


Affective Knowledge-enhanced Emotion Detection in Arabic Language: A Comparative Study


Jesus Serrano-Guerrero

(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain)

 <https://orcid.org/0000-0002-6177-8188>, jesus.serrano@uclm.es


Bashar Alshouha

(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain)

 <https://orcid.org/0000-0001-6475-4248>, bashar.alshouha@alu.uclm.es


Francisco P. Romero

(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain)

 <https://orcid.org/0000-0002-6993-2434>, franciscop.romero@uclm.es

Jose A. Olivas

(University of Castilla-La Mancha, Information Technologies and Systems Dept., Ciudad Real, Spain)

 <https://orcid.org/0000-0003-4172-4729>, joseangel.olivas@uclm.es

Abstract Online opinions/reviews contain a lot of sentiments and emotions that can be very useful, especially, for Internet suppliers which can know whether their services/products are meeting their customers' expectations or not. To detect these sentiments and emotions, most applications resort to lexicon-based approaches. The major issue here is that most well-known emotion lexicons have been developed for English language; nevertheless, in other languages such as Arabic, there are fewer available tools, and many times, the quality of them is poor.

The goal of this study is to compare the performance of two different types of algorithms, shallow machine learning-based and deep learning-based, when dealing with emotion detection in Arabic language. To improve the performance of the algorithms, two lexicons, which were originally developed in other languages and translated into Arabic language, have been used to add emotional features to different information models used to represent opinions. All approaches have been tested using the dataset SemEval 2018 Task 1: Affect in Tweets and the dataset LAMA+DINA. The semantic approaches outperform the classical algorithms, that is, the information provided by the lexicons clearly improves the results of the algorithms. Particularly, the BiLSTM algorithm outperforms the rest of the algorithms using word2vec. On the contrary to other languages, the best results were obtained using the NRC lexicon.

Keywords: Emotion detection, affective feature detection, machine learning, deep learning, affective lexicons

Categories: H.3.0, H.3.3, J.7, J.4

DOI: 10.3897/jucs.72590

1 Introduction

Affective computing is an emerging research field whose goal is to detect, model and interpret human emotions. It is a multidisciplinary area related to fields such as psychology, sociology, artificial intelligence (AI), natural language processing (NLP), etc. Fields such as sentiment analysis and emotion detection, which are recently gaining a lot of interest in social media, are key parts of affective computing, in the sense they allow computers to deal with sentiments and emotions coming from different users [Poria et al. 2017].

To detect the users' emotional states, multiple information sources can be used which span from classical multimedia resources (videos, images, texts) to body gestures, face or voice recognition, etc. Due to the increasing development of Internet applications, opinions/reviews are becoming an interesting source to find user emotions in multiple languages. To detect these sentiments or emotions, artificial emotional intelligence can resort to classification algorithms based on classical strategies such as shallow machine learning (ML) techniques, or more recent and trendier, such as deep learning (DL) techniques. Supervised classification is one of the most used emotion detection techniques due to the availability of a large number of labeled datasets. It primarily consists in, given a number of training examples associated with the corresponding outcomes, finding the relationship between the patterns and the outcomes using solely the training samples. These training data must be represented by their attributes which are mainly captured by feature extraction techniques, which model and select the best features. The selection of the best features is vital for the classification process. Furthermore, the use of lexicons to provide these techniques with semantic capabilities is a key resource to improve their performance.

When dealing with lexicons, the major issue is to find high-quality resources. Regarding emotion analysis, there are several well-known tools such as SenticNet [Cambria et al. 2020], WordNet-Affect (WNA) [Strapparava and Valitutti, 2004], SentiStrength [Baccianella et al. 2010], ANEW [Bradley and Lang 1999], AFFIN [Nielsen 2011], NRC Word-Emotion Association Lexicon [Mohammad and Turney 2013], among others. Nonetheless, these tools are mainly based on English language. In other languages, the quality of this type of tools is not so high, for that reason, in many cases, the researchers resort to translate these tools, assuming some quality loss. One of these cases is the Arabic language.

There are several attempts to develop Arabic lexicons in the literature [Al-Moslmi et al. 2018, Guellil et al. 2018]; nevertheless, most of them are especially focused on sentiments (positiveness or negativeness), and not on emotions such as anger, rage, happiness, etc. Moreover, most of them do not include words from dialects and are not even written in Arabic standard but Arabizi, that is, Arabic text written using Latin characters. For that reason, it is more difficult to develop algorithms for emotion categorization in Arabic language. Hence, the aim of this study is to compare the performance of supervised classifiers including emotional features extracted from several lexicons when detecting emotions in Arabic texts. To sum up, the main contributions of this study are:

- To compare the performance of supervised classification methods applied to emotion recognition.
- To compare the use of different feature extraction techniques applied to the previous methods.

- To study the effect of using lexicons to provide emotional lexical features to improve the information used by these previous techniques for classifying emotions.
- To test the proposals using well-known datasets and compare the obtained results in Arabic with other languages such as Spanish.

The rest of the paper is organized as follows: Next section summarizes the state of the art, and Section 3 describes the necessary methods and materials to be used in this study. Section 4 explains the methodology to be followed whereas Section 5 discusses the results. Finally, some conclusions and future work are pointed out.

2 Literature review

Microblogging platforms and social networks are some of the most popular online applications that allow people to share their ideas over several topics and communicate with each other. On these platforms, users can publish posts including their activities, photos, videos, thoughts, and so on. This information could include indicators about a person's emotional state related to anxiety, stress, depression, satisfaction, among others. Therefore, most textual posts contain emotional information that can be collected and analyzed to develop practical applications exploiting this information such as recommendation systems or analytical dashboards. Most of the approaches on emotion detection from texts, also for sentiment detection, follow strategies based on shallow ML, DL, lexicons, or hybrid approaches [Birjali et al. 2021].

2.1 Shallow machine learning approaches

In [Al-Khatib and El-Beltagy 2018], Al-Khatib et al. proposed a naïve Bayes approach to detect emotions in Arabic texts. As input features, they used a bag of words (BOW) representation model consisting of different n-grams. The proposed approach was evaluated using their own dataset derived from Twitter and based on Ekman's basic emotions model [Ekman and Friesen 1969]. The proposed approach achieved 68.1% of accuracy. [Duppada and Hiray 2017] implemented a parallel tree boosting by XGBoost¹ to deal with the dataset WASSA 2017 Task 1, using multiple features such as syntactic and lexicon features, and word vectors. [Mohammad and Kiritchenko 2015] used the datasets "Hashtag Emotion Corpus" and "Headlines" for detecting emotions from tweets using emotion-word hashtags. They used an emotion manually-labeled tweet corpus to create a vast lexicon of word-emotion associations. The experimental results of the support vector machine (SVM) algorithm on six basic emotions indicated an improvement in emotion association accuracy.

Aside from English, shallow ML classifiers have been also applied to other languages for emotion detection, for instance, Roman Urdu. [Majeed et al. 2020] developed a large corpus of sentences from different fields, classified into six classes and applied word embeddings for Roman Urdu. For emotion detection, several shallow ML algorithms were applied, and the results indicated that SVM achieved the best accuracy (69.4%) and F1-score (69%). [Jayakrishnan et al. 2018] applied SVM algorithms to classify emotions in Indian languages. Many syntactic features were used to improve the classification such as n-grams, negation related, POS related, and level related features. Furthermore,

¹ <https://xgboost.readthedocs.io/en/latest/>

the proposed approach specified whether the sentence was a conversation, a question, or not. [Suhasini and Srinivasu 2020] compared and applied two types of shallow ML algorithms, Naïve Bayes and K-nearest neighbor algorithm (KNN), to classify emotions of Twitter messages into four emotional categories. The results showed that the Naïve Bayes outperformed KNN.

2.2 Deep learning approaches

Many DL approaches to text emotion detection have been recently proposed, for instance, [Wu et al. 2018] proposed an attention-based CNN (convolutional neural network)-LSTM (long short-term memory) model to predict the intensity score of the emotions and sentiments. A CNN layer with different kernel sizes was used to extract the features and long-term contextual information was extracted from texts using LSTM. Another similar approach is SEDAT [Abdullah et al. 2018], a system combining feed-forward, LSTM, and CNN to predict the sentiments/emotions of a tweet. This system deals with the problem of SemEval 2018 Task 1, combining different features such as word embedding vectors [Mikolov et al. 2013] and semantic features acquired from the AffectiveTweets package [Bravo-Marquez et al. 2014]. In [Baali and Ghneim 2019], a CNN architecture was proposed to detect emotions, in which the information of several tweets was modeled at different levels: word vectorization by a word2vec model, sentence vectorization and document vectorization. The proposed classifier outperformed other shallow ML algorithms.

[Zahiri and Choi 2017] introduced a new corpus for emotion detection tasks obtained from spoken dialogues. They proposed four novel types of sequence-based convolutional neural network models for contextual emotion detection. [AlZoubi et al. 2020] applied several models such as CNNs, bidirectional GRU_CNN, and XGBoost regressor combining many feature extraction methods were used such as Term Frequency - Inverse Document Frequency (TF-IDF), word-level embedding or lexicon features to calculate the emotion intensity for a given tweet.

[Jabreel and Moreno 2018] applied shallow ML techniques such as SVM unigrams and XGboot regressor based on a set of embeddings and lexicons-based features, and an N-Channels ConvNet to cope with the emotion intensity and valence regression (El-reg, V-reg) task of the SemEval-2018 Task 1: Affect in Tweets (AIT) in Arabic and English language. They built an ensemble model combining two different techniques: N-Stream ConvNets and XGBoost regressor. The study concluded that N-Channels ConvNet's performance was close to the results of the ensemble models. To test the same dataset, [Abdullaand Shaik2018] proposed a fully connected neural network structure whose layers were fed by the concatenation of doc2vec embeddings, word2vec and a set of psycholinguistic features. [Ma et al. 2019] proposed a BiLSTM to determine the contextual emotion in text. An emotion-oriented attention network was added to the proposed approach, which could extract emotional information from an utterance.

[Polignano et al. 2019] combined BiLSTM and CNN with a self-attention layer providing the model with the capability to weigh the vectors of single words in a sentence, depending on the similarity between the neighboring tokens. Moreover, they applied different pre-trained word-embedding techniques, concluding that FastText word-embedding technique achieved the best results. [Rayhan et al. 2020] developed two DL models to identify emotions in Bangla texts. These models were a Bidirectional Gated Recurrent Unit (BiGRU) and a CNN-BiLSTM network. The results showed that the CNN-BiLSTM model outperformed the other model.

2.3 Hybrid and lexicon-based approaches

When regard to hybrid or pure lexicon-based approaches, several well-known lexicons can be found in English language, but most of them are focused on sentiments, rather than emotions, for instance, WordNet-Affect (WNA) [Strapparava and Valitutti, 2004], SentiStrength [Baccianella et al. 2010], ANEW [Bradley and Lang 1999], AFFIN [Nielsen 2011], and NRC Word-Emotion Association Lexicon [Mohammad and Turney 2013]. Nonetheless, for other languages such as Arabic or Spanish, the researchers, many times, must resort to translating these English resources.

Duwairi et al. used a SentiStrength-based approach to detect emotions in tweets [Duwairi et al. 2015], whereas Bandhakavi et al. used EmoSenticNet, WordNet-Affect, and the NRC Emotion Lexicon in an emotion detection task, achieving NRC the best performance [Bandhakavi et al. 2017]. Rabie et al. proposed a lexicon-based approach using an extracted sample word-emotion lexicon obtained from a manually annotated corpus in Arabic [Rabie and Sturm 2014].

[Plaza-del-Arco et al. 2020] proposed a mechanism to integrate knowledge from three different affective lexical resources into shallow ML classifiers. The results demonstrated that the incorporation of lexical features led to substantial improvements over most of the shallow ML classifiers.

[Alswaidan and Menai 2020] proposed three DL models for classifying emotions in Arabic texts. The first model was a human-engineered feature-based model which contained three dense neural network layers concatenated with many human-engineered feature methods including linguistic features, lexical sentiment features from an Arabic Twitter sentiment lexicon, an Arabic hashtag lexicon, an Arabic emoticon lexicon, and an Arabic hashtag dialectal lexicon, and lexical emotion features from the NRC lexicon, syntactic features, and semantic features from SenticNet [Cambria et al. 2020]. The second model contained an LSTM layer with a GRU layer. The third model was a hybrid model that concatenated the previous models. The result demonstrated that the hybrid model outperformed the other two.

3 Methods and materials

3.1 Lexicons

As it is not easy to find specific emotion lexicons in Arabic language, to provide affective information, the following two non-Arabic emotion lexicons have been used. Both have been developed in different languages to compare whether this fact can also affect the classification process or not.

- **Translated Improved Spanish Emotion Lexicon (TISEL)** is the result of translating into Arabic, using Google translator, the lexicon SEL [Plaza-del-Arco et al. 2020], which was specifically developed for Spanish emotions. It contains 2,036 Spanish words and each word is related to a measure of the Probability Factor of Affective Use (PFA) for at least one of Ekman's basic emotions: joy, sadness, fear, anger, disgust, and surprise.
- **NRC** contains a set of English words that are related to one or more of the following emotions: joy, anger, sadness, fear, trust, anticipation, surprise, and anger. It is possible that a single word is related to multiple emotions. This lexicon was developed specifically for English language and translated into over one hundred other languages such as Arabic by Google translator [Mohammad and Turney 2013].

3.2 Methods

3.2.1 Classification of machine learning algorithms

- **SVM:** is a statistical classifier which has proven a good performance both for classification and regression. It performs classification by initializing hyperplanes in a multi-dimensional space. The value of each feature is also the value of a specified coordinate in the hyperplanes. It is a boundary method for separating many classes by optimizing the ideal hyperplanes by means of the use of functions called “kernels” [Tong and Koller 2001].
- **Multinomial Naïve Bayes (MultiNB):** is another learning algorithm used for text classification problems. Multinomial Naïve Bayes is often referred as “multivariate event model”. Particularly, in document classification, the events represent the occurrence of a term in a document. The process assumes that the input values represent the frequencies with which certain terms have been generated by a multinomial distribution (p_1, p_2, \dots) , being p_i the probability that term i occurs. The Naïve Bayes theorem calculates the posterior probability for each class and predicts the class with the highest probability, being able to carry out both binary classification and multi-class classification problems [Kibriya et al. 2004].
- **Multilayer perceptron (MLP):** is a type of feedforward artificial neural network (ANN). MLP is made up of three layers of nodes: input layer, hidden layer and an output layer. Under MLP, a supervised learning technique known as backpropagation is used for training [Singh and Shahid Husain 2014].
- **Logistic Regression (LR):** is a binary classification algorithm that assumes the input of variables are numeric and has a Gaussian distribution. It is expected that an algorithm learns a coefficient for each input value, which are linearly combined into the regression function and transformed using a logistic function. It is a popular classification algorithm which belongs to the class of generalized linear models [Dreiseitl and Ohno-Machado 2002].

3.2.2 Classification of deep learning algorithms

- **CNN:** is network whose typical architecture consists of three classes of layers: convolutional layer, pooling layer and fully connected layer. The goal of the convolutional layer is to detect similarities of the features from the previous layer and learn the feature representation of the input through the use of filters, which are also known as kernels. A filter is a matrix of weights specifically trained to detect particular features. The second layer is the pooling layer, which achieves shift variance by reducing the resolution of the feature map. And the third one is the fully connected layer, which performs high-level reasoning and connects all neurons in the previous layer with all neurons in the current layer [Kim 2014].
- **Single LSTM:** is a type of recurrent neural network (RNN) architecture whose goal is to avoid the “long-term dependencies” problem. The typical architecture has an input layer, hidden layers, and output layer, and contains a set of memory blocks connected with each other via gates. Each hidden layer consists of an LSTM cell that applies the necessary steps to generate the new state and move it to the next hidden layer and so on, to finally reach the output layer [Hochreiter and Schmidhuber 1997].

- **Bidirectional LSTM:** is an extension of the traditional LSTMs that has the capability to improve the performance of sequence classification problems. Bidirectional LSTMs train two LSTMs instead of one, both of which are connected to the same output layer [Zhou et al. 2016]. This provides additional context to the network, which is one of the limitations of LSTM, considering contextual information from the future.

3.2.3 Feature extraction

The accuracy of a learning system relies on the followed representation model. Regarding the text classification task, it is important to convert the document into a format that the learning classifier can understand. The aim of the feature selection technique is to remove irrelevant, redundant, and noisy data to find the most relevant features. The second goal of feature selection is to reduce both the dimensionality of the feature space and the processing time [Oussous et al. 2020]. In our experiments, to train the shallow ML classifiers and DL algorithms, each tweet has been represented by a vector of numerical features weighted by TF-IDF [Robertson 2004] for the shallow ML classifiers, and word embeddings [Lebret et al. 2016] for the DL algorithms.

- **Bag-of-words (BOW) using TF-IDF:** BOW is just a technique for representing the terms whose importance has been weighted by the statistical measure TF-IDF. This measure is based on how many times the words appear in a document and the inverse document frequency of these words across the documents [Robertson 2004]. Different formulas can be found to compute TF and IDF. Specifically, in our experiments, the weight of a term in a document has been computed as $TF_{ij} * IDF_i$, being TF_{ij} the frequency of term i in document j and $IDF_i = \log(N/n_i) + 1$, where N is total number of documents in the collection and n_i is the number of documents where term i appears.
- **Word Embedding:** Unlike BOW, this technique tries to capture the semantics of the words, preserving the context and relationship between them, and not just counting the number of words. It is a representation model in the vector space model which utilizes distributed representations of a word to capture both semantic and syntactic features of that word [Xu et al. 2018].

It is possible to use well-known pre-trained word embeddings or develop your own one for a specific dataset. In this case, Keras embedding layer² has been used to provide a dense representation of our dataset, and also word2vec. Word2vec is a technique developed by [Mikolov et al. 2013] at Google, which is considered one of the most common techniques to learn word embeddings using a shallow neural network. It can follow two approaches: continuous Bag-of-Words (CBOW) which aims to learn the embeddings by predicting the key word in a context given the other words in the context without considering their order in the sentence, and Skip-Gram model which aims to predict the surrounding context words given a word. In our experiments, a pretrained word2vec model based on CBOW with 300 dimensions for Arabic was used.

² https://keras.io/api/layers/core_layers/embedding/

3.3 Dataset

To conduct our experiments, one of the used dataset is the one proposed for the the EI-oc (emotion intensity ordinal classification) task of SemEval 2018 Task 1 AIT [Mohammad et al. 2018], composed of a set of 5, 612 Arabic tweets divided into four emotions: anger, fear, joy, and sadness. The EI-oc task consists in detecting the intensity of every tweet, which can be classified from no-intensity to high-intensity. Nevertheless, this is not the purpose of this study, but emotion classification. For that reason, those tweets labeled as “0-no emotion can be inferred”, that is, they are not conveying any emotion, have been removed, keeping only those tweets conveying emotions from each subset (train, dev and test). The distribution of the tweets of the resulting dataset is shown in Table 1.

Class Label	Train	Dev	Test	Total
Anger	629	109	299	1,037
Fear	530	92	274	896
Joy	624	193	373	1,190
Sadness	501	96	242	839

Table 1: Number of tweets per emotion in the SemEval-2018 AIT dataset.

Aside from this dataset, LAMA+DINA has been also used in our experiments. This dataset was proposed to assess the performance of the DL toolkit for arabic social media called AraNet [Abdul et al. 2020]. It represents the combination of another two previous Twitter datasets, DINA [Abdul et al. 2016] and LAMA [Alhuzali et al. 2018]. All tweets are labeled using 8 emotions following the distribution shown in Table 2. 80% of the tweets were used for training the classification algorithms and 20% for testing.

Class Label	# tweets
Anger	916
Anticipation	922
Disgust	998
Fear	1,392
Happy	1,281
Sad	990
Surprise	1,142
Trust	853
Total	8494

Table 2: Number of tweets per emotion in the LAMA+DINA dataset.

3.4 Evaluation measures

To evaluate the performance of the proposed models, four classical metrics were employed in order to compare and evaluate the results: recall, precision, F1-score, and

accuracy. Let TP , TN , FP and FN be the number of correctly classified tweets, correctly rejected tweets, misclassified tweets and incorrectly rejected tweets, these metrics are mathematically defined as follows:

- **Accuracy:** is the ratio of correctly classified instances over the total number of instances, as shown in Eq. 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

- **Precision:** is the ratio of the correctly identified tweets (TP) over the number of expected tweets, as shown in Eq. 2:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- **Recall:** is the ratio of the correctly identified tweets (TP) over the total of detected tweets, as shown in Eq. 3:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

- **F1-score:** is the harmonic mean of precision and recall defined by Eq. 4:

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \quad (4)$$

4 Methodology

This is an outline of the various steps and processes that have been conducted to accomplish this research. The followed methodology includes: preprocessing the dataset, extracting the main features for both shallow ML and DL algorithms, modeling the information to perform the classification algorithms, and finally evaluating the obtained output. Fig 1 is a diagrammatic representation of these steps.

The proposed emotion recognition methods as well as the evaluation system have been implemented in Python using the following libraries: Natural Language ToolKit (NLTK) [Bird et al. 2009], scikit-learn [Pedregosa et al. 2011], and Keras using TensorFlow backend [Brownlee 2019].

4.1 Data preprocessing

Text preprocessing is relevant to the data analysis process as it deals with the noise and colloquial nature of Twitter data, and especially, considering the inherited ambiguity of the Arabic language and the use of dialectal Arabic. It is also helpful as it can reduce the number of iterations, making the models converge faster.

The following preprocessing steps have been applied in this order to transform the data that will feed the categorization models:

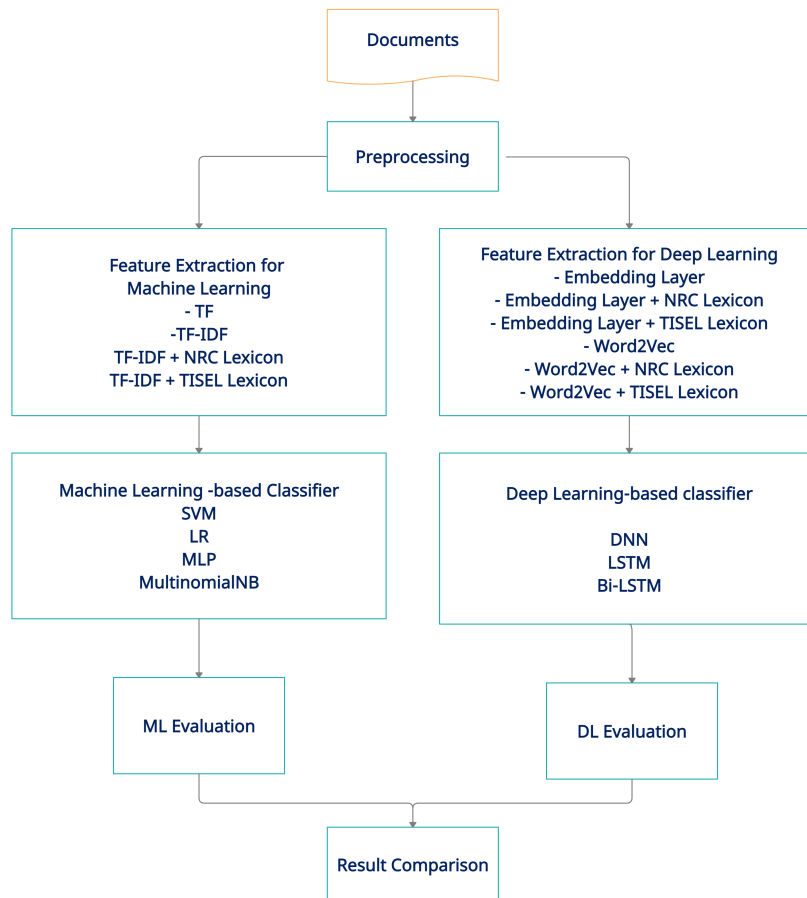


Figure 1: Steps followed to assess the emotion classifiers

- **Data cleaning:** It involves the removal of punctuation, additional whitespaces, non-Arabic characters, numbers, underscores, and diacritics such as Fatha, Tanwin Fath, Damma, Tanwin Damm, Kasra, and Tatwil. Two tools were used to preprocess the tweets: NLTK and regular expressions using python [Stubblebine 2007].
- **Tokenization:** The tweets are divided into multiple tokens based on separator characters such as a white space, comma, tab, etc. NLTK TweetTokenizer was used to tokenize the tweets.
- **Stop words:** It is performed the removal of words that contain little information such as conjunctions and prepositions. NLTK was used to remove stop words.
- **Lemmatization:** It consists in the vocabulary and morphological analysis of words with the aim of removing inflectional endings. NLTK WordNetLemmatizer was used to apply lemmatization for Arabic language.

- **Stemming:** It is the reduction of words to their stems and the removal of suffixes of the terms based on some grammatical rules. NLTK ISRIStemmer was used to implement it.

4.2 Approach integrating affective knowledge

Feature selection is a key step for emotion recognition. We claim that using emotional external knowledge can improve emotion classification and investigate the effectiveness of various affective lexical features applied to several shallow ML and DL algorithms.

These features are provided by the lexicons as described in Section 3. Each sentence is preprocessed using Stanford Core NLP library and then, the TF-IDF scheme is computed for the shallow ML algorithms, and word2vec or Keras embedding layer for the DL algorithms, with the aim of transforming each text into a numerical vector representation. To incorporate the affective lexical features, the presence of lexicon words in the sentence is checked and a vector is obtained representing each emotional category. Finally, to perform the classification, the TF-IDF vector representation and the affective features are concatenated and used as an input for the different shallow ML algorithms, whereas the concatenation of the embedding layer or word2vec, and the affective features is the input for the different DL algorithms. The followed mechanism for computing the affective lexical features is:

1. **TISEL.** After detecting the presence of the lexicon words in a sentence, then the addition of the intensity values of the words grouped by the emotional category (joy, fear, anger, sadness) is computed. As a result, a vector of four emotional values for each sentence is generated and concatenated [Plaza-del-Arco et al. 2020].
2. **NRC.** The Arabic version of this resource has been used. In a similar manner, the presence of lexicon words in a sentence has been computed as well as the sum of the emotion intensity scores grouped by the emotional categories (anger, joy, fear, sadness, surprise, disgust, trust, anticipation). This manner, a vector of eight values (eight emotions) for each sentence has been generated and concatenated.

4.3 Emotion classification

The last step consists in using the features extracted in the previous phase to classify the emotions from the dataset by means of the methods explained in subsection 3.2.

5 Results

After implementing all approaches³ following the methodology explained in the previous section, the results are described and compared in this section.

³ <https://github.com/201190024/Emotion-detection-In-Arabic-language>

5.1 Baseline machine learning classifiers

Firstly, the four shallow ML-based classifiers (SVM, LR, MultiNB, and MLP) were executed without using any lexicon, obtaining the results shown in Tables 3 and 4, respectively. In this case, regarding both datasets, LR outperforms the rest of algorithms in terms of F1-score, precision, recall, and accuracy. On the contrary, MLP obtains the worst performance among all of the shallow ML algorithms. And among the targeted classes, “joy” in SemEval and “anticipation” in LAMA+DINA obtained the best results for all of the shallow ML algorithms in comparison with the other classes. Nevertheless, those tweets related to “sadness” were the most difficult to classify by all approaches.

	Anger			Fear			Joy			Sadness			Average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Ac
SVM	0.51	0.54	0.52	0.53	0.53	0.53	0.83	0.88	0.86	0.51	0.44	0.47	0.60	0.60	0.59	0.62
LR	0.51	0.65	0.57	0.61	0.55	0.58	0.86	0.88	0.87	0.57	0.43	0.49	0.64	0.63	0.63	0.64
MultiNB	0.47	0.49	0.48	0.47	0.51	0.49	0.83	0.87	0.85	0.51	0.42	0.46	0.57	0.57	0.57	0.59
MLP	0.50	0.48	0.49	0.46	0.49	0.47	0.85	0.82	0.83	0.39	0.39	0.40	0.55	0.55	0.55	0.57

Table 3: Results for shallow ML classifiers using the SemEval dataset

	MultiNB			SVM			LR			MLP		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.59	0.42	0.49	0.52	0.49	0.50	0.54	0.52	0.53	0.47	0.42	0.44
Sad	0.43	0.32	0.37	0.42	0.43	0.42	0.49	0.37	0.42	0.37	0.38	0.37
Fear	0.49	0.52	0.51	0.56	0.56	0.56	0.48	0.68	0.56	0.51	0.50	0.51
Disgust	0.42	0.51	0.46	0.42	0.52	0.46	0.54	0.49	0.51	0.37	0.47	0.41
Surprise	0.52	0.53	0.52	0.64	0.58	0.61	0.76	0.57	0.65	0.56	0.54	0.55
Happy	0.44	0.64	0.52	0.50	0.58	0.54	0.43	0.61	0.51	0.47	0.53	0.50
Trust	0.55	0.35	0.43	0.59	0.47	0.52	0.61	0.40	0.48	0.53	0.46	0.49
Anticipation	0.62	0.56	0.59	0.72	0.61	0.66	0.78	0.61	0.69	0.64	0.54	0.58
Average	0.51	0.48	0.49	0.55	0.53	0.54	0.58	0.53	0.54	0.49	0.48	0.48
Acc	0.49			0.54			0.54			0.48		

Table 4: Results for the shallow ML classifiers using the LAMA+DINA dataset

5.2 Machine learning algorithms with lexical affective features

Secondly, the shallow ML classifiers have been tested enriching their data representations adding lexical affective features from TISEL and NRC. Comparing the results, the use of NRC’s features obtained the best results compared to TISEL for both datasets. Moreover, regarding the use of TISEL, the results improved by around 1% for the best shallow ML technique, LR, in terms of F1-score and accuracy, as shown in Table 5 and 6 for both datasets.

	Anger			Fear			Joy			Sadness			Average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
SVM	0.56	0.59	0.56	0.56	0.50	0.53	0.83	0.88	0.85	0.52	0.50	0.51	0.62	0.62	0.62	0.64
LR	0.54	0.67	0.60	0.63	0.52	0.57	0.85	0.87	0.86	0.56	0.48	0.51	0.64	0.63	0.63	0.65
MultiNB	0.50	0.50	0.50	0.49	0.51	0.50	0.79	0.85	0.82	0.49	0.41	0.45	0.57	0.57	0.57	0.61
MLP	0.49	0.59	0.53	0.54	0.42	0.47	0.84	0.84	0.84	0.49	0.48	0.48	0.59	0.58	0.58	0.60

Table 5: Results for the shallow ML classifiers with TISEL affective lexical features using the SemEval dataset

	MultiNB			SVM			LR			MLP		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.56	0.44	0.49	0.49	0.48	0.49	0.57	0.51	0.54	0.44	0.44	0.44
Sad	0.50	0.31	0.38	0.41	0.37	0.39	0.46	0.41	0.43	0.38	0.39	0.38
Fear	0.49	0.61	0.54	0.52	0.60	0.55	0.49	0.70	0.58	0.53	0.57	0.55
Disgust	0.45	0.46	0.46	0.44	0.51	0.48	0.54	0.44	0.49	0.40	0.49	0.44
Surprise	0.69	0.61	0.65	0.73	0.69	0.66	0.74	0.65	0.69	0.69	0.62	0.66
Happy	0.40	0.60	0.48	0.49	0.53	0.51	0.44	0.59	0.50	0.46	0.48	0.47
Trust	0.58	0.45	0.51	0.59	0.53	0.56	0.67	0.47	0.55	0.58	0.51	0.54
Anticipation	0.65	0.59	0.62	0.70	0.60	0.65	0.82	0.56	0.67	0.67	0.56	0.61
Average	0.54	0.51	0.52	0.55	0.54	0.54	0.59	0.54	0.56	0.52	0.51	0.51
Acc	0.52			0.54			0.55			0.51		

Table 6: Results for shallow ML classifiers with TISEL affective lexical features using LAMA+DINA dataset

Nevertheless, when integrating the lexical affective features extracted from NRC, the best results were improved by between 1 and 2% in terms of F1-score and accuracy respectively, as shown in Table 7 and 8 for each dataset.

Therefore, LR performance achieved the most effective results in absolute terms compared to the other shallow ML models as in the previous cases. Nonetheless, it is worth highlighting the fact that LR is not the shallow ML algorithm which obtained the most remarkable improvement; the rest of techniques achieved improvements between 2 and 4% in relative terms, whereas LR just around 2%.

	Anger			Fear			Joy			Sadness			Average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
SVM	0.57	0.53	0.55	0.54	0.52	0.53	0.86	0.93	0.89	0.56	0.56	0.56	0.63	0.63	0.63	0.65
LR	0.59	0.58	0.58	0.61	0.55	0.58	0.77	0.90	0.83	0.61	0.52	0.56	0.64	0.64	0.64	0.66
MultiNB	0.54	0.52	0.53	0.50	0.48	0.49	0.79	0.87	0.82	0.65	0.53	0.54	0.60	0.60	0.60	0.62
MLP	0.51	0.51	0.51	0.52	0.47	0.50	0.87	0.89	0.88	0.53	0.56	0.54	0.61	0.61	0.61	0.63

Table 7: Results for shallow ML classifiers with NRC affective lexical features using the SemEval dataset

	MultiNB			SVM			LR			MLP		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.49	0.38	0.43	0.50	0.47	0.49	0.51	0.54	0.52	0.41	0.47	0.44
Sad	0.51	0.30	0.38	0.40	0.56	0.47	0.45	0.43	0.44	0.44	0.45	0.44
Fear	0.50	0.58	0.54	0.48	0.59	0.53	0.47	0.65	0.55	0.54	0.56	0.55
Disgust	0.52	0.45	0.48	0.41	0.44	0.43	0.62	0.47	0.53	0.49	0.44	0.46
Surprise	0.60	0.61	0.60	0.69	0.53	0.60	0.69	0.61	0.64	0.64	0.65	0.65
Happy	0.38	0.60	0.47	0.53	0.52	0.53	0.52	0.61	0.56	0.49	0.53	0.51
Trust	0.57	0.39	0.46	0.63	0.42	0.50	0.63	0.50	0.56	0.57	0.52	0.54
Anticipation	0.53	0.55	0.54	0.75	0.62	0.68	0.77	0.61	0.68	0.65	0.58	0.61
Average	0.51	0.48	0.49	0.55	0.52	0.53	0.58	0.55	0.56	0.53	0.52	0.53
Acc	0.49			0.53			0.56			0.53		

Table 8: Results for the shallow ML classifiers with NRC affective lexical features using the LAMA+DINA dataset

5.3 Deep learning classifiers

To detect the emotions in the SemEval dataset, three different DL algorithms were applied: CNN, LSTM and BiLSTM. To tune the hyper-parameters, a 10-fold cross validation over different parameter combinations was performed to select the best structures for the CNN, LSTM, and BiLSTM models.

5.3.1 Convolutional neural network (CNN)

After defining the CNN architecture, the hyper-parameters were tuned obtaining the best results for the values shown in Table 9.

Hidden layers	Hidden neurons	Epochs	Batch size	Optimizer
2	128, 64	50	8	RMSprop

Table 9: Tuned parameters for the CNN architecture

Several word embedding techniques were applied to extract features for classifying the emotions, and two lexicons, TISEL and NRC, to extract lexical affective features. Combining embeddings and lexicons, various configurations were executed for the CNN model. A summary of the results of these combinations is shown in Tables 10 and 11, respectively. Looking at Tables 10 and 11, the combination “embedding layer+NRC” obtained the highest accuracy and F1-score for both datasets, outperforming the rest of approaches by 1% approximately.

5.3.2 Long short-term memory (LSTM) network

In a similar manner, after defining the LSTM baseline architecture, the best hyper-parameters were tuned, obtaining the best results using the values shown in Table 12.

	Anger			Fear			joy			Sadness			Average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
Embedding Layer	0.54	0.61	0.57	0.64	0.49	0.55	0.84	0.86	0.85	0.57	0.59	0.58	0.65	0.64	0.64	0.65
Word2Vec	0.53	0.60	0.56	0.54	0.58	0.56	0.81	0.83	0.82	0.69	0.51	0.58	0.64	0.63	0.63	0.65
Embedding Layer+TISEL	0.55	0.65	0.59	0.60	0.49	0.54	0.88	0.84	0.86	0.58	0.59	0.59	0.65	0.64	0.65	0.66
Embedding Layer+NRC	0.57	0.64	0.60	0.52	0.63	0.57	0.89	0.84	0.87	0.71	0.51	0.59	0.67	0.66	0.66	0.67
Word2Vec+TISEL	0.59	0.58	0.58	0.54	0.51	0.52	0.81	0.88	0.84	0.61	0.60	0.61	0.64	0.64	0.64	0.65
Word2Vec+NRC	0.54	0.72	0.62	0.61	0.60	0.60	0.90	0.82	0.85	0.65	0.49	0.56	0.67	0.65	0.66	0.66

Table 10: Results for the CNN model using several embeddings and feature extraction methods using the SemEval dataset

	Embedding layer			Word2vec			Embedding layer+TISEL			Embedding layer+NRC			Word2vec+TISEL			Word2vec+NRC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.52	0.46	0.49	0.53	0.54	0.54	0.56	0.51	0.53	0.53	0.49	0.51	0.50	0.52	0.51	0.47	0.32	0.38
Sad	0.46	0.50	0.48	0.48	0.40	0.44	0.42	0.47	0.44	0.45	0.52	0.48	0.47	0.46	0.46	0.41	0.49	0.45
Fear	0.57	0.67	0.62	0.67	0.52	0.59	0.65	0.61	0.63	0.59	0.61	0.60	0.51	0.64	0.57	0.65	0.73	0.69
Disgust	0.59	0.44	0.51	0.54	0.43	0.48	0.50	0.58	0.54	0.52	0.49	0.51	0.59	0.45	0.51	0.45	0.46	0.46
Surprise	0.52	0.74	0.61	0.65	0.63	0.64	0.72	0.67	0.70	0.69	0.62	0.65	0.65	0.71	0.67	0.70	0.61	0.65
Happy	0.50	0.56	0.53	0.49	0.58	0.53	0.49	0.48	0.49	0.50	0.55	0.52	0.51	0.53	0.52	0.46	0.55	0.50
Trust	0.61	0.46	0.52	0.49	0.55	0.52	0.52	0.55	0.54	0.53	0.57	0.55	0.60	0.49	0.54	0.57	0.53	0.55
Anticipation	0.89	0.53	0.67	0.51	0.70	0.59	0.58	0.57	0.57	0.73	0.59	0.65	0.62	0.60	0.61	0.79	0.64	0.71
Average	0.58	0.55	0.55	0.55	0.54	0.55	0.56	0.55	0.55	0.57	0.56	0.56	0.56	0.55	0.55	0.56	0.54	0.55
Acc	0.55			0.55			0.56			0.57			0.56			0.56		

Table 11: Results for the CNN model using several embeddings and feature extraction methods for LAMA+DINA

The same combinations used for the CNN model were executed for the LSTM model, obtaining the results described in Table 13 and 14 for both datasets, respectively.

As in the previous case, NRC obtained the best results, but in this case, with the combination “word2vec+NRC”, achieving 0.57 and 0.67 of F1-score for LAMA+DINA and SemEval, respectively. Observing all of the previous results, the LSTM model performance was more satisfactory, achieving more significant results than the shallow ML and CNN models.

5.3.3 Bidirectional long short-term memory (BiLSTM) network

Finally, the BiLSTM model was optimized to achieve better results. In this case, the best evaluation results were obtained when the number of layers is 2 with 128 hidden neurons per layer. The rest of tuned hyper-parameters for conducting these experiments can be seen in Table 15.

Hidden Layers	Hidden Neurons	Learning Rate	Epochs	Batch Size	Dropout	Optimizer
2	128, 64	0.002	30	8	0.5	RMSprop

Table 12: Tuned hyper-parameters for the LSTM model

	Anger			Fear			joy			Sadness			average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
Embedding Layer	0.58	0.62	0.60	0.53	0.48	0.50	0.85	0.85	0.85	0.63	0.64	0.63	0.65	0.65	0.65	0.65
Word2Vec	0.55	0.60	0.57	0.62	0.51	0.55	0.85	0.85	0.85	0.59	0.57	0.58	0.65	0.63	0.64	0.65
Embedding Layer+ TISEL	0.54	0.69	0.61	0.64	0.36	0.46	0.82	0.87	0.84	0.64	0.64	0.64	0.66	0.64	0.64	0.66
Embedding Layer+ NRC	0.56	0.66	0.61	0.55	0.59	0.57	0.89	0.88	0.88	0.67	0.48	0.56	0.67	0.65	0.65	0.67
Word2Vec +TISEL	0.57	0.64	0.60	0.52	0.62	0.57	0.82	0.83	0.83	0.71	0.50	0.59	0.66	0.65	0.65	0.67
Word2Vec +NRC	0.54	0.67	0.60	0.62	0.58	0.60	0.86	0.87	0.86	0.70	0.53	0.60	0.68	0.66	0.67	0.68

Table 13: Results for the LSTM model using several embeddings and feature extraction methods using the SemEval Dataset

The same combinations were again executed for the BiLSTM model obtaining the results shown in Table 16 and 17 for SemEval and LAMA+DINA, respectively.

Analyzing the results, the combination “word2vec+NRC” obtained again the highest F1-scores, 0.58 and 0.68, for SemEval and LAMA+DINA, respectively. Furthermore, its performance achieved the most effective results compared to the rest of models.

5.4 Discussion

Comparing all of the shallow ML results, LR slightly outperformed the rest of algorithms for all possible combinations, obtaining the best results when adding affective lexical features from the NRC lexicon. SMV achieved similar results to LR, whereas MLP and MultiNB obtained the worst results as it can be seen in Fig. 2 and Fig. 3 in terms of accuracy for the SemEval and LAMA+DINA datasets, respectively, with respect to the accuracy measure. Overall, the effect of NRC is more positive than TISEL.

On the other hand, after applying all of the DL approaches (CNN, LSTM and BiLSTM), it is possible to conclude that the BiLSTM model achieved the best performance. Particularly, working with the combination “word2vec+NRC”, it outperformed the rest of the DL-based models in terms of accuracy as is depicted in Fig. 4 and Fig. 5 using the SemEval and LAMA+DINA datasets, respectively.

Analyzing all of the results, regarding the feature extraction methods, the integration of affective lexical features into the classifiers clearly led to improve the scores achieved by the baseline classifiers. The affective knowledge provided by NRC appears to be richer than TISEL’s, as the results show in Tables 10-17.

	Embedding layer			Word2vec			Embedding layer+ TISEL			Embedding layer+ NRC			Word2vec +TISEL			Word2vec +NRC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.67	0.42	0.52	0.50	0.45	0.47	0.50	0.49	0.50	0.60	0.46	0.52	0.56	0.51	0.53	0.65	0.47	0.55
Sad	0.44	0.45	0.45	0.46	0.49	0.47	0.46	0.45	0.45	0.44	0.52	0.47	0.51	0.47	0.49	0.46	0.55	0.50
Fear	0.55	0.59	0.57	0.61	0.67	0.64	0.61	0.57	0.59	0.67	0.65	0.66	0.66	0.55	0.60	0.62	0.68	0.65
Disgust	0.55	0.46	0.50	0.49	0.42	0.46	0.47	0.54	0.50	0.48	0.42	0.45	0.59	0.45	0.51	0.43	0.61	0.50
Surprise	0.69	0.66	0.68	0.63	0.63	0.63	0.61	0.68	0.64	0.62	0.69	0.65	0.83	0.59	0.69	0.66	0.58	0.63
Happy	0.46	0.61	0.53	0.48	0.53	0.50	0.52	0.54	0.53	0.46	0.57	0.51	0.40	0.68	0.50	0.56	0.51	0.54
Trust	0.49	0.57	0.53	0.53	0.53	0.53	0.60	0.51	0.55	0.68	0.47	0.56	0.65	0.52	0.58	0.58	0.56	0.57
Anticipation	0.71	0.62	0.64	0.68	0.66	0.65	0.69	0.64	0.66	0.63	0.66	0.64	0.56	0.71	0.63	0.77	0.60	0.68
Average	0.57	0.55	0.55	0.55	0.54	0.54	0.56	0.55	0.55	0.57	0.55	0.56	0.60	0.56	0.57	0.59	0.57	0.58
Acc	0.55			0.55			0.56			0.57			0.57			0.58		

Table 14: Results for the LSTM model using several embeddings and feature extraction methods using the LAMA+DINA dataset

Hidden Layers	Hidden Neurons	Learning Rate	Epochs	Batch Size	Dropout	Optimizer
2	128, 128	0.002	30	8	0.5	RMSprop

Table 15: Tuned hyper-parameters for the BiLSTM model

In regard to the used embedding model, word2vec provides, overall, a more accurate representation of the tweets, obtaining the best performance over the rest of algorithms.

Analyzing the results for both the DL and shallow ML methods, those based on DL outperformed based-shallow ML; nonetheless, it is worth mentioning LR can act in a similar manner to CNN. Therefore, in case of having to select one of them, it is good to mention the training process for LR is considerably shorter.

Comparing the conclusions here pointed out over the shallow ML algorithms against others published in [Plaza-del-Arco et al. 2020] regarding Spanish language, LR and SVM appear to be the best mechanisms for both languages (Spanish and Arabic). Nevertheless, comparing the lexicon performance, TISEL achieved better results than NRC for Spanish language, whereas in Arabic language, NRC should be the selected tool according to its results. For instance, using different algorithms, the use of NRC makes it possible to find sentences such as these ones, that cannot be correctly classified by TISEL:

مش إنسان اللي يعمل كذا مناظر مروعه للأطفال

“Not a person who does these horrific scenes for children”

الشوق يقتلنا دون أدنى رحمة والحنين كافر

“Longing kills us without the lowest mercy and the nostalgia is an infidel”

This fact can be explained because the original version of TISEL is written in Spanish, whereas both NRC and TISEL are translated lexicons, and the accuracy loss in the translation process can decrease the quality of the resulting vocabulary. Particularly, the translation process is very challenging for Arabic language because of the use of different terms depending on the type of used Arabic: modern standard Arabic (official language), dialectal Arabic or classical Arabic, language of Islam’s holy book. For that

	Anger			Fear			joy			Sadness			average			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	Acc
Embedding Layer	0.55	0.67	0.60	0.60	0.51	0.55	0.86	0.86	0.86	0.59	0.51	0.55	0.65	0.64	0.64	0.66
Word2Vec	0.58	0.69	0.63	0.52	0.64	0.57	0.86	0.80	0.83	0.74	0.52	0.61	0.68	0.66	0.66	0.66
Embedding Layer+ TISEL	0.58	0.65	0.61	0.57	0.60	0.59	0.90	0.85	0.87	0.71	0.56	0.65	0.69	0.67	0.68	0.68
Embedding Layer+ NRC	0.52	0.71	0.60	0.58	0.50	0.54	0.91	0.79	0.83	0.67	0.57	0.62	0.67	0.65	0.65	0.67
Word2Vec +TISEL	0.54	0.66	0.60	0.65	0.51	0.57	0.85	0.85	0.85	0.64	0.60	0.62	0.67	0.66	0.67	0.68
Word2Vec +NRC	0.56	0.68	0.62	0.60	0.64	0.62	0.88	0.86	0.87	0.74	0.55	0.64	0.69	0.68	0.68	0.69

Table 16: Results for the BiLSTM model using several embeddings and feature extraction methods using the SemEval dataset

	Embedding layer			Word2vec			Embedding layer+ TISEL			Embedding layer+ NRC			Word2vec +TISEL			Word2vec +NRC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Anger	0.54	0.47	0.50	0.49	0.46	0.47	0.49	0.46	0.47	0.59	0.38	0.46	0.65	0.47	0.55	0.79	0.40	0.54
Sad	0.43	0.50	0.46	0.45	0.48	0.47	0.45	0.48	0.47	0.47	0.66	0.55	0.46	0.55	0.50	0.50	0.54	0.52
Fear	0.52	0.64	0.57	0.68	0.56	0.62	0.68	0.56	0.62	0.57	0.58	0.58	0.62	0.68	0.65	0.57	0.72	0.64
Disgust	0.49	0.39	0.43	0.45	0.49	0.47	0.45	0.49	0.47	0.53	0.51	0.52	0.43	0.61	0.50	0.50	0.58	0.53
Surprise	0.71	0.65	0.68	0.61	0.64	0.63	0.61	0.64	0.63	0.84	0.59	0.69	0.66	0.59	0.63	0.73	0.62	0.67
Happy	0.55	0.57	0.56	0.48	0.58	0.52	0.48	0.58	0.52	0.56	0.55	0.56	0.56	0.51	0.54	0.50	0.60	0.54
Trust	0.46	0.55	0.50	0.57	0.53	0.55	0.57	0.53	0.55	0.44	0.69	0.54	0.58	0.56	0.57	0.69	0.59	0.63
Anticipation	0.72	0.57	0.63	0.72	0.64	0.68	0.72	0.64	0.68	0.71	0.60	0.65	0.77	0.60	0.68	0.72	0.60	0.66
Average	0.55	0.52	0.54	0.56	0.55	0.55	0.56	0.55	0.55	0.59	0.57	0.57	0.59	0.57	0.58	0.62	0.58	0.59
Acc	0.55			0.56			0.56			0.58			0.58			0.59		

Table 17: Results for the BiLSTM model using several embeddings and feature extraction methods for LAMA+DINA

reason, it is vital to develop specific tools for Arabic language instead of using translated dictionaries.

Focusing individually on the four emotions, both lexicons provided richer information regarding “joy” in SemEval because, for all possible combinations, the results are the best in our implementation and were also in [Plaza-del-Arco et al. 2020]. In LAMA+DINA, “anticipation” is the feeling which obtains higher accuracy. Furthermore, in Arabic the lowest scores were achieved for the class “sadness” and so were also they in Spanish, for both collections.

Finally, regarding the time to process all of the algorithms, the difference between the shallow ML and DL approaches is substantial. All of the algorithms were implemented and executed using Google Colab⁴, and while the shallow ML algorithms needed less

⁴ <https://colab.research.google.com>

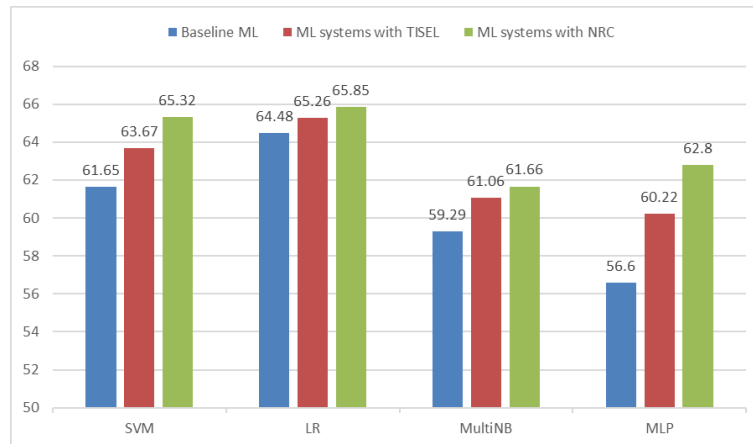


Figure 2: Accuracy for the different combinations of the shallow ML algorithms using the SemEval dataset

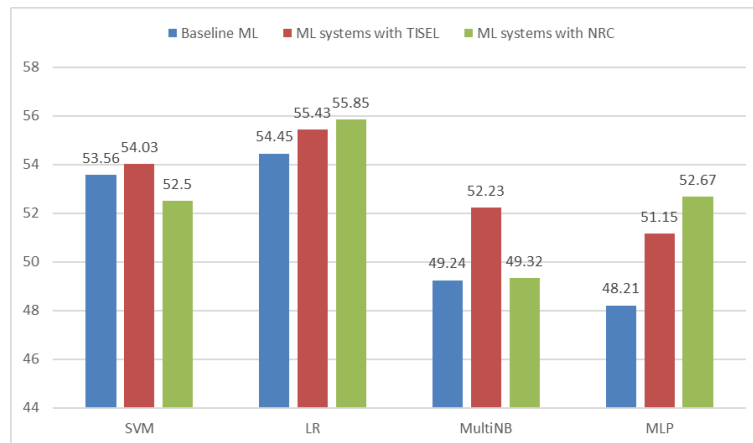


Figure 3: selectfontAccuracy for the different combinations of the shallow ML algorithms using the LAMA+DINA dataset

than one minute to be trained for both datasets, CNN needed around 30 minutes, LSTM around one hour and half, and BiLSTM 3 hours approximately, for SemEval, and twice as much time for LAMA+DINA for each algorithm. Therefore, although the DL approaches obtain better results, it is necessary to seriously consider the training time they involve when making a decision about which one to select.

The lemmatization and stemming steps considerably affect the training time because they sharply reduce the size of the datasets. On the one hand, the SemEval dataset was composed of almost 20,000 unique terms, and after applying lemmatization and stemming, 36.5% of the vocabulary was removed. On the other hand, the LAMA+DINA dataset was reduced by 38.1%, having initially almost 31,000 unique terms.

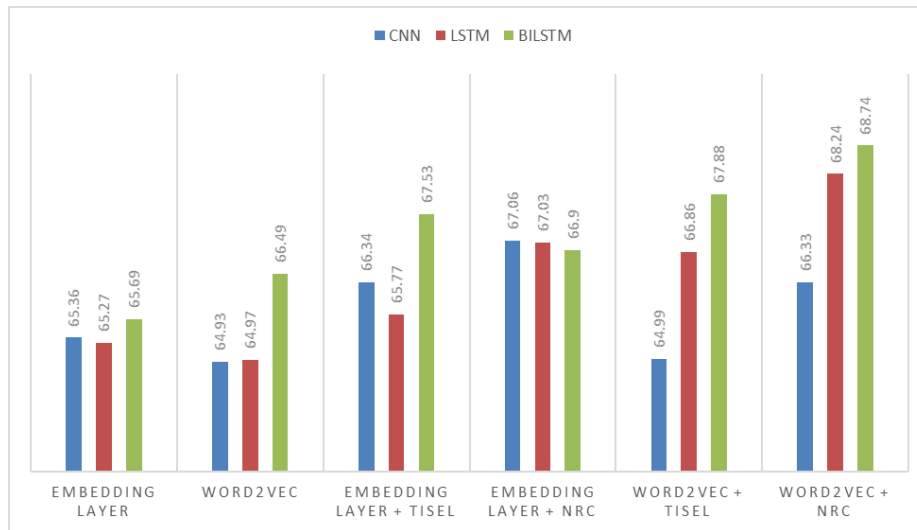


Figure 4: Accuracy for the different combinations of the DL algorithms using the SemEval dataset

6 Conclusions and future work

The present study compares different mechanisms for improving the task of emotion categorization on social media applied to Arabic language, by means of diverse combinations of AI techniques, features extraction techniques and emotional lexicons. For this comparison, two standard datasets in Arabic have been utilized.

From the experiments carried out, it is possible to conclude that the DL methods outperform the shallow ML ones, except for LR, which can achieve results close to CNN. Regarding the feature extraction, word2vec appears to be most adequate for the DL approaches, whereas with respect the affective lexicons, NRC provides more informative affective features than TISEL for Arabic language, in contrast to other studies in Spanish language.

Although it is difficult to find specific lexicons developed for Arabic language, the use of lexicons translated from other languages such as Spanish or English, clearly improves the results of the tested approaches. Nonetheless, the same dictionaries not necessarily perform in a similar manner in different languages, it is necessary to analyze every case in detail. Hence, the development and use of specifically designed lexicons for every language, and especially for Arabic, is a need of the utmost urgency when coping with social media texts.

The major limitation of this study is the quality of the used lexicons as well as the translation process. Most of the translation tools are not able to translate accurately many terms/expressions for several reasons, for instance, because the lexicons do not provide any context about the terms and it is not easy to disambiguate their meaning or, because the terms/expressions do not even exist in the other languages or, because there exist several dialects, among other reasons.

As future work, the development and use of a specific lexicon to test in Arabic is clearly necessary. In addition, the study of the effect of other lexical and syntactical aspects such as Arabizi terms or negations in Arabic is necessary.

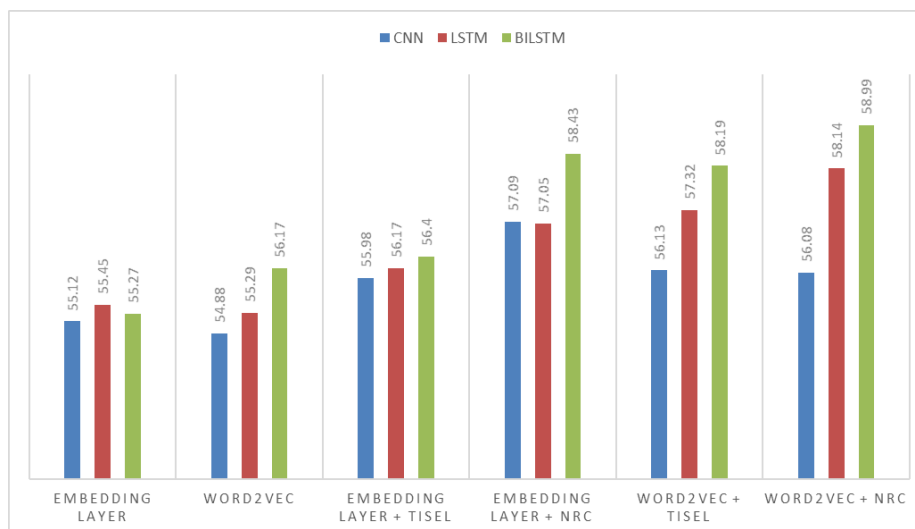


Figure 5: Accuracy for the different combinations of the DL algorithms using the LAMA+DINA dataset

Acknowledgments

This work has been partially supported by FEDER and the State Research Agency (AEI) of the Spanish Ministry of Economy and Competition under grant SAFER: PID2019-104735RB-C42 (AEI/FEDER, UE).

References

- [Abdul et al. 2016] Abdul-Mageed, M., AlHuzli, H., Abu-Elhij'a, D., Diab, M.: "Dina: A multi-dialect dataset for arabic emotion analysis"; In the 2nd Workshop on Arabic Corpora and Processing Tools, (2016), 29-37.
- [Abdul et al. 2020] Abdul-Mageed, M., Zhang, C., Hashemi, A., Nagoudi, E.M.B.: "AraNet: A Deep Learning Toolkit for Arabic Social Media"; LREC 2020 Workshop Language Resources and Evaluation Conference, (May 2020), 16-23.
- [Abdullah et al. 2018] Abdullah, M., Hadzikadicy, M., Shaikhz, S. SEDAT: "Sentiment and Emotion Detection in Arabic Text Using CNN-LSTM Deep Learning"; Proceedings of the 17th IEEE International Conference on Machine Learning and Applications, (2018), 835–840. <https://doi.org/10.1109/ICMLA.2018.00134>
- [Abdullaand Shaik2018] Abdullah, M., Shaikh, S.: "TeamUNCC at SemEval-2018 Task 1: Emotion Detection in English and Arabic Tweets using Deep Learning." (2018), 350–357. <https://doi.org/10.18653/v1/s18-1053>
- [Alhuzali et al. 2018] Alhuzali, H., Abdul-Mageed, M., Ungar, L.H.: "Enabling Deep Learning of Emotion With First-Person Seed Expressions"; Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, (2018), 25-35.
- [Alswaidan and Menai 2020] Alswaidan, N., Menai, M. E. B.: "Hybrid Feature Model for Emotion Recognition in Arabic Text"; IEEE Access, (2020), (Vol 8), 37843–37854. <https://doi.org/10.1109/ACCESS.2020.2975906>

- [AlZoubi et al. 2020] AlZoubi, O., Tawalbeh, S. K., AL-Smadi, M.: “Affect detection from arabic tweets using ensemble and deep learning techniques”; *Journal of King Saud University - Computer and Information Sciences*, (2020). <https://doi.org/10.1016/j.jksuci.2020.09.013>
- [Al-Khatib and El-Beltagy 2018] Al-Khatib, A., El-Beltagy, S. R.: “Emotional tone detection in Arabic tweets.”; In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*: Vol. 10762 LNCS. Springer International Publishing. (2018). https://doi.org/10.1007/978-3-319-77116-8_8
- [Al-Moslmi et al. 2018] Al-Moslmi, T., Albared, M., Al-Shabi, A., Omar, N., Abdullah, S.: “Arabic senti-lexicon: Constructing publicly available language resources for Arabic sentiment analysis”; *Journal of information science*, (2018), 44(3), 345-362. <https://doi.org/10.1177/0165551516683908>
- [Baccianella et al. 2010] Baccianella, S., Esuli, A., Sebastiani, F.: “SENTIWORDNET 3.0: An enhanced lexical resource for sentiment analysis and opinion mining”; *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, (2010), 2200–2204.
- [Baali and Ghneim 2019] Baali, M., Ghneim, N.: “Emotion analysis of Arabic tweets using deep learning approach”; *Journal of Big Data*, (2019), 6(1), 1–12. <https://doi.org/10.1186/S40537-019-0252-X/TABLES/7>
- [Bandhakavi et al. 2017] Bandhakavi, A., Wiratunga, N., Padmanabhan, D., Massie, S.: “Lexicon based feature extraction for emotion text classification”; *Pattern Recognition Letters*, (2017), (Vol 93), 133–142. <https://doi.org/10.1016/j.patrec.2016.12.009>
- [Bird et al. 2009] Bird, S., Klein, E., Loper, E. “Natural language processing with Python: analyzing text with the natural language toolkit”; “O’Reilly Media, Inc.”. (2009).
- [Birjali et al. 2021] Birjali, M., Kasri, M., Beni-Hssane, A.: “A omprehensive survey on sentiment analysis: Approaches, challenges and trends”; *Knowledge-Based Systems*, (2021), (Vol. 226), 107134. <https://doi.org/10.1016/j.knosys.2021.107134>
- [Bradley and Lang 1999] Bradley, M. M., Lang, P. J.: “Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings”, Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, (1999), 1-45.
- [Bravo-Marquez et al. 2014] Bravo-Marquez, F., Mendoza, M., Poblete, B.: “Meta-level sentiment models for big social data analysis”; *Knowledge-Based Systems*, 69(1), (1999), 86–99. <https://doi.org/10.1016/J.KNOSYS.2014.05.016>
- [Brownlee 2019] Brownlee, J.: “Develop Deep Learning Models On Theano And TensorFlow Using Keras”; *Journal of Chemical Information and Modeling*, (2019), 53(9), 1689–1699.
- [Cambria et al. 2020] Cambria, E., Li, Y., Xing, F. Z., Poria, S., Kwok, K. SenticNet 6: “Ensemble application of symbolic and subsymbolic AI for sentiment analysis”; *Proceedings of the 29th ACM international conference on information & knowledge management*, (October 2020), 105-114. <https://doi.org/10.1145/3340531.3412003>
- [Dreiseitl and Ohno-Machado 2002] Dreiseitl, S., Ohno-Machado, L.: “Logistic regression and artificial neural network classification models: A methodology review”; *Journal of Biomedical Informatics*, (2002), 35(5–6), 352–359. [https://doi.org/10.1016/S1532-0464\(03\)00034-0](https://doi.org/10.1016/S1532-0464(03)00034-0)

- [Duwairi et al. 2015] Duwairi, R. M., Ahmed, N. A., Al-Rifai, S. Y.: “Detecting sentiment embedded in Arabic social media—a lexicon-based approach”; *Journal of Intelligent & Fuzzy Systems*, (2015), (Vol. 29(1)), 107-117. <https://doi.org/10.3233/IFS-151574>
- [Duppada and Hiray 2017] Duppada, V., Hiray, S.: “SeerNet at EMOINT-2017: Tweet emotion intensity estimator”; *Proceedings of the EMNLP 2017 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)*, Copenhagen, Denmark, (September 2017), 205-211. <https://doi.org/10.18653/v1/w17-5228>
- [Ekman and Friesen 1969] Ekman, P., Friesen, W. V.: “The Repertoire of Non Verbal Behaviour-Categories, Origins Usage and Coding”; *Semiotica*, (1969), (Vol. 1), 49-98.
- [Guellil et al. 2018] Guellil, I., Azouaou, F., Mendoza, M.: “Arabic sentiment analysis: studies, resources, and tools”; *Social Network Analysis and Mining*, (2019), (Vol 9(1)), 1-17. <https://doi.org/10.1007/s13278-019-0602-x>
- [Hochreiter and Schmidhuber 1997] Hochreiter, S., Schmidhuber, J.: “Long Short-Term Memory”; *Neural Computation*, 9(8), (1997), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- [Jabreel and Moreno 2018] Jabreel, M., Moreno, A.: “Eitaka at semeval-2018 Task 1: An ensemble of n-channels convnet and xgboost regressors for emotion analysis of tweets”; *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018*, New Orleans, Louisiana, USA, June 5-6, (2018), 193-199. <https://doi.org/10.18653/v1/s18-1029>
- [Jayakrishnan et al. 2018] Jayakrishnan, R., Gopal, G. N., Santhikrishna, M. S.: “Multi-Class Emotion Detection and Annotation in Malayalam Novels”; *International Conference on Computer Communication and Informatics, ICCCI*, (2018), 6–10. <https://doi.org/10.1109/ICCCI.2018.8441492>
- [Kibriya et al. 2004] Kibriya, A. M., Frank, E., Pfahringer, B., Holmes, G.: “Multinomial naïve bayes for text categorization revisited”; *Lecture Notes in Artificial Intelligence*, (2004), 3339, 488–499. https://doi.org/10.1007/978-3-540-30549-1_43
- [Kim 2014] kim, Y.: “Convolutional Neural Networks for Sentence Classification”; *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, (2014), 1746–1751. <https://doi.org/10.3115/v1/D14-1181>
- [Larochelle et al. 2009] Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: “Exploring strategies for training deep neural networks”; *Journal of Machine Learning Research*, (2009), 10, 1–40. <https://doi.org/10.1145/1577069.1577070>
- [Lebret et al. 2016] Lebret, R. P., Bourlard, H., Collobert, R.: “Word Embeddings for Natural Language Processing”; *Ecole Polytechnique Fédérale de Lausanne*. (2016).
- [Ma et al. 2019] Ma, L., Zhang, L., Ye, W., Hu, W.: “PKUSE at SemEval-2019 Task 3: Emotion Detection with Emotion-Oriented Neural Attention Network”; *Proceedings of the 13th International Workshop on Semantic Evaluation*, (2019), 287–291. <https://doi.org/10.18653/v1/s19-2049>
- [Majeed et al. 2020] Majeed, A., Mujtaba, H., Beg, M. O.: “Emotion detection in roman urdu text using machine learning”; *Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering Workshops, ASEW*, (2020), 125–130. <https://doi.org/10.1145/3417113.3423375>
- [Mikolov et al. 2013] Mikolov, T., Chen, K., Corrado, G., Dean, J.: “Efficient estimation of word representations in vector space”; *1st International Conference on Learning Representations, ICLR 2013*, (2013).
- [Mohammad et al. 2018] Mohammad, S., Bravo-Marquez, F., Salameh, M., Kiritchenko, S.: “Semeval-2018 task 1: Affect in Tweets”; *In Proceedings of the 12th international workshop on semantic evaluation*, (2018), 1–17.
- [Mohammad and Kiritchenko 2015] Mohammad, S. M., Kiritchenko, S.: Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, (2015), 31(2), 301–326. <https://doi.org/10.1111/coin.12024>

- [Mohammad and Kiritchenko 2019] Mohammad, S. M., Kiritchenko, S.: “Understanding emotions: A dataset of tweets to study interactions between affect categories”; LREC 2018 - 11th International Conference on Language Resources and Evaluation, (2019), 198–209. <http://saifmohammad.com/WebPages/EmoInt2017.html>
- [Mohammad and Turney 2013] Mohammad, S. M., Turney, P. D.: “Nrc emotion lexicon”; National Research Council, Canada, 2. (2013).
- [Nielsen 2011] Nielsen, F. A.: “A new ANEW: Evaluation of a word list for sentiment analysis in microblogs”; CEUR Workshop Proceedings, 718, (2011), 93–98.
- [Oussous et al. 2020] Oussous, A., Benjelloun, F. Z., Lahcen, A. A., Belfkih, S.: “ASA: A framework for Arabic sentiment analysis”; Journal of Information Science, (2020), 46(4), 544–559. <https://doi.org/10.1177/0165551519849516>
- [Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É.: “Scikit-learn: Machine learning in Python”; Journal of Machine Learning Research, (2011), 12, 2825–2830.
- [Plaza-del-Arco et al. 2020] Plaza-del-Arco, F. M., Martín-Valdivia, M. T., Ureña-López, L. A., Mitkov, R.: “Improved emotion recognition in Spanish social media through incorporation of lexical knowledge”; Future Generation Computer Systems, (2020), 110, 1000–1008. <https://doi.org/10.1016/j.future.2019.09.034>
- [Polignano et al. 2019] Polignano, M., De Gemmis, M., Basile, P., Semeraro, G.: “A comparison of Word-Embeddings in Emotion Detection from Text using BiLSTM, CNN and Self-Attention”; ACM UMAP 2019 Adjunct - Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization. (2019). 63–68. <https://doi.org/10.1145/3314183.3324983>
- [Poria et al. 2017] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: “A review of affective computing: From unimodal analysis to multimodal fusion”; Information Fusion. (2017). (Vol.37), 98-125. <https://doi.org/10.1016/j.inffus.2017.02.003>
- [Rabie and Sturm 2014] Rabie, O., Sturm, C.: “Feel the heat: Emotion detection in Arabic social media content”; The International Conference on Data Mining, Internet Computing, and Big Data (BigData2014), (2014), 37–49.
- [Rayhan et al. 2020] Rayhan, M. M., Musabe, T. Al, Islam, M. A.: “Multilabel Emotion Detection from Bangla Text Using BiGRU and CNN-BiLSTM”; Proceedings of ICCIT 2020 - 23rd International Conference on Computer and Information Technology, (2020), 19–21. <https://doi.org/10.1109/ICCIT51783.2020.9392690>
- [Robertson 2004] Robertson, S.: “Understanding inverse document frequency: on theoretical arguments for IDF”; Journal of Documentation; Emerald Group Publishing Limited, (2004).
- [Singh and Shahid Husain 2014] Singh, P. K., Shahid Husain, M.: “Methodological Study Of Opinion Mining And Sentiment Analysis Techniques”; International Journal on Soft Computing, (2014), 5(1), 11–21. <https://doi.org/10.5121/ijsc.2014.5102>
- [Strapparava and Valitutti, 2004] Strapparava, C., Valitutti, A.: “WordNet-Affect: An affective extension of WordNet”; Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC (August 2004), 1083–1086.
- [Stubblebine 2007] Stubblebine, T.: “Regular Expression Pocket Reference: Regular Expressions for Perl, Ruby, PHP, Python, C, Java and .NET”; O’Reilly Media, Inc., (2007).
- [Suhasini and Srinivasu 2020] Suhasini, M., Srinivasu, B.: “Emotion Detection Framework for Twitter Data Using Supervised Classifiers”; Advances in Intelligent Systems and Computing, 1079(2016), (2020), 565–576. https://doi.org/10.1007/978-981-15-1097-7_47
- [Tong and Koller 2001] Tong, S., Koller, D.: “Support Vector Machine Active Learning with Applications to Text Classification”; Journal of Machine Learning Research, (2001), (Vol. 2), 45-66.

[Wu et al. 2018] Wu, C., Wu, F., Liu, J., Yuan, Z., Wu, S., Huang, Y.: “THU_NGN at SemEval-2018 Task 1: Fine-grained Tweet Sentiment Intensity Analysis with Attention CNN-LSTM”; Proceedings of the 12th International Workshop on Semantic Evaluation. (2018), 186–192. <https://doi.org/10.18653/v1/s18-1028>

[Xu et al. 2018] Xu, H., Liu, B., Shu, L., Yu, P. S.: “Double embeddings and cnn-based sequence labeling for aspect extraction”; ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, (2018), 2, 592–598. <https://doi.org/10.18653/v1/p18-2094>

[Zahiri and Choi 2017] Zahiri, S. M., Choi, J. D.: “Emotion detection on TV show transcripts with sequence-based convolutional neural network”; Workshops at the thirty-second AAAI conference on artificial intelligence, (2018), 44-51.

[Zhou et al. 2016] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: “Attention-based bidirectional long short-term memory networks for relation classification”; 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, (2016), 207–212. <https://doi.org/10.18653/v1/p16-2034>