4,5]. Psychological impacts include for instance: losing concentration, anxiousness, stress, and negative thoughts which can lead to an advanced level of nervousness, depression, and fear [6–9]. In addition to their effect on the humans' moods, some news was found to be engineered to interrupt the reader's level of mental concentration [3]. Hence, continuous exposure to such type of news can affect readers' memories at the long term [6]. News has also some physical impacts that include for example digestion problems, hair, cells and bones growth problems, susceptibility to infections [8]. Daily exposure of repeated news may also turn readers to be passive toward their own daily life problems [9].

Accordingly, it is important for news publishers to realize their responsibility in providing a good basis to reflect everyday events in such a way that does not hurt the public. Thus, analyzing the severity level of published news is an essential step that can help in enhancing the news headlines' tones positively. Sentiment Analysis is one way to do such analysis and, yet it is considered so powerful in estimating how readers may feel towards the published news. Sentiment analysis serves as a process of identifying whether specific subjective context expresses positive, neutral, or negative feelings

Figure 1. Words generated from the root 'read' in Arabic.

through extensive syntactic and semantic patterns recognition [10]. Emotions evolving from reading news can be classified into various categories [11].

The world witnessed a technological revolution recently characterized by the involvement of Artificial Intelligence (AI) in analyzing the existing data and deriving actionable insights accordingly [12]. Many researchers spot the light on the sentiment analysis that uses natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Compared to English work, few researches performed sentiment analysis on Arabic data and its dialects. We believe that there is still a great need for contribution to Arabic sentiment analysis mainly because of the challenging nature of Arabic text mining. Arabic is a morphologically rich language and based on root pattern structure. A single Arabic root can generate hundreds of words (noun, verbs). For example, the word read as shown in Figure 1 can generate many words such as: he read, her reading, and he reads. News headlines are different from the regular tweets in the sense they allow explicitly subjective content and hence, are difficult to classify and process. Sentiment analysis of news headlines that are written in Arabic which is considerably new in this field is a challenging task due to the complex morphological, structural, and grammatical nature of Arabic.

This research investigates the effectiveness of using sentiment analysis in classifying the severity level of Arabic news headlines on readers' emotions. Accordingly, 1698 news headlines were collected from top news' website sources and annotated by three unbiased persons on rounds, whereas machine learning (ML) techniques were used to train different classifiers for emotions classification (e.g. anger, sadness, happiness, disgust, fear, surprise, or neutral). The issue that remains under scrutiny in this study is 'how different ML will perform while classifying the readers' emotions based on the collected Arabic news'.

This paper is organized as follows: Section 2 presents the related work that served as a reference for this study. Section 3 presents the adopted methodology of emotions extraction, feature set, emotional classification, as well as the data collection process and undertaken data preparation stages. Section 4 summarizes the classification results, obtained through this study. Finally, Section 5 presents the conclusion to our study and findings as well as study limitations and possible future work.

2. Arabic sentimental analysis

In the literature, studies that have been performed on Arabic language were mainly focused on social media. Most of the early work in Arabic sentiment analysis considered data collected from Twitter, Facebook, or YouTube regarding different topics.

A study done by Alaa et al. in 2013 [13] analyzed the level of agreement and disagreement based on users' comments on political news posts on Facebook by developing a corpus for sentiment analysis and opinion mining using three types of ML techniques. These techniques are: (1) Decision Tree (DT), (2) Support Vector Machines (SVM), and (3) naïve Bayes (NB). Their system was designed to preprocess the trained data by ignoring all stop-words that do not affect the meaning of the phrases written, as well as the very long comments, special characters (#, @, !, %) and meaningless redundant letters. The authors used data from posts on the following news pages on Facebook: (Rassd, Silvio Costa, 6 April, we are all Khaled Said, official page for the presidency of the Council of Ministers, Egypt News Network, Tahrir Square News, Egyptian today). The classification then took place to the extracted posts to either supportive, attacking or neutral comments. It was reported that SVM achieved the best result of 73% of precision.

Another study done by Al-Ayyoub et al. in 2014 [14] evaluated users' feedback and comments related to Arabic news. The authors reported that it was a challenge due to the limited number of available datasets, collected and analyzed to apply sentiment analysis to it. In their work, Al-Ayyoub et al. used mainly the unsupervised lexicon-based approach to determine the polarity of reviews. The methodology adopted in their work was divided into two stages: (1) building a crawler to gather user comments and in this stage, they applied morphological and Khoja's stemming process [15] to preprocess comments by remove stop-words and extracting the roots of words. (2) building a sentiment lexicon in which each word has a sentiment value (positive, negative or neutral) assigned based on the given percentage (0–100)%. They collected the stems from Abuaiadh's available dataset [16] and to extend their lexicon, they used a crawler to collect articles from different Arabic news websites like Aljazeera.net. Moreover, a translator was used to convert English articles to Arabic ones. Afterwards, they built a tool to process these articles and extract distinct stems from them.

In a study done by Mona Diab in 2009 [17], the problem of processing modern standard Arabic was addressed and a second generation of tools that process Arabic, AMIRA, was presented which is a successor suite to the ASVM Tools. The AMIRA toolkit includes a tokenizer, part of speech tagger and base phrase chunker and more

functionalities were added to the earlier version as it can take in text in different encodings and produce the output in the input encoding, accordingly. AMIRA tool was able to handle 2 main issues which caused some inaccuracy in the previous version: 'Taa Marbuta', the word final feminine marker on nominals indicating syntactic gender of the words and differentiating between 'Alef Maqsuura' and 'Alef Mad' by alternating between a 'y' and an 'A' to classify the words by final letters. This distinction is a lexical distinction and it relates to the underlying root of the word. The technology of AMIRA is based on supervised learning with no explicit dependence on knowledge of deep morphology; hence, in contrast to MADA, it relies on surface data to learn generalizations. That study was limited to process Arabic words without subjective analysis, but it was the basic inspiration for the SAMAR system in 2012 [18].

In [19], opinions including emotions corresponding to weather, food, and sport are collected from Facebook where about 45% are standard Arabic and 55% dialect Arabic. The opinions were then identified as positive, negative, or neutral using two online tools and a third proposed tool named AOPI. Three ML techniques were used to classify the opinions and evaluate the used tools: KNN, NB, SVM. The dataset was classified stemmed and non-stemmed. The results suggested that the stemmed dataset gives better results.

In [20], the authors performed emotion classification of emojis features using two approaches. The first approach uses KNN, NB, and DT separately to classify the dataset, while the second approach divides the dataset to subsets where each subset passes through several estimators. This method, namely, conventional bagging, was constructed with the same principle but with three different classifiers at each time. Meaning that the first conventional bagging uses DT, the second uses KNN and the third uses NB.

In [21], the study tackled Facebook comments in standard and dialect Arabic. The features are obtained using n-grams and weighting schemes. Feature selection was used to reduce the dimensionality of the features. It is noticed that the feature selection maintained the performance with less features and improved it in some cases. Three ML techniques were experimented: NB, SVM, and RF, where SVM and NB showed the highest results. In [22], YouTube comments were obtained and annotated using a developed tool. The dataset was preprocessed and normalized before passing it to the three used classifiers that are SVM, KNN and NB. SVM-RBF achieved the highest accuracy that is 88%. In [23] and [24], both studies used NB and SVM to classify standard and dialect Arabic in book reviews and Facebook comments. Both studies concluded that NB outperformed SVM with the used datasets.

In [25], the authors proposed hierarchical classification approach applied using four classifiers (SVM, DT, NB, and KNN). The approach can have different number of levels where each level is meant to recognize specific classes. For example, to classify an instance in a four classes problem, the first level recognizes if the instance is class one or two or other. If the instance is assigned to other it passes to the second level that recognizes if it is class three or four. The approach used was compared to the flat normal classification using the four classifiers (SVM, DT, NB, and KNN). The results showed that KNN coupled with the hierarchical approach performed better than other techniques. In [26], the study classifies tweets retrieved from UCI repository using SVM, NB, and DT. The obtained results show that DT performs better that SVM and NB. Table 1 summarizes the related work.

Deep learning has spread widely in research community to target natural language processing and support sentiment analysis of many languages. In [27] recursive deep learning was utilized to predict sentence level of label distributions. The authors has developed a ML framework to learn vector space representations for multi-word phrases, trained on an unlabeled corpus of the English Wikipedia.

Table 1. Summary of the related work.

Authors and year	Features	Methods	Data Source	Domain
Al-Azani et al. [20]	–	• NB • KNN • *DT*	Twitter	–
Elouardighi et al. [21]	n-gram TF TF-IDF	• NB • *SVM* • RF	Facebook	Elections
Al-Tamimi et al. [22]	n-gram TF TF-IDF	• *SVM* • KNN • NB	YouTube	–
El Ariss et al. [23]	TF-IDF N-gram	• *NB* • SVM	Book reviews	–
Elouardighi et al. (2017) [21]	Unigram Bigram TF-IDF	• *NB* • SVM	Facebook	News
Nuseir et al. [25]	BOW	• SVM • DT • NB • *KNN*	Book reviews	–
Altawaier et al.[26]	TF-IDF	• SVM • NB • *DT*	Twitter	Politics and art
Al-Kabi et al. [19]	–	• KNN • NB • *SVM*	Facebook	Food, sport, and weather

Figure 2. Emotional extraction methodology.

Socheretal et al. [28] then decided in 2011 to introduce a max-margin framework applicable on natural language and computer vision tasks that relies on recursive neural networks technique for label prediction that requires labeled tree structures. In [29], the same authors used a trained recursive neural tensor network (RNTN), on a built sentiment tree bank to enhance the sentiment analysis on English datasets. In [30], a model based on deep neural network architectures was developed including convolutional neural networks (CNNs) and long-short term memory (LSTM) to achieve the first rank on all of the five English sub-tasks in SemEval-2017. Other recent deep learning methods for sentiment analysis include [31]. Similarly, the authors of [32] applied RNTN to predict the sentiment on Arabic tweets by using sentiment treebank. They reported that using deep learning has achieved the state-of-the-art results for the Arabic sentiment analysis. In [33], deep learning framework was proposed for text sentiment classification in Arabic. They used Recursive Auto Encoder and their results showed a high improvement compared to the state of the art with the advantage of using no lexicon resources that are costly.

From the literature, we conclude the following:

- Many studies address the approach of analyzing Arabic posts from social media. Most of these studies aimed to classify the data into either negative or positive but none of the studies address classification into psychological stages.
- Most of studies focus on social media users' comments, which are usually expressed in dialect Arabic that considerably differs from standard Arabic. None of the studies addressed standard Arabic only.
- Most studies used existing tools to annotate the dataset.

This study analyzes the Arabic news headlines annotated by domain experts to identify their impacts on readers' emotions. Moreover, this study is not aimed to analyze posts power of the public but rather we analyze the power authorized of trusted news providers that are accredited to share news with public.
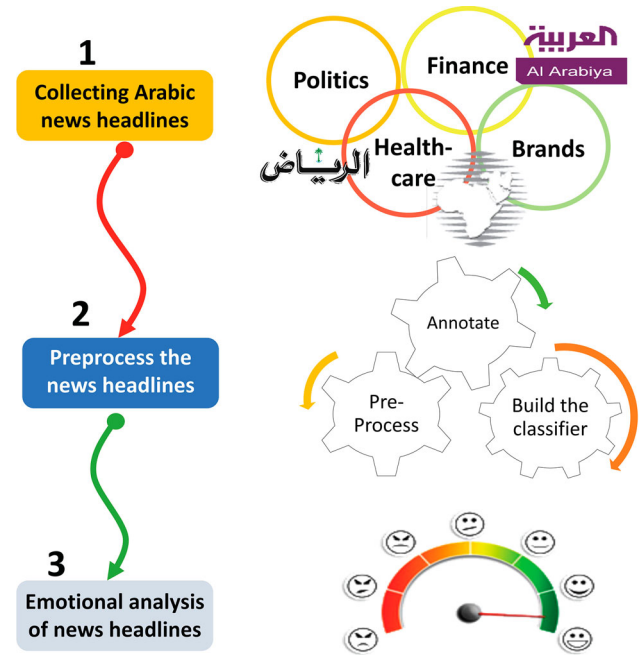
3. Emotional extraction methodology

The proposed methodology used in conducting this research is illustrated in Figure 2. The proposed approach suggests that data represented by Arabic news headlines is extracted from news websites, cleaned (e.g. remove duplicates) and annotated by three annotators on multiple rounds. Seven emotional labels are used to annotate the news headlines including the neutral label as well as the six universal emotions: disgust, anger, sadness, fear, surprise, and happiness [34]. Only headlines that two or more annotators agreed on are considered in the analysis. Emotional classification of extracted text requires choosing the relevant features to consider. In the coming sections, the possible extraction and selection features of data classifications are discussed further.

3.1. Data collection and annotation

3.1.1. Data collection

To conduct this research, a dataset [35] was collected across multiple phases from mainly some Arabic news websites using some already existing xml sheets or manual collection process between September 2017 and April 2018. Besides their being frequently updated, these sites (e.g. 'Al-Arabiya' [36], 'Al-Riyadh' [37] and 'Alsharq Alawsat' [7]) are of the top frequently visited sites by readers across the Middle Eastern countries and thus were used to extract data needed for this study.

A total of 1698 news headlines that address various news categories and many life events that took place in the past few years, were extracted accordingly. These categories exceeded 25 news classes in each website, some of which are: (1) Culture and art, (2) Arab world news, (3) Local and international banking, finance and economy, (4) International news, (5) Research, medicine and health, (6) natural resources (oil and gas), (7) Technology, and (8) Tourism. Data cleansing was applied to the extracted data afterward. Data cleansing involved removing all duplications found as well as the unnecessary details (e.g. the writer name, date, and place of publication). Data cleansing was followed by manual emotional annotation to ensure more reliable results.
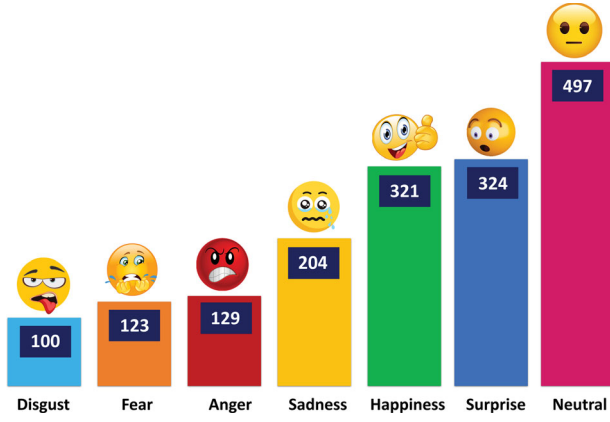
Figure 3. Visual distribution of annotators' emotions on news.

3.1.2. Data annotation

The manual emotional annotation to the extracted news headlines was carried out with the help of three unbiased annotators on multiple rounds of data annotation over time. Emotional classes were identified in isolation with each of them to be one of the following categories: happiness, neutral, sadness, anger, fear, surprise, and disgust for each news headline. The majority voted class was then considered as a final decision to serve as a the label of the dataset instance. The final emotional classes assigned to the 1698 news headlines, based on the annotation stage, were distributed as shown in Figure 3.

In order to measure the reliability of agreement between the three annotators, the Kappa score is considered. It is a measure of the ability of several annotators to make the same decision in labeling the dataset instances. There are two known Kappa scores namely, Cohen Kappa and Fleiss' Kappa. Cohen Kappa is used when there are two annotators only while Fleiss' Kappa is the generalized form that is used for any numbers of annotators. Fleiss' Kappa is calculated according to the formula in Equation (1)

$$k = \frac{P_0 - P_e}{1 - P_e} \qquad (1)$$

where P_0 is the agreement between the annotators and P_e is the hypothetical probability that the annotators have agreed by chance. If the annotators totally agree, then we obtain $k = 1$ whereas a $k \leq 0$ indicates no agreement between the annotators. The terms P_0 and P_e are defined as in Equations (2) and (3)

$$P_0 = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{n(n-1)} \sum_{j=1}^{k} (n_{ij}^2 - n_{ij}) \right) \qquad (2)$$

$$P_e = \sum_{j=1}^{k} P_j^2 \qquad (3)$$

where N is the number of dataset samples, n is the number of annotators, and k is the number of classes. The term P_0 can be simplified and calculated in two steps by first computing P_i that is defined in Equation (4) and then obtaining the mean of P_i as in Equation (5).

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1) \qquad (4)$$

$$P_0 = \frac{1}{N} \sum_{i=1}^{N} P_i \qquad (5)$$

Calculating the Fleiss' Kappa in our dataset leads to the term in Equation (6)

$$k = \frac{0.802 - 0.198}{1 - 0.198} = 0.753 \qquad (6)$$

The interpretation of the Fleiss' Kappa score is known to be as follows:

- < 0 : Agreement less than chance
- $0.01 - 0.2$: Slight agreement
- $0.21 - 0.4$: Fair agreement
- $0.41 - 0.6$: Moderate agreement
- $0.61 - 0.8$: Substantial agreement
- $0.81 - 0.99$: Almost perfect agreement

According to the interpretation of the Fleiss' Kappa score, the value obtained for our dataset that is 0.753 represents a substantial agreement between annotators.

3.2. Data preparation

After the data collection and annotation phase, the dataset underwent two phases which are (1) preprocessing and feature extraction, then (2) classification (as shown in Figure 2). In the preprocessing phase, Unicode signs were removed including all comas, single and double quotes, plus, minus, equal, percent signs, colons, and full stops. In this phase, the following unsupervised filters were adopted, respectively, which are: (1) Reorder, (2) ReplaceMissingValues, (3) Standardize, (4) StringToNominal, and (5) StringToWordVector.

Reorder filter was used to randomize the attributes order and thus generate the output with the newly given order. This was useful in making a better combination of data in the training dataset. Missing nominal values were replaced with the modes and means of the training dataset using the ReplaceMissingValues filter. Standardization filter was used on the nominal Arabic news sample to normalize all the whitespace so that multiple spaces in between words are normalized to a single space character. StringToNominal filter was used to convert the string attributes (e.g. a given year in the news headline) to nominal values. StringToWordVector filter was used to convert string attributes into a set of attributes representing word occurrence (depending on the tokenizer used) information from the text contained in the strings. In this paper, n-gram Tokenizer was used to break the Arabic text into small words with a min vector distribution as unigram (to account for the Arabic prepositions) and max of trigrams. In addition, Snowball Stemmer, a small string processing language designed for creating stemming algorithms to retrieve information was used.

After preprocessing the annotated data (saved in .csv format), the dataset was divided into training and testing sets to perform classification. A 50-fold cross-validation was used on the training set on Weka tool. n-fold cross-validation basically divides the data into n-fold, using a single fold for testing and the other n-1 for training, while it continues to do so until no more folds are found. Different classifiers and algorithms were used to obtain the best possible classification performance as discussed earlier. In the coming section, the resulting performance of these classifiers is compared and discussed further.

3.2.1. Data validation and key performance measures

After annotation and cleansing phases, n-gram features and BOW features are applied to extract features. Different classifiers, RF, Bayesian techniques, DT, ZeroR, SVM and CNN with multilevel

$$
\begin{array}{c c}
 & D_0 \;\; D_1 \;\; D_2 \;\cdots\; D_{n-1} \;\; D_n \\
\begin{array}{c}
W_0 \\ W_1 \\ W_2 \\ \vdots \\ W_{m-2} \\ W_{m-1} \\ W_m
\end{array}
&
\begin{bmatrix}
1 & 0 & 1 & \cdots & 0 & 0 \\
0 & 0 & 1 & \cdots & 0 & 1 \\
0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & 0 & 1 & \cdots & 1 & 1 \\
1 & 1 & 0 & \cdots & 1 & 0 \\
0 & 1 & 0 & \cdots & 0 & 0
\end{bmatrix}
\end{array}
$$

Figure 4. BOW features representation.

architecture of deep learning, are used. These classifiers were chosen in specific as they proved high performance and accuracy rates in many similar studies [21–23,26].

3.3. Feature extraction

Many features can be used for classifying news headlines:

(1) Bag-Of-Words (BOW): headlines could be represented by a feature vector composed of Boolean attributes for each word that occurs in the headline. If a word occurs in each headline D, its corresponding attribute w is set to 1; otherwise it is set to 0 (as shown in Figure 4). BOW considers words as independent entities. It does not consider any semantic information from the text.

(2) N-grams : sequences of words of length n that can be used for catching syntactic patterns in text and may include important text features such as negations, e.g. 'not happy'. Negation is an important feature for the analysis of emotion in text because it can totally change the expressed emotion of a sentence. Like 'not' in English, Arabic language supports the negation given usually in form shown in Figure 5 As an example, for negation from the corpus: "A cure you may not expect to defeat cancer" is shown in the Figure 5.

(3) Stems: Stemming is the process for transforming words in the BOW representation into their stem, i.e. by removing terminations. For example, for the word 'sadness' is reduced to its stem 'sad'. With such a method, the features are generalized for classifying new sentences. However, reducing some words would result in incorrect feature, for example, transforming 'fearless' to fear. Arabic stems are no different compared to English as it removes all suffixes and prefixes from the Arabic word before deriving its root. Arabic stems remove the prefix and ignores the ending suffix. Examples of prefixes and suffixes are shown in Figure 5 [38,39]. As an example, the Arabic news headline from the corpus: 'Kuwait and Egypt sign an oil contract that is worth 4 billion dollars' will appear without pre/suf-fixes as shown in Figure 5.

3.4. Classification techniques

ML techniques have been used for deriving useful knowledge from data and generalizing it for classification purposes [40]. ML is either supervised or unsupervised wherein the latter, the structure of the data is the key concern [41,42]. Supervised learning is being trained through examples in terms of data structure over several iterations. In unsupervised learning, however, no desired output is provided therefore categorization is done. As we note from the literature, SVM was proven effective in [19,21] and [22], DT outperformed other classification techniques in [20] and [26], NB outperformed SVM in [23]

Negation in Arabic language uses:

لا، لن، لم ... etc.

علاج قد **لا** تتوقعه لهزيمة السرطان

"A cure you may not expect to defeat cancer"

Prefixes in Arabic language: ال، أ، فـ ، كـ

Suffixes in Arabic language: ـه ، هـا ، ـات ، ـان ،

ـون ، ـين ، ـة

الكويت ومصر توقعان عقدا نفطيا بقيمة 4 مليارات دولار

"Kuwait and Egypt sign an oil contract that is worth 4 billion dollars"

without pre/suf-fixes
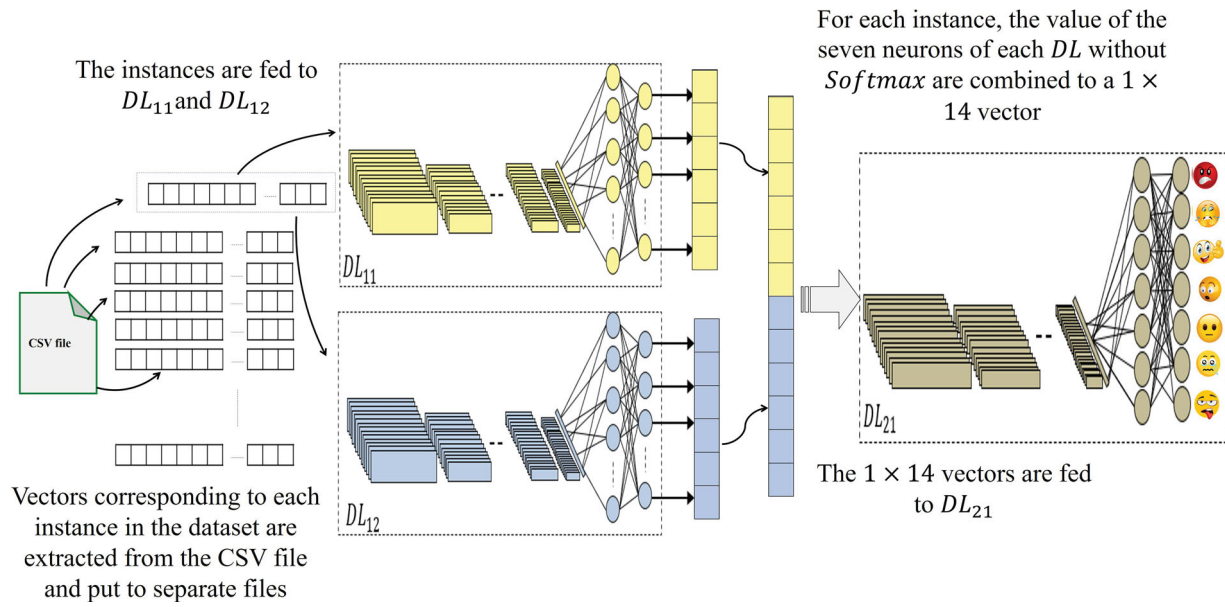
كويت ومصر وقع عقد نفط قيم 4 مليار دولار

Figure 5. Negation, prefixes, and suffixes in Arabic language with examples.

Table 2. Supervised learning techniques used.

Classifier	Description
ZeroR	Simplest classification method that relies on the target and ignores all predictors. It predicts the majority class and is used mostly as a benchmark for other techniques [43]
KNN	Classify a new instance in the test set based on the shortest distance between it and numeric neighbors (k) stored in the training set using the Euclidian Distance [44] equation. For nominal values, Hamming Distance is used to measure the minimum number of substitutions needed to change one word into the other [45]
Decision Tree	Identify a variable and the corresponding threshold, and then split the input observation into two or more subgroups in forms of binary trees [46]
naïve Bayes	Assign the given class labels to problem instances, represented by vectors of feature values, where the class labels are drawn from some finite set. It is a multiple set of algorithms based on a common principle if the value of certain feature is independent of the value of its siblings [47]
Random Forest	It takes a dataset as an input and produces n decision trees where Random set is generated for each tree. New entry classification depends on combining the results from the *n* trees generated [44,48]
SVM	This classifier transforms the data to a higher dimensional space using the kernel function and creates a classification boundary [49]
CNN	This classifier is a category that falls under neural networks. It passes filters on the dataset instances in order to identify patterns [50]

and [24], and KNN showed the highest performance in [25]. These results give the potential to experiment different methods. Hence, the adopted techniques in this study are: (1) ZeroR, (2) KNN (at K = 1, 2 and 3), (3) DT (J48, Id3), (4) Bayesian techniques (naïve, simple, updateable), (5) RF at $n = 100, 200,$ and 300, (6) SVM, and (7) CNN (with ReLU activation, shifted ReLU, and multilevel architecture of deep learning) as shown in Table 2.

Concerning the to seventh approach, regular CNN with ReLU activation and 11 convolutional blocks was trained for 25 epochs with SGD and an initial learning rate of 0.01. The second CNN used the

The instances are fed to DL_{11} and DL_{12}

For each instance, the value of the seven neurons of each DL without $Softmax$ are combined to a 1×14 vector

DL_{11}

DL_{21}

Vectors corresponding to each instance in the dataset are extracted from the CSV file and put to separate files

CSV file

DL_{12}

The 1×14 vectors are fed to DL_{21}

**Figure 6.** The two-level architecture approach used.

Table 3. Classifiers' validation accuracy and variance across the validation folds using n-grams features.

ML technique	$(1 - \text{Validation error}) * 100$	Variance of 50 folds
ZeroR	29.41	0.11
KNN ($k = 1$)	61.76	0.21
KNN ($k = 2$)	29.41	0.13
KNN ($k = 3$)	29.42	0.11
Decision Trees (J48)	61.76	0.30
Naïve Bayes	50	0.45
Naïve Bayes Simple	50	0.45
Naïve Bayes Updateable	50	0.45
Random Forests ($n = 100$)	61.76	0.33
Random Forests ($n = 200$)	61.76	0.33
Random Forests ($n = 300$)	61.76	0.33
SVM (RBF)	76.4	0.53
CNN (ReLU)	89.1	0.79
CNN (shifted ReLU)	88.3	0.82
Two-level CNN	90.4	0.53

Table 4. Classifiers' accuracy, recall and precision on test set using n-grams features.

ML techniques	Accuracy %	Recall %	Precision %
ZeroR	29.27	29.30	08.60
KNN ($k = 1$)	68.49	68.50	84.80
KNN ($k = 2$)	29.98	30.00	58.60
KNN ($k = 3$)	29.86	29.90	58.60
Decision Trees (J48)	68.49	68.50	84.80
Naïve Bayes	54.95	54.90	61.50
Naïve Bayes Simple	54.95	54.90	61.50
Naïve Bayes Updateable	54.95	54.90	61.50
Random Forests ($n = 100$)	68.49	68.50	84.80
Random Forests ($n = 200$)	68.49	68.50	84.80
Random Forests ($n = 300$)	68.49	68.50	84.80
SVM (RBF)	75.6	78.8	81.2
CNN (ReLU)	88.2	91	91.4
CNN (shifted ReLU)	87	90.9	91.1
Two-level CNN	89.3	92.5	92.6

same parameters with shifted ReLU to -0.1. The multilevel architecture of deep learning combines the output of n networks from the first level and feed them to the second level with i networks. The output of the second level can be, in turn, combined and fed to the third level and so forth till the kth level. The experimented approach here is a two-level architecture shown in Figure 6.
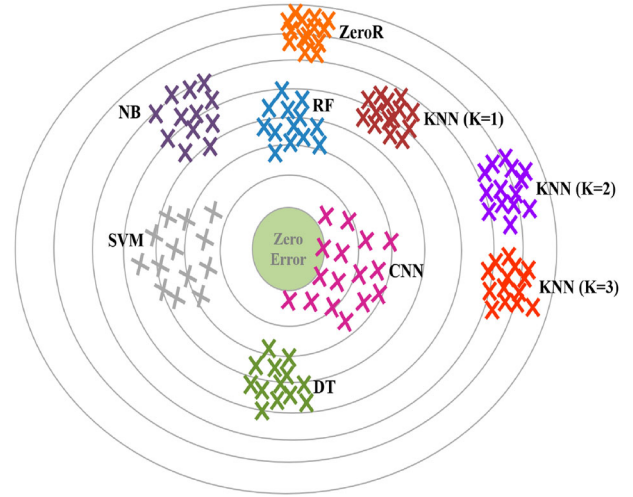
Figure 7. Interpretation of the variance/validation error of the classifiers using n-grams features.

Table 5. Classifiers' accuracy, recall and precision on test set using BOW features.

ML techniques	Accuracy %	Recall %	Precision %
CNN (ReLU)	98.2	98.84	98.17
CNN (shifted ReLU)	99.4	98.83	98.16
Two-level CNN	100	100	100

4. Results and discussion

In an earlier stage of this work, 1698 Arabic news headlines were collected from some popular news sites in the Middle East, added to a corpus and then annotated on different rounds by three annotators. However, Figure 3 elicited that the data reflected an obvious data imbalance that is caused by mainly three emotions categories which are (1) disgust, (2) fear and (3) anger annotations. The imbalance was caused due to the enormous difference in numbers between them and the other emotions expressed. Such difference may cause the classifiers to get biased while being trained on sample data. Thus,

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	
Anger	12	0	0	0	0	0	0	100%
Disgust	0	10	0	0	0	0	0	100%
Fear	0	0	12	1	0	0	0	92.3%
Happiness	1	0	0	31	0	0	1	94%
Neutral	0	0	0	0	50	0	0	100%
Sadness	0	0	0	0	0	20	0	100%
Surprise	0	0	0	0	0	0	31	100%
	92.8%	100%	100%	96.9%	100%	100%	96.9%	98.2%

Figure 8. Confusion matrix of CNN with ReLU (DL_{11}.

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	
Anger	13	0	0	0	0	0	0	100%
Disgust	0	10	0	0	0	0	0	100%
Fear	0	0	12	0	0	0	0	100%
Happiness	0	0	0	32	0	0	0	100%
Neutral	0	0	0	0	49	0	0	100%
Sadness	0	0	0	0	0	20	0	100%
Surprise	0	0	0	0	1	0	32	97%
	100%	100%	100%	100%	98%	100%	100%	99.4%

Figure 9. Confusion matrix of CNN with shifted ReLU to -0.1 (DL_{12}.

testing different classifiers' performance cannot rely on the accuracy merely, but rather other performance factors should be interfered.

These performance measures can be extracted from the confusion matrix. In the confusion matrix, the False Positives (FP) are the negative tuples that were incorrectly labeled as positive, whereas False negatives (FN) are the positive tuples that were mislabeled as negative. True Positives (TP) and True Negatives are righteous classifications. Both Recall and precision will be used for evaluating the best classifier performance rather than accuracy as it might be biased towards high-value labels. Recall can be calculated as TP / (TP+FN) while precision is calculated as TP / (TP+FP).

In order to deal with the problem of imbalanced classes and the possibility of the classifiers to be biased, increasing the number of folds can be useful. This helps to use more training data in each iteration (e.g. Leave-one-out cross-validation) which serves in lowering the bias towards estimating the generalization error and also improves the run-time and variance of the estimate as the overlap between training sets increases with an increasing size of folds. Hence, a 50-fold cross-validation is used as balancing act between bias and variance and computational efficiency. Table 3 shows the training error of the classification methods used along with the variance across the 50 folds.

The unbalanced nature of the dataset may cause some classifiers to be biased toward certain classes. The bias of a classifier refers to the difference between the model's predictions and the actual values. The variance however, indicates how the predictions of the model vary

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise	
Anger	13	0	0	0	0	0	0	100%
Disgust	0	10	0	0	0	0	0	100%
Fear	0	0	12	0	0	0	0	100%
Happiness	0	0	0	32	0	0	0	100%
Neutral	0	0	0	0	50	0	0	100%
Sadness	0	0	0	0	0	20	0	100%
Surprise	0	0	0	0	0	0	32	100%
	100%	100%	100%	100%	100%	100%	100%	100%

Figure 10. Confusion matrix of DL_{21}.

when varying the training set. A low variance means that the data points are relatively close to the mean and hence to each other, while a high variance indicates that the data points are more scattered. The bias/variance relationship allows us to understand the performance of the model. A high bias and a low variance are a sign of model underfitting. Underfitting occurs when the model is unable to learn useful patters from the features. On the other hand, a low bias and a high variance indicates that the model is overfitting. This happens when the model pays too much attention to the training features that it captures the noise and the random fluctuations as useful features. In a perfect classification, both the bias and the variance have low values. In real world, however, there is usually a tradeoff between the bias and the variance. This tradeoff is related to the complexity of the model. A model that is too simple tends to have low variance but high bias. The more complex the model gets the lower the bias becomes and the higher the variance grows. So there is a tradeoff that allows for reaching a balance between bias and variance. According to the results in Table 3, ZeroR and KNN ($K = 2$ and $K = 3$) have a high bias and a low variance which signals a model underfitting. DT, NB, RF, SVM and CNN tend to have higher variance but lower bias. In overall, the variance is approximately between 0.1 and 0.8 which is a relatively low variance which suggests that with varying the training data, the model results do not vary greatly. Figure 7 shows a visual interpretation of the results in Table 3. The closer the points are to the center, the lower the bias is, and the more the points are clustered the lower is the variance. ZeroR and KNN ($K = 2$ and $K = 3$) are clustered but far away from the target. SVM and CNN are more scattered but closer to the target point.

As test the performance of the models on unseen data, the test set was used. Table 4 shows the accuracy, precision and recall of the test set using n-grams features. The table shows that KNN, (with $k = 3$) gave the lowest accuracy, that is 29.86%. NB classifier with all its versions has achieved an overall accuracy of 54%. RF with different values of n, gave an accuracy of 68.49% while SVM reached an accuracy of 75.6%. The highest performance was reached by the two-level CNN. As illustrated in section 3, the two-level CNN combines the CNN with ReLU (DL_{11}) and the CNN with shifted ReLU (DL_{12}) to a new CNN (DL_{21}). We notice from Table 4 that the two-level CNN

has an improved accuracy, precision, and recall compared to CNN with ReLU and CNN with shifted ReLU on their own. this model has achieved an accuracy of 89.3%.

With the aim to further experiment the performance of the two-level CNN, BOW features were used to train CNN with ReLU, CNN with shifted ReLU, and with the combined two-level CNN. The obtained results, shown in Table 5, manifest that the two-level CNN has increased the accuracy from 99.4% to 100% and the precision and recall from 98.83% and 98.16%, respectively, to 100%. The confusion matrices of DL_{11}, DL_{12}, and DL_{21} are demonstrated in Figures 8–10, respectively.

5. Conclusion

In this study, Arabic news headlines were extracted from the most popular news websites in the Middle East and annotated by three unbiased experts to seven classes: anger, disgust, fear, happiness, sadness, surprise, and neutral. The Fleiss' Kappa score was used to measure the agreement across different annotators. The obtained score was 0.75 which indicate a substantial agreement between annotators and hence promotes the reliability of the dataset. The collected news headlines underwent a preprocessing stage to clean the data using several unsupervised filters. The cleaning process involved removing signs (commas, semicolons, periods, etc.) and normalizing the data. Two feature extraction techniques were adopted, n-grams and BOW. For classification, a two-level deep learning approach was proposed and compared to six other ML classifiers (ZeroR, KNN, DT, NB, RF, and SVM). The proposed approach is based on assembling two CNNs and feeding their output to a third CNN. The two assembled CNNs consider using slightly different activation function where the first CNN uses ReLU and the second CNN used a shifted ReLU to -0.01. The variation of the activation function may help extracting different features in each CNN and thus leads to an improved performance when assembling the two CNNs. The experimental results showed that ZeroR and NB performed poorly in classifying the news headlines. KNN, DT, and RF showed an average accuracy of 68.5%, SVM resulted in an accuracy of 75.6% and CNN gave an accuracy of 88% and 87% using ReLU and shifted ReLU,

respectively. While the proposed assembled multilevel CNN resulted in an accuracy of 89.3% that is an improvement of more than 1% over the single CNN. Further experiments manifested that using BOW features, the two-level CNN gives higher results, that is an accuracy, precision and recall of 100% for the used dataset.

The study limitation lies in the random classifiers' guesses that were caused in the sample with insufficient training possibilities. In other words, emotional classes like fear '123 headlines', anger '129 headlines' and disgust '100 headlines' (as shown in Figure 3) affected the data training on these labels. Moreover, the labels with bigger samples have overcome others while getting assigned. Accordingly, some classifiers were biased towards the label with bigger training sample due to the obvious but manageable data imbalance.

In the future, the data size can be increased on these three labels to overcome the classifiers' bias problem. Moreover, some annotation of some headlines, especially for Anger category, were subject to ethnic bias due to annotators' culture, preferences, religion, beliefs and objectives (e.g. Muslims' angry annotations towards Al-Quds declaration as the capital of Israel related news headlines where non-Muslim might annotate them as Neutral). Such news headlines can be filtered in the future to reduce results bias. To reduce the labor of the annotation process, semi-automatic annotation that uses a GUI interface can be introduced. In addition, we plan to extend this study further to build a mobile-based solution that is able to recognize textual impact on people's emotions to help reduce news' harsh impact on readers' mental and psychological health.

Acknowledgments

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Notes on contributors

Hoda Elsayed received a bachelor's degree in software engineering with Honors at Alfaisal University in 2015 and a master's degree in Software Engineering with Honors at Prince Sultan University, KSA, in 2017. She is a former certified co-op trainee at Oracle and is currently an instructor of Software Engineering at Alfaisal university, KSA and a registered professional educator at IBM institute. Her research interests include big data analysis and machine learning in addition to development and management. She is an author of publications in academic journals and conferences and has been awarded by many entities for her research contribution.

Soumaya Chaffar graduated from the University of Montreal, where she received in 2006 and 2009, respectively, her Master and her PhD in Computer Science. Then, she has been employed as a postdoctoral researcher at the School of Electrical Engineering and Computer Science of the University of Ottawa, her main concern was to propose and experiment promising ideas and approaches related to analysis and generation of emotion in textual data. After that, as a postdoctoral researcher in Nuance Communications, she worked on industrial projects related to sentiment analysis for speech technologies. She joined Prince Sultan University as an assistant professor and received several excellence-based awards in Canada.

Samir Brahim Belhaouri received a master's degree in Telecommunications and Network from the InstitutNationale Polytechnique of Toulouse, France, in 2000, and the Ph.D. Degree in Mathematics from the Federal Polytechnic School of Lausanne-Switzerland, in 2006. He is currently an associate professor at the University of Hamad Bin Khalifa, Qatar foundation, in the Division of Information and Communication Technologies, College of Science and Engineering, HBKU. During past years, he also holds several Research and Teaching positions at Innopolis University-Russia, Alfaisal university-KSA, University of Sharjah-UAE, University Technology PETRONAS-Malaysia, and EPFL Federal Swiss School-Switzerland. His research interests include vary areas from Applied Mathematics, Statistics, and Data analysis until Artificial Intelligence, Image & Signal Processing (Biomedical, Bioinformatics, Forecasting…), due to both Mathematics and Computer Science backgrounds.

Hafsa Raissouli received a bachelor's degree in Computer Science, Taibah University, KSA in 2019. Master student in Faculty of Computer Science and Information Technology, University Putra Malaysia. Her research interests include data science, machine learning, and deep learning.

References

[1] Griffiths F, Dobermann T, Cave J, et al. The impact of online social networks on health and health systems: a scoping review and case studies. Policy Internet. 2015;7(4):473–496.

[2] Bessière K, Pressman S, Kiesler S, et al. Effects of internet use on health and depression: a longitudinal study. J Med Internet Res. 2010;12(1):e6.

[3] Johnston WM, Davey GC. The psychological impact of negative TV news bulletins: The catastrophizing of personal worries. Br J Psychol. 1997;88(1):85–91.

[4] McHugo GJ, Smith CA, Lanzetta JT. The structure of self-reports of emotional responses to film segments. Motiv Emot. 1982;6(4):365–385.

[5] Sternbach RA. Assessing differential autonomic patterns in emotions. J Psychosom Res. 1962;6(2):87–91.

[6] Mathews A. Why worry? The cognitive function of anxiety. Behav Res Ther. 1990;28(6):455–468.

[7] Mathews A, MacLeod C. Cognitive approaches to emotion and emotional disorders. Annu Rev Psychol. 1994;45:25–50.

[8] Mogg K, Mathews A, Eysenck M. Attentional bias to threat in clinical anxiety states. Cogn Emot. 1992;6(2):149–159.

[9] Wells A, Morrison AP. Qualitative dimensions of normal worry and normal obsessions: a comparative study. Behav Res Ther. 1994;32(8):867–870.

[10] Ibrahim HS, Abdou SM, Gheith M. MIKA: A tagged corpus for modern standard Arabic and colloquial sentiment analysis. In 2015 IEEE 2nd International Conference on Recent Trends in Information Systems, ReTIS 2015 – Proceedings; 2015. p. 353–358. Available from: https://doi.org/10.1109/ReTIS.2015.7232904

[11] Philippot P. Inducing and assessing differentiated emotion-feeling states in the laboratory. Cogn Emot. 1993;7(2):171–193.

[12] Gibbs W, McKendrick J. Contemporary research methods and data analytics in the news industry. Hershey (PA): IGI Global; 2015.

[13] Hamouda AEA, El-taher FE. Sentiment analyzer for Arabic comments system. Int J Adv Computer Sci Appl. 2013;4(3):99–103.

[14] Al-Ayyoub M, Bani Essa S, Alsmadi I. Lexicon-based sentiment analysis of Arabic tweets. Int J Soc Netw Min. 2015;2(2):101–114.

[15] Khoja S, Garside R. Stemming Arabic Text. Lancaster, UK: Computing Department, Lancaster University; 1999.

[16] Abuaiadh D. Dataset for Arabic Document Classification [online]. 2011. Available from: http://diab.edublogs.org/dataset-for-arabic-document-classification/

[17] Diab M. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. Proceedings of the Second International Conference on Arabic Language Resources and Tools; Cairo, Egypt: 2009. p. 285–288.

[18] Abdul-Mageed M, Kuebler S, Diab M. Samar: A system for subjectivity and sentiment analysis of arabic social media. Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis. Association for Computational Linguistics; 2012. p. 19–28.

[19] Al-Kabi MN, Khasawneh RT, Wahsheh HA. Evaluating social context in Arabic opinion mining. Int Arab J Inform Technol. 2018;15(6):974–982.

[20] Al-Azani S, El-Alfy E-SM. Imbalanced sentiment polarity detection using emoji-based features and bagging ensemble. Proceedings of the 2018 1st International Conference on Computer Applications and Information Security (ICCAIS); Riyadh, Saudi Arabia: IEEE; 2018. p. 1–5.

[21] Elouardighi A, Maghfour M, Hammia H, et al. A machine learning approach for sentiment analysis in the standard or dialectal Arabic facebook comments. Proceedings of the 2017 3rd International Conference of Cloud Computing Technologies and Applications (CloudTech); Rabat, Morocco: IEEE; 2017. p. 1–8.

[22] Al-Tamimi A-K, Shatnawi A, Bani-Issa E. Arabic sentiment analysis of youtube comments. Proceedings of the 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT); Amman, Jordan: IEEE; 2017. p. 1–6.

[23] El Ariss O, Alnemer LM. Morphology based Arabic sentiment analysis of book reviews. Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing; Budapest, Hungary: Springer; 2017. p. 115–128.

[24] Maghfour M, Elouardighi A. Standard and dialectal Arabic text classification for sentiment analysis. Proceedings of the International Conference on Model and Data Engineering; Marrakesh, Morocco: Springer; 2018. p. 282–291.

[25] Nuseir A, Al-Ayyoub M, Al-Kabi M. Improved hierarchical classifiers for multi-way sentiment analysis. International Arab Conference on Information Technology (ACIT'2016) 2016.

[26] Altawaier MM, Tiun S. Comparison of machine learning approaches on Arabic twitter sentiment analysis. Int J Adv Sci Eng Inform Technol. 2016;6(6):1067–1073.

[27] Socher R, Pennington J, Huang EH, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions. Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics; 2011. p. 151–161.

[28] Socher R, Lin CCY, Ng AY, et al. Parsing Natural Scenes and Natural Language with Recursive Neural Networks. ICML; 2011.

[29] Socher R, Perelygin A, Wu J, et al. Recursive deep models for semantic compositionality over a sentiment treebank. Proceedings of the 2013 conference on empirical methods in natural language processing; 2013. p. 1631–1642.

[30] Cliche M. BB twtr at SemEval-2017 Task 4: Twitter Sentiment Analysis with CNNs and LSTMs. 2017. Preprint arXiv:1704.06125.

[31] Maas L, Daly RE, Pham PT, et al. Learning accurate, compact, and interpretable tree annotation. Proceedings of ACL; 2011.

[32] Baly R, Badaro G, El-Khoury G, et al. A characterization study of Arabic twitter data with a benchmarking for state-of-the-art opinion mining models. Proceedings of the Third Arabic Natural Language Processing Workshop; 2017. p. 110–118.

[33] Al Sallab A, Hajj H, Badaro G, et al. Deep learning models for sentiment analysis in Arabic. Proceedings of the Second Workshop on Arabic Natural Language Processing; 2015. p. 9–17.

[34] Ekman P. An argument for basic emotions. Cogn Emot. 1992;6:169–200.

[35] Chaffar S, ElSayed H, Belhouari S, et al. Arabic news headlines. IEEE Dataport. 2020; Available from: http://dx.doi.org/10.21227/7e79-nt12

[36] Alarabiya RSS service. (n.d.). [updated 2017 Jun 12]. Available from: https://www.alarabiya.net/tools/mrss

[37] Al-Riyadh RSS service. (n.d.). [updated 2017 May 7]. Available from: http://www.alriyadh.com/page/rss

[38] Boudelaa S. Is the Arabic mental lexicon morpheme-based or stem-based? Implications for spoken and written word recognition. Handbook of Arabic literacy. Springer; 2014. p. 31–54.

[39] Moftah M, Fakhr W, Abdou S, et al. Stem-based Arabic language models experiments. Proceedings of the Second International Conference on Arabic Language Resources and Tools, Cairo, Egypt, April. The MEDAR Consortium; 2009.

[40] Aghtar S. A New Incremental Classification Approach Monitoring: The Risk of Heart Disease. 2012. Available from: https://macsphere.mcmaster.ca/bitstream/11375/12637/1/fulltext.pdf.

[41] Ramaswamy S, Golub TR. DNA microarrays in clinical oncology. Clin Oncol. 2002;20:1932–1941.

[42] Zhao Z, Liu H. Spectral feature selection for supervised and unsupervised learning. Proceedings of the 24th international conference on Machine learning; ACM; 2007. p. 1151–1157.

[43] Lakshmi Devasena C, Proficiency comparison of ZeroR, RIDOR and PART classifiers for Intelligent Heart Disease Prediction. Oper Res Appl: Int J Adv in Computer Sci Technol (IJACST). 2014;3(11):12–18. Special Issue.

[44] Elsayed H, Syed L. An automatic early risk classification of hard coronary heart diseases using Framingham scoring model. In second International Conference on Internet of Things, Data and Cloud Computing (ICC 2017); ACM; 2017.

[45] Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. 1950;29(2):147–160. DOI: 10.1002/j.1538-7305. 1950.tb00463.x, MR 0035935.html?cref = navdesk

[46] Xing Y, Wang J, Zhao Z. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In 2007 International Conference on Convergence Information Technology; IEEE; 2007. p. 868–872.

[47] Calders T, Verwer S. Three naive Bayes approaches for discrimination-free classification. Data Min Knowl Discov. 2010;21(2):277–292. Available from: https://doi.org/10.1007/s10618-010-0190-x

[48] Ani R, Augustine A, Akhil NC, et al. Random forest ensemble classifier to predict the coronary heart disease using risk factors. In Proceedings of the International Conference on Soft Computing Systems; India: Springer; 2016. p. 701–710.

[49] Evgeniou T, Pontil M. Support vector machines: theory and applications. Machine learning and its applications. 2001. p. 249–257.

[50] KASI HG. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. DOI:10.1145/3065386