

# Dialectal Arabic sentiment analysis based on tree-based pipeline optimization tool

Soukaina Mihi<sup>1</sup>, Brahim Ait Ben Ali<sup>1</sup>, Ismail El Bazi<sup>3</sup>, Sara Arezki<sup>2</sup>, Nabil Laachfoubi<sup>1</sup>

<sup>1</sup>IR2M Laboratory, Faculty of Science and Technology, University Hassan First, Settat, Morocco

<sup>2</sup>MISI Laboratory, Faculty of Science and Technology, University Hassan First, Settat, Morocco

<sup>3</sup>Systems Engineering Laboratory, University Sultan Moulay Sliman, Beni Mellal, Morocco

## Article Info

### Article history:

Received Feb 14, 2021

Revised Mar 25, 2022

Accepted Apr 15, 2022

### Keywords:

AutoML

Informal Arabic

Polarity detection

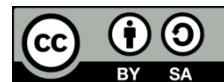
Sentiment analysis

Tree-based optimization tool

## ABSTRACT

The heavy involvement of the Arabic internet users resulted in spreading data written in the Arabic language and creating a vast research area regarding natural language processing (NLP). Sentiment analysis is a growing field of research that is of great importance to everyone considering the high added potential for decision-making and predicting upcoming actions using the texts produced in social networks. Arabic used in microblogging websites, especially Twitter, is highly informal. It is not compliant with neither standards nor spelling regulations making it quite challenging for automatic machine-learning techniques. In this paper's scope, we propose a new approach based on AutoML methods to improve the efficiency of the sentiment classification process for dialectal Arabic. This approach was validated through benchmarks testing on three different datasets that represent three vernacular forms of Arabic. The obtained results show that the presented framework has significantly increased accuracy than similar works in the literature.

*This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.*



## Corresponding Author:

Soukaina Mihi

IR2M Laboratory, Faculty of Science and Technology, University Hassan First

Settat, Morocco

Email: soukaina.mihi@uhp.ac.ma

## 1. INTRODUCTION

Sentiment analysis is a very active field of research positioned at the crossroads of natural language processing (NLP), text mining, and machine learning techniques. It involves analyzing a piece of text to retrieve the attitudes [1] and behavioral insights about an entity (product, feature, and service) or a feature of an entity. Subsequently, sentiment analysis tackles the study of the text in order to attribute a category of sentiment orientation [2], most commonly by detecting positive (favorable) or negative (unfavorable) polarity.

Since the substantial explosion of the world wide web, Internet users are not just information recipients anymore, but they contribute to effectively building up a myriad of publicly available content. This opinionated data is shared over social media within different user communities, where it spreads viewpoints about various topics such as politics, education, health systems, and product quality. Such subjective opinions may even alter the perception of reality and lead to a contention regarding controversial subjects. An example of this was the Arab spring when the Arab world was vigorously shaken in 2011 by a movement emanating from claims on the social networks from societies contesting authoritarian modalities of governance they have been undergoing over the past decades

Therefore, states and organizations no longer take for granted the data disseminated on the web, and special attention is paid to social media to track and even monitor the commonly shared information. In

particular, the content created by Arab internet users is increasingly growing due to the rising use of the internet and online services spurred by the coronavirus disease (COVID-19) context. As though, scientific papers dealing with sentiment analysis for the Arabic language have experienced a strong growth in recent years [3]–[5]. Researchers are addressing the issues of sentiment analysis to overcome the scarcity of tools and corpora available for the three forms of Arabic language, namely Standard Arabic, Classical Arabic and Dialectal Arabic [6]. Translated with [www.DeepL.com/Translator](http://www.DeepL.com/Translator) (free version) this paper aims to propose a model based on the tree-based optimization tool (TPOT) [7] to improve hyperparameters of machine learning algorithms. We use this model for the polarity classification of sentiments derived from tweets in the dialectal Arabic language.

The remainder of this paper is organized: section 2 gives an overview of the existing works related to sentiment analysis methods then highlights the most challenging issues of processing the Arabic language. In section 3, the related works are presented with emphasis on AutoML frameworks, especially the TPOT. Next, in section 4, we detail the materials and proposed method to show the results in section 5. Lastly, we finish with a conclusion and future works.

## 2. BACKGROUND RESEARCH

### 2.1. Sentiment analysis overview

The applications of sentiment analysis have a broad landscape ranging from marketing to politics. With the era of the contextual pandemic COVID-19, almost all activities have become digital, forcing people to search for online services and thus to consult the opinions of other users who may have already tested them. People and companies are interested in knowing how many positive reviews there are about a product, a company, or a feature. Sentiment analysis encompasses the study of different granularities of a piece of text. Namely, we distinguish the document level, the sentence level, and the aspect level. The document level refers to an overall sentiment as expressed through the entire text, whether the document consists of review, comment, and tweet. A document can be written in several phrases, and we assume that a document only expresses a unique sentiment regarding a specific subject. Conversely, the sentence level is defined by studying a sentence to infer its subjectivity and subsequently identify the sentiment and opinion it conveys. On the other side, the aspect level, sometimes referred to as the feature level, is the most fine-grained layer since it allows highlighting the sentiment towards a given target.

In all three levels, sentiment analysis can be conducted through three methods, automated machine-learning techniques, lexicon-based approaches, and hybrid ones [8]. Machine learning methods are further categorized into supervised, unsupervised, and semi-supervised. Supervised learning considers the use of training documents to train algorithms that will classify test documents. The most known supervised algorithms for the sentiment analysis task are support vector machines (SVM), naïve Bayes (NB), decision tree (DT), and maximum entropy (ME) [9]–[12]. In contrast, unsupervised learning allows detecting common elements to group similar documents into clusters without having a training set. The k-means clustering algorithm is prominently used for that. Semi-supervised learning incorporates both labeled and unlabeled data to perform sentiment classification. The most commonly used algorithms for semi-supervised learning are the ensemble approaches such as boosting, bagging, and random forest (RF). The lexicon-based approach is based on a lexicon composed of a collection of terms. Each term conveys a known sentiment. This approach includes dictionary-based and corpus-based techniques. Finally, the hybrid approaches rely on using machine learning methods combined with sentiment lexicons to enhance classification accuracy.

Sentiment classification addresses the task of classifying documents under two or multiple categories. When two categories are involved, then the task consists of detecting polarity (positive/negative). Polarity detection is referred to as the binary classification task [13]. The ternary classification is also widely applied by adding the neutral/objective class. When there are more than three classes, it is called multi-way classification [14], where we can classify the materials according to emotional intensity. Moreover, it is possible to use other classes, including, for example, the sarcasm class or the mixed class.

Figure 1 summarizes the different techniques used for sentiment analysis. More recently, efforts have been made to create resources and tools in the discipline of affective computing and sentiment analysis (ACSA) [15], which focuses on emotion recognition, subjectivity detection and opinion target identification. SenticNet is among the most used resources in ACSA, it is interested in developing intelligent algorithms based on the concept-level knowledge, the objective is to tackle the cognitive and affective aspect of natural language that is not covered by only machine learning algorithms. The last released version of SenticNet [16] covers more than 100,000 commonsense concepts in the English language, It represents data by a semantic perspective instead of using a syntactical methodology.

Although the SenticNet initiative has significantly advanced the handling of ACSA-related tasks, one major constraint remains that it focuses entirely on English [17]. With the expansion of social

networking worldwide, many other languages are becoming more prominent in the internet databank. However, these languages have not gained enough interest and the works dedicated to them are still barely noticeable in comparison with English. Vilares *et al.* [18] proposed BabelSenticNet to handling 40 languages including Arabic. BabelSenticNet is created in two steps, a first version is based on statistical machine translation, then a matching between senticnet concepts and wordnet synsets is performed to ensure accuracy at the concept level.

A similar work was done in [19] where a concept based sentiment analysis system is proposed to handle Arabic concepts. The construction of Arabic SenticNet lexicon consists of two stages, the two-way translation and the process of extending the Arabic version of Wordnet senses. The system embeds the usage of a rule based semantic parser to comply with grammatical and morphology requirements of Arabic language. Although the proposed system achieved a 93% F-score by using the concept, the lexical and the word 2vec features, the tests were carried out on data set of news articles written in the modern standard Arabic, which unlike the informal dialect Arabic, have more structured and convenient phrases to be transformed in concepts.

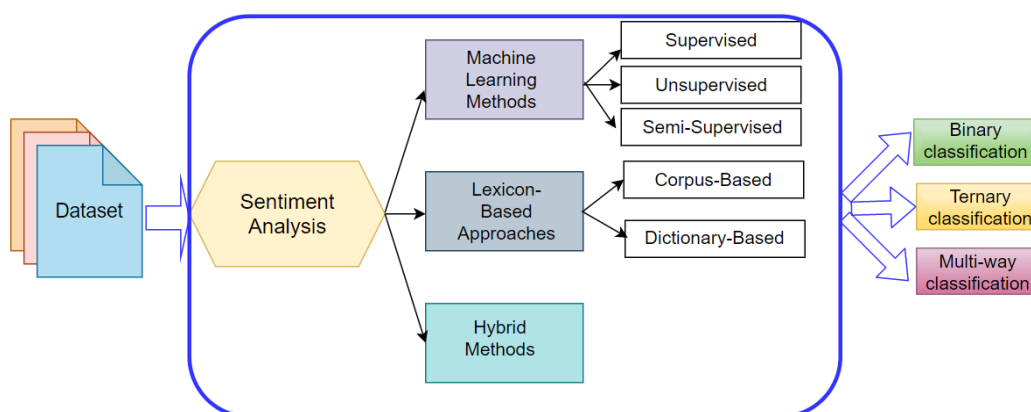


Figure 1. Techniques of sentiment analysis

## 2.2. Informal Arabic challenges

Arabic is a Semitic language widely used by more than 400 million people worldwide, of whom around 183 million are active internet users, placing Arabic in the fourth rank of the most used languages on the web. Arabic consists of 28 letters with different shapes according to their place, characterized by the absence of upper case letters and written from right to left, unlike English. There are three types of Arabic, classic Arabic being the Holy Quran's language, modern standard Arabic, the official language for education, the news, and all formal circumstances or events, and colloquial or informal Arabic is the simple way people talk to one another. All three Arabic types have some common morphological characteristics and are different in orthography, grammar, and lexicon [20]. The colloquial Arabic varies greatly depending on the geographical area. Generally, we consider five significant dialectal Arabic, the Egyptian, the Levantine, the Maghrebian, the Peninsular, and the Mesopotamian. Also, within every single dialect, several varieties exist, according to regions. Below, we flesh out some of the main challenges in processing colloquial Arabic text to analyze sentiments:

- Morphology: Arabic is considered to be a morphologically rich language (MRL) [21] due to its agglutinative and highly inflectional character compared to other languages [22]. This strong agglutination in Arabic generates an abundance of new words based upon a single morpheme such as stem or root. A morpheme represents the smallest significant letters unit [23] by appending clitics to a root; multiple words can be produced change form and shape according to their position within a sentence. For example, from the root kharaja/خرج, we can add affixes to create other verbs or nouns. Those affixes can be infixes, prefixes, or suffixes. The following Table 1 shows different words deduced from this root.
- Orthography and transliteration: Given the diverse variety of Arabic dialects, every dialect has distinctive orthography and lexicon. A unique word may be spelled in different manners for each dialect and even inside the same dialect. Thus, one may find several words that have the same meaning but spelled differently. Such a problem also arises when transliterating a word from another foreign language.

- Arabizi: Arabizi is a recent phenomenon in writing the Arabic language related to expanding social media and live chats. It consists of writing Arabic by using Latin alphabets while adding digits to match existing Arabic letters missing in Latin.
- Unstructured words: The most prevalent form of Arabic used on the internet is dialectal since users are increasingly providing more content in their communication language. While there is no standardized format for the various Arabic dialects, two persons can spell a given word very differently, thereby rendering the operation of getting the root of the word quite arduous. For instance, in the Moroccan dialect, one can write makanbghish/مكاتبغيش or makanbish/مكاتبيش, which means "I do not appreciate". Another common practice is to duplicate letters or tatweel [24] to stretch some Arabic letters.

Table 1. Example of the derivational character of the Arabic language

Root	Affixe	Affixe type	New word	Meaning
Kharaja/خرج	Ist/است	Prefix	Istakhraja/استخرج	To extract
	Aa/ا	Infixe	Kharij/خارج	Outside
	Ta/ت	Prefix	Takharaja/تخرج	To graduate
	Ou/و	Suffix	Kharajou/خرجوا	They went out

### 3. RELATED WORK

The use of the Internet, social networks, and the internet of things has become inevitable for everyone, and the produced data represents an essential component for decision-making and analysis on general questions and concerns. Without data, no organization nor business can function today. It helps improve the decision making and give more insights about strategies and future projects. However, the large volume and the format of these data, derived from various sources, constitute significant difficulties for researchers and machine learning practitioners.

Machine learning is not a trivial task. A thorough study is needed to detect the most accurate algorithms, hyperparameters, and efficient feature selection methods depending on each domain. That is the main reason that AutoML [25], [26] is a convenient alternative to achieve outstanding performance while saving time and effort of searching for the appropriate parameters. An interesting analysis is given by [27] comparing four AutoML tools with human performance over 13 commonly used datasets, and the obtained results were impressive as they show that AutoML tools outperform the machine learning process achieved by human data scientists in 4 of 13 tasks.

Moreover, [28], [29] present two TPOT based methods for radar signal recognition, aiming to solve the existing problems of radar feature extraction and low recognition rate. TPOT is used to select and optimize classifier parameters to improve recognition accuracy. The experimental results of [28] enhance the overall radar signal recognition that reaches 94.42%. In contrast, Zhang *et al.* [29] managed to maintain a TPOT accuracy beyond 96% under different Signal-to-noise ratio changes.

Howard *et al.* [30] benchmark different feature text representation methods for social media posts derived from health forums to predict mental health states. They used TPOT and Auto-Sklearn to generate classifiers with features extracted from textual data. Another study was conducted in [31] to predict the clinical diagnosis of depressive individuals, and this study introduces the feature set selector (FSS).

To specify subsets of the features as separate datasets in order to reduce the computational time of TPOT. Their study indicated that TPOT exceeded the tuned extreme gradient boosting (XGBoost) model while implementing FSS improved the results significantly. The work presented in [32] applied TPOT-based machine learning (ML) to predict angiographic diagnoses of coronary artery disease (CAD). It compared TPOT-generated ML pipelines with selected ML classifiers, optimized with a grid search approach, and demonstrated through experiments the power of agnostic model selection performed with AutoML TPOT for predicting CAD diagnosis. Similarly, the study of [33] built a radionics model with TPOT to predict molecular parameters essential in diagnosing tumor entities. Most relevant features were extracted from fluid-attenuated inversion re-recovery (FLAIR) images and used to generate ten separate TPOT models. We accomplish the steps of feature selection, model selection, and parameter optimization using TPOT. According to model comparison, TPOT helped to optimize the model parameters automatically and found valuable features to enhance the model performance. It predicted the lethal brain tumor encoded by the mutation of a histone named diffuse midline glioma (DMG) mutation status in patients with an accuracy of 91.1%.

Other studies of automated machine learning using TPOT concern different fields. The study conducted in [34] attempted to establish a learning architecture for forecasting and trading stock indices. Zhou *et al.* proposed a cascaded model and evaluated its effectiveness by comparing its performance with

other TPOT models. Additionally, Ahlgren *et al.* [35] proposed a machine learning approach based on TPOT to predict the dynamic fuel oil consumption. They demonstrated that an optimized model is a reliable tool for decision support systems.

#### 4. METHOD

Machine learning has led to exciting results in many NLP tasks, pointing out the necessity of finding optimal hyper-parameters to optimize the algorithms and the preprocessing phase and selecting appropriate features. It requires specific knowledge and experience that depend on the domain of study, data type, and expected results. Nevertheless, the AutoML system has shown its effectiveness in many problems [36]. As far as sentiment analysis is concerned, we propose using an approach built on the TPOT [36], an iterative and powerful system that uses genetic programming techniques to optimize the pipeline and models. The components of the framework are described in Figure 2.

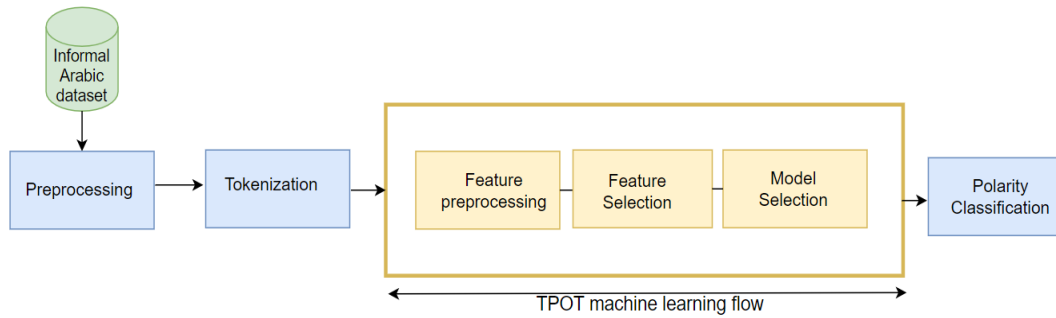


Figure 2. The system components for sentiment classification

##### 4.1. Tree-based pipeline

TPOT will search across a wide range of preprocessors, feature constructors, feature selectors, models, and parameters to find a set of operators that minimize the error of model predictions. Some of these operators are complex and can be time-consuming to perform, especially for large datasets. In this study, we consider four operators:

- Preprocessors: this operator scales the features using the mean and variance of the sample (StandardScaler), scales the features with the sample and the interquartile range (RobustScaler), and generates the interacting features by the polynomial combination of numerical features (PolynomialFeatures). When the number of characteristics is 4, and the degree is 2, the conversion by PolynomialFeatures can be expressed:

$$\sum_{k=1}^{15} x'_k = x_i \times x_j \quad (1)$$

where  $i \leq j$  and  $j = 0, 1, 2, 3, 4$ . When  $i = j = 0$  then  $x_i = x_j = 1$ .

- Decomposition: randomized principal component analysis [37] is applied to decompose the dimensionality reduction, using approximated singular value decomposition of the data and keeping only the most significant singular vectors to project the data to a lower-dimensional space.
- Feature selection: in TPOT, many feature selection methods are implemented, such as Select KBest, SelectPercentile, and VarianceThreshold. It uses a linear pipeline that follows a specified structure starting by feature selection where we can specify the method and enable the FeatureSetSelector parameter to reduce TPOT computation time.
- Model selection: TPOT was designed for supervised learning. The models integrate decision tree classifier, random forest classifier, gradient boosting classifier, support vector machine, logistic regression, and k-nearest neighbors classifier.

##### 4.2. Genetic programming

Genetic programming is a wonderfully powerful technology that emerged in the 90s [38]. It is a type of evolutionary algorithms that addresses automatic programming and machine learning problems. The genetic programming paradigm is founded on natural selection and biological breeding derived from the Darwinian evolution of living organisms. In this paper, optimizing TPOT pipelines is performed with genetic

programming, as presented in Figure 3. Each pipeline operator consists of a set of functions, with each function receiving a set of parameters. Parameters settings are specified as shown in Table 2.

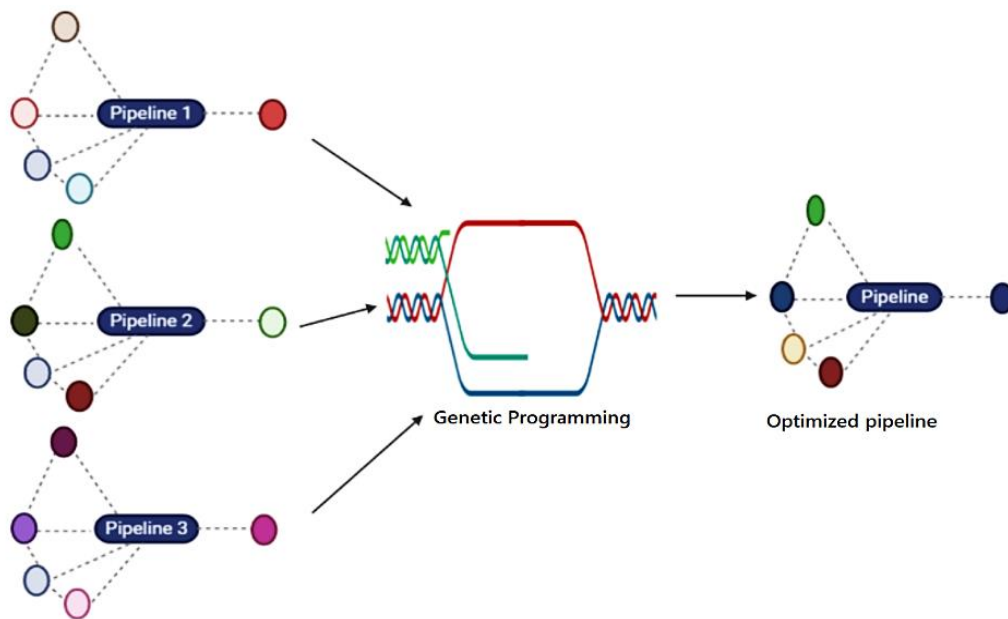


Figure 3. Optimizing machine learning pipelines with genetic programming

Table 2. Genetic programming parameters

Parameter	Value
Population Size	15
Generations	5
Mutation Rate	0.9
Crossover Rate	0.1
Selection	0.1
Scoring	Accuracy

- Population size: number of the tree-based pipelines to retain in the genetic programming population at the start of every generation.
- Generations: number of iterations to run the tree pipelines optimization process. The algorithm repeats this evaluation-selection-crossover-mutation process for 5 generations.
- Mutation rate: this parameter is in the range of [0.0, 1.0], it tells the algorithm how many pipelines to apply random changes for every generation.
- Crossover rate: this parameter is in the range of [0.0, 1.0]. It represents the number of times a crossover occurs for tree-based pipelines in one generation.
- Selection: this process determines which pipelines are allowed to survive and which pipelines are allowed to reproduce. Once a set of tree pipelines has been selected for further reproduction, the following operators are applied: reproduction, mutation, and crossover. 10% of the population is created from the best individuals that will constitute the new offsprings, and the tournament selection is used to determine the success rate of the population.
- Scoring: a scoring operation also called a fitness function, is applied to the process's outcome. We use accuracy in order to evaluate the quality of pipelines for classifying sentiments.

## 5. EVALUATION

There are mainly five vernacular forms of Arabic: Maghrebian Arabic in the North African region, Egyptian Arabic in the Nile region, Levantine Arabic, Peninsular Arabic in the Gulf region and Mesopotamian Arabic in the Iraqi region. This section presents the baseline of the pre-treatment phase performed on two datasets representing Arabic dialects among the most representative in the Arab World,

namely Moroccan and Egyptian dialects. We detail the vectorization method used and give the results for each configuration.

### 5.1. Datasets

Moroccan sentiment Twitter dataset (MSTD) [39] is a Moroccan dataset retrieved from tweets covering four-way sentiment classification. We are interested in the binary dataset. The second dataset Arabic sentiment Twitter dataset (ASTD) handles the Egyptian dialect and is characterized by its unbalanced settings as the positive class counts 799 tweets whilst the negative class counts more than 1,600 tweets. The third dataset addresses the Jordanian dialect, which we refer to as ArTwitter, and is a balanced dataset with 1,000 tweets for each class. We consider 70% for training and 30% for test purposes for the three datasets. Table 3 shows the description of MSTD, ASTD [40], and ArTwitter [41] datasets:

Dataset	Dialect type	Sentiment	
		Positive	Negative
MSTD	Moroccan	866	2769
ASTD	Egyptian	799	1684
ArTwitter	Jordanian	1000	1000

### 5.2. Preprocessing phase

Preprocessing text is so essential for all NLP tasks, especially when dealing with informal language. We first clean the text by removing noisy data such as punctuation, repeated letters, Urls, Html code, Hashtags, Usernames, and non-Arabic letters [42]. Next, we normalize text by replacing similar alphabet letters with one unique form, for example, "ا ا ا" are three forms of the same letter "ا", then we deleted diacritics and stop words. For that purpose, we built in a stop words list from both Moroccan and Egyptian dialects that match the most common conjunctions, prepositions, and non sentimental words also used in Jordanian dialect. After tokenizing the documents, we carry out the last phase of preprocessing by stemming words to their root. The individual stress response index (ISRI) stemmer [43] is employed to reduce words into their stem.

### 5.3. Term frequency-inverse document frequency (TF-IDF) vectorizer

The TF-IDF weighting approach is commonly used in information retrieval and text mining. This statistical metric can be used to assess the significance of a term in a document in relation to a collection or corpus. The weight grows in direct proportion to the number of times the term appears in the document. It also fluctuates depending on the word's frequency in the corpus. TF-IDF aims to convert a collection of raw documents to a matrix of TF-IDF features to express the importance of a word to a document while considering the relation to other documents in the corpus. The following formula calculates the TF-IDF score:

$$tfidf(w, d, D) = tf(w, d) \times idf(w, D) \quad (2)$$

while

$$f(w, d) = \log(1 + f(w, d)) \quad (3)$$

and

$$idf(w, D) = \log\left(\frac{N}{f(w, D)}\right) \quad (4)$$

with  $f(w, d)$  is the frequency of word  $w$  in document  $d$ ,  $N$  is the number of documents, and  $D$  is the collection of all documents.

## 6. RESULTS AND DISCUSSION

In this section, we present the experiments and the corresponding results. We first study the effect of changing the hyperparameters of the proposed framework on classification accuracy. All three datasets were prepared in the same manner, as explained in section 5. We show the effect of the presence or absence of the

stemming module, the TF-IDF with n-grams, and changing the evolutionary process parameters in TPOT. To make our results comparable with similar work in the literature, we used the accuracy metric, defined by:

$$Accuracy = \frac{(TP+TN)}{TP+TN+FP+FN} \quad (5)$$

where TP: true positive, TN: true negative, FP: false positive, and FN: false negative

For validation, the K-fold cross-validation method with k=5 was adopted. Cross-validation is a technique used in applied machine learning to estimate a machine learning model's skill on unknown data. That is, to use a small sample to assess how the model will perform in general when used to generate predictions on data that was not utilized during the model's training. It helps avoid the underfitting and overfitting of the proposed model by dividing the data into five subsets, each time one of the five subsets is used for validation, and the four other subsets are used to form a training set. K-fold cross-validation is a popular strategy since it's straightforward to grasp and produces a less biased or optimistic estimate of model competence than other approaches, such as a simple train/test split.

The tables of results show outcoming by dataset and benchmark test. We analyze the effect of using the n-grams and TF-IDF and how this affects the framework's performance. Corresponding results are shown in Tables 4, 5, and 6. According to outcomes, we perceived that the best accuracies were related to the combination 1g+2g for the three benchmarked datasets. The morphology of the Arabic language relatively explains this result, and this is because the shortest type of sentence having a semantic meaning comprises two words, such as the negation case, where a negation term precedes a verb to reverse sentiment shared between the various Arabic dialects.

Tables 7, 8, and 9 show accuracy measurements while using the root-stemming method and without stemming. The achieved results clearly show the benefits of stemming since the model performance is significantly improved. For the dataset MSTD, stemming words helped improve the accuracy by 0.024. These exciting findings arise from minimizing irrelevant features related to sentiment analysis by reducing the words to their root, thereby limiting the inflection level that is very high in Arabic and selecting the optimum feature set to be used during the evolution process by the TPOT. Moreover, we attempted to measure how far the evolutionary algorithm's parameters have been impacted by varying mutation rates, crossover rates, and the initial population size as shown in Tables 10, 11, and 12. The results were not affected by the variation of the mutation rate and the crossover rate, whereas the obtained measurements showed that accuracy was increased when increasing the population size. This enables that the model may explore novel pipelines through a selection of new offset springs. Due to the computational limitation, we have initially performed the process with a population size of 15 and later with 30. The sum of mutation rate and crossover rate needs to be lower than 1. We run other benchmarks with mutation rate=0.9 and crossover rate=0.1.

Table 4. Effect of TF-IDF with ngrams on ArTwitter

	TF-IDF 1g	TF-IDF 1g+2g	TF-IDF 1g+2g+3g
Accuracy	0.857	0.862	0.857

Table 5. Effect of TF-IDF with ngrams on ASTD

	TF-IDF 1g	TF-IDF 1g+2g	TF-IDF 1g+2g+3g
Accuracy	0.784	0.792	0.784

Table 6. Effect of TF-IDF with ngrams on MSTD

	TF-IDF 1g	TF-IDF 1g+2g	TF-IDF 1g+2g+3g
Accuracy	0.821	0.826	0.816

Table 7. Effect of stemming on ArTwitter

	Stemming=1	Stemming=0
Accuracy	0.862	0.812

Table 8. Effect of stemming on ASTD

	Stemming=1	Stemming=0
Accuracy	0.792	0.758

Table 9. Effect of stemming on MSTD

	Stemming=1	Stemming=0
Accuracy	0.826	0.802

Table 10. Results of changing TPOT parameters on ArTwitter

	Mutationrate=0.5	Cross-over=0.4	Population size=30
Accuracy	0.862	0.872	0.863

Table 11. Results of changing TPOT parameters on ASTD

	Mutation rate=0.5	Cross-over=0.4	Population size=30
Accuracy	0.784	0.771	0.793



Table 12. Results of changing TPOT parameters on MSTD

	Mutation rate=0.5	Cross-over=0.4	Population size=30
Accuracy	0.826	0.8207	0.829

Table 13 presents a comparison of the TPOT-based approach against other related approaches from the literature concerning accuracy. The comparison shows that for both datasets MSTD and ArTwitter, our proposed approach increased the accuracy considerably and outperforms other approaches based on convolutional neural network (CNN) and recurrent neural networks (RNN). On the other hand, our system gives comparable results for the ASTD dataset, with an accuracy of 79.3%, while the best-reported accuracy was given by the combined long short-term memory (LSTM) with Adam optimizer and reached 81.6%.

Table 13. Comparison with other related works

Dataset	Approach	Technique	Accuracy
MSTD	[39]	Farasa [44]+SVM	0.776
	Our System	Root Stemmer+TF-IDF+TPOT	0.829
ASTD	[45]	CNN non-static	0.759
	[46]	Combined-LSTM-Mul, non-static, continuous bag of words (CBOW), Adam optimizer	0.816
	[47]	Lexicon+SVM	0.751
	Our System	Root stemmer +TF-IDF+TPOT	0.793
ArTwitter	[45]	CNN non static	0.85
	[41]	Root stemmer+SVM	
	[47]	Lexicon+RNN	0.85
	Our System	Root stemmer +TF-IDF+TPOT	0.872

## 7. CONCLUSION




Machine-learning techniques have been widely exploited for NLP. It is challenging to find the correct hyperparameters and select the appropriate features. Concerning sentiment analysis, the Arabic language has gained widespread interest given its prevalence, prevalence, and difficulty as a morphologically complex language. Therefore, we have proposed a comprehensive framework for classifying sentiments as positive or negative and written in informal Arabic using different dialectal forms. This framework comprises a data preparation phase, a document cleaning process, and a stemming module. Afterward, we introduce preprocessed data into a TPOT-based module for the development of pipeline optimization. The results obtained are promising since we succeeded at improving the accuracy for the three benchmarked datasets. This work can be expanded to cover larger datasets involving multiple dialects at once. We also intend to handle evolutionary algorithms considering their significant contribution to the optimization of the sentiment analysis process.

## REFERENCES




- [1] B. Liu, *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge: Cambridge University Press, 2015.
- [2] T. Nasukawa and J. Yi, "Sentiment analysis: capturing favorability using natural language processing," in *Proceedings of the 2nd International Conference on Knowledge Capture, K-CAP 2003*, 2003, pp. 70–77, doi: 10.1145/945645.945658.
- [3] A. Ghallab, A. Mohsen, and Y. Ali, "Arabic sentiment analysis: a systematic literature review," *Applied Computational Intelligence and Soft Computing*, vol. 2020, pp. 1–21, Jan. 2020, doi: 10.1155/2020/7403128.
- [4] O. Oueslati, E. Cambria, M. Ben HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems*, vol. 112, pp. 408–430, Nov. 2020, doi: 10.1016/j.future.2020.05.034.
- [5] A. AlOwisheq, S. AlHumoud, N. AlTwaresh, and T. AlBuhairi, "Arabic sentiment analysis resources: a survey," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9742, 2016, pp. 267–278.
- [6] W. Zaghouni, "Critical survey of the freely available arabic corpora," *arXiv preprint arXiv:1702.07835*, Feb. 2017.
- [7] R. S. Olson and J. H. Moore, "TPOT: a tree-based pipeline optimization tool for automating machine learning," in *Automated Machine Learning*, Springer International Publishing, 2019, pp. 151–160.
- [8] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
- [9] R. M. Duwairi and I. Qarqaz, "Arabic sentiment analysis using supervised classification," in *International Conference on Future Internet of Things and Cloud*, Aug. 2014, pp. 579–583, doi: 10.1109/FiCloud.2014.100.
- [10] A. El Abdouli, L. Hassouni, and H. Anoun, "Sentiment analysis of noroccan Tweets using naive bayes algorithm," *International Journal of Computer Science and Information Security*, vol. 15, no. 12, pp. 191–200, 2017.
- [11] L. Almuqren, A. Alzammam, S. Alotaibi, A. Cristea, and S. Alhumoud, "A review on corpus annotation for Arabic sentiment analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10283, Springer International Publishing, 2017, pp. 215–225.
- [12] M. Elarnaoty, "A machine learning approach for opinion holder extraction in Arabic language," *International Journal of Artificial Intelligence and Applications*, vol. 3, no. 2, pp. 45–63, Mar. 2012, doi: 10.5121/ijaia.2012.3205.
- [13] M. N. Al-kabi, A. H. Gigieh, I. M. Alsmadi, and H. A. Wahsheh, "Opinion mining and analysis for Arabic language," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181–195, 2014.

- [14] E. Refaee and V. Rieser, "An Arabic twitter corpus for subjectivity and sentiment analysis," in *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, 2014, pp. 2268–2273.
- [15] B. Calabrese and M. Cannataro, "Sentiment analysis and affective computing: methods and applications," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10087 LNCS, Springer International Publishing, 2016, pp. 169–178.
- [16] E. Cambria, "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, no. 2, pp. 102–107, Mar. 2016, doi: 10.1109/MIS.2016.31.
- [17] H. G. Hassan, H. M. A. Bakr, and I. E. Ziedan, "A framework for arabic concept-level sentiment analysis using SenticNet," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 4015–4022, Oct. 2018, doi: 10.11591/ijece.v8i5.pp4015-4022.
- [18] D. Vilares, H. Peng, R. Satapathy, and E. Cambria, "BabelSenticNet: a commonsense reasoning framework for multilingual sentiment analysis," in *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Nov. 2018, pp. 1292–1298, doi: 10.1109/SSCI.2018.8628718.
- [19] A. Nasser and H. Sever, "A concept-based sentiment analysis approach for Arabic," *The International Arab Journal of Information Technology*, vol. 17, no. 5, pp. 778–788, Sep. 2020, doi: 10.34028/iajit/17/5/11.
- [20] M. Abdul-Mageed, M. Diab, and S. Kübler, "SAMAR: subjectivity and sentiment analysis for Arabic social media," *Computer Speech and Language*, vol. 28, no. 1, pp. 20–37, Jan. 2014, doi: 10.1016/j.csl.2013.03.001.
- [21] R. Aayed, A. Chouigui, and B. Elayeb, "A new morphological annotation tool for Arabic texts," in *IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA)*, Oct. 2018, pp. 1–6, doi: 10.1109/AICCSA.2018.8612798.
- [22] N. Y. Habash, *Introduction to Arabic natural language processing*, vol. 3, no. 1. Morgan and Claypool, 2010.
- [23] B. Joseph-gabriel, "The morphological disambiguation of Arabic (in French)," *Algerian Scientific Journal Platfroms*, vol. 6, no. 1, pp. 197–224, 2008.
- [24] S. Al-Osaimi and M. Badruddin, "Sentiment analysis challenges of informal Arabic language," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 2, 2017, doi: 10.14569/IJACSA.2017.080237.
- [25] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: automatic hyperparameter configuration for Scikit-learn," in *Proceedings of the 13th Python in Science Conference*, 2014, pp. 32–37, doi: 10.25080/Majora-14bd3278-006.
- [26] F. Mohr, M. Wever, and E. Hüllermeier, "ML-Plan: automated machine learning via hierarchical planning," *Machine Learning*, vol. 107, no. 8–10, pp. 1495–1515, Sep. 2018, doi: 10.1007/s10994-018-5735-z.
- [27] M. Blohm, M. Hanussek, and M. Kintz, "Leveraging automated machine learning for text classification: evaluation of AutoML tools and comparison with human performance," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*, 2021, vol. 2, pp. 1131–1136, doi: 10.5220/0010331411311136.
- [28] J. Wan, X. Yu, and Q. Guo, "LPI radar waveform recognition based on CNN and TPOT," *Symmetry*, vol. 11, no. 5, p. 725, May 2019, doi: 10.3390/sym11050725.
- [29] W. Zhang, P. Ge, W. Jin, and J. Guo, "Radar signal recognition based on TPOT and LIME," in *Chinese Control Conference (CCC)*, Jul. 2018, pp. 4158–4163.
- [30] D. Howard, M. M. Maslej, J. Lee, J. Ritchie, G. Woollard, and L. French, "Transfer learning for risk classification of social media posts: model evaluation study," *Journal of Medical Internet Research*, vol. 22, no. 5, May 2020, doi: 10.2196/15371.
- [31] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, Jan. 2020, doi: 10.1093/bioinformatics/btz470.
- [32] A. Orlenko *et al.*, "Model selection for metabolomics: predicting diagnosis of coronary artery disease using automated machine learning," *Bioinformatics*, vol. 36, no. 6, pp. 1772–1778, Mar. 2020, doi: 10.1093/bioinformatics/btz796.
- [33] X. Su *et al.*, "Automated machine learning based on radiomics features predicts H3 K27M mutation in midline gliomas of the brain," *Neuro-Oncology*, vol. 22, no. 3, pp. 393–401, Sep. 2019, doi: 10.1093/neuonc/noz184.
- [34] F. Zhou, Q. Zhang, D. Sornette, and L. Jiang, "Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices," *Applied Soft Computing*, vol. 84, Nov. 2019, doi: 10.1016/j.asoc.2019.105747.
- [35] F. Ahlgren, M. E. Mondejar, and M. Thern, "Predicting dynamic fuel oil consumption on ships with automated machine learning," *Energy Procedia*, vol. 158, pp. 6126–6131, Feb. 2019, doi: 10.1016/j.egypro.2019.01.499.
- [36] M.-A. Zöllner and M. F. Huber, "Benchmark and survey of automated machine learning frameworks," *Journal of Artificial Intelligence Research*, vol. 70, pp. 409–472, Jan. 2021, doi: 10.1613/jair.1.11854.
- [37] P.-G. Martinsson, V. Rokhlin, and M. Tytgert, "A randomized algorithm for the decomposition of matrices," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 47–68, Jan. 2011, doi: 10.1016/j.acha.2010.02.003.
- [38] J. R. Koza, "Genetic programming: on the programming of computers by means of natural selection," *Biosystems*, vol. 33, no. 1, pp. 69–73, 1992.
- [39] S. Mihi, B. Ait, I. El, S. Arezki, and N. Laachfoubi, "MSTD: moroccan sentiment Twitter dataset," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 10, pp. 363–372, 2020, doi: 10.14569/IJACSA.2020.0111045.
- [40] M. Nabil, M. Aly, and A. Atiya, "ASTD: Arabic sentiment tweets dataset," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2515–2519, doi: 10.18653/v1/D15-1299.
- [41] N. A. Abdulla, N. A. Ahmed, M. A. Shehab, and M. Al-Ayyoub, "Arabic sentiment analysis: Lexicon-based and corpus-based," in *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT)*, Dec. 2013, pp. 1–6, doi: 10.1109/AEECT.2013.6716448.
- [42] T. Zerrouki, "PyArabic an Arabic language library for Python," *Pyarabic*, 2010. .
- [43] M. G. Syarief, O. T. Kurahman, A. F. Huda, and W. Darmalaksana, "Improving Arabic stemmer: ISRI stemmer," in *IEEE 5th International Conference on Wireless and Telematics (ICWT)*, Jul. 2019, pp. 1–4, doi: 10.1109/ICWT47785.2019.8978248.
- [44] H. Mubarak, "Build fast and accurate lemmatization for Arabic," in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 2019, pp. 1128–1132.
- [45] A. Dahou, S. Xiong, J. Zhou, M. H. Haddoud, and P. Duan, "Word embeddings and convolutional neural network for Arabic sentiment classification," in *26th International Conference on Computational Linguistics (COLING)*, 2016, pp. 2418–2427.
- [46] S. Al-Azani and E.-S. M. El-Alfy, "Hybrid deep learning for sentiment polarity determination of Arabic microblogs," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10635, Springer International Publishing, 2017, pp. 491–500.
- [47] K. Elshakankery and M. F. Ahmed, "HILATSA: a hybrid Incremental learning approach for Arabic tweets sentiment analysis," *Egyptian Informatics Journal*, vol. 20, no. 3, pp. 163–171, Nov. 2019, doi: 10.1016/j.eij.2019.03.002.




**BIOGRAPHIES OF AUTHORS**

**Soukaina Mihi**    holds an Engineer Degree from Cadi Ayyad University and a Masters degree from INSA Lyon in Artificial Intelligence. She is a Ph.D. student at the IR2M Laboratory, which stands for Informatics, networks, Mobility and Modeling in Faculty of Sciences and Technologies Hassan 1<sup>st</sup> University, Settat, Morocco. Her research interests deep learning and machine learning. Her current research focus on NLP and sentiment analysis especially in Arabic. She can be contacted at email: soukaina.mihi@uhp.ac.ma.






**Brahim Ait Ben Ali**    is a Computer Science Engineer, graduated from National School of Applied Sciences (ENSA) at Cadi Ayyad University of Marrakesh, Morocco). Since 2019, He is preparing his Ph.D. in the IR2M Laboratory, Department of Computer Science. Faculty of Sciences and Techniques, Settat, Morocco, at Hassan First University of Settat. He has published several papers in reputed journals and international conferences. His research interest is machine learning and deep learning for natural language processing and its application. He can be contacted at email: b.aitbenali@uhp.ac.ma.






**Ismail El Bazi**    holds a Doctorate in Computer Science from Hassan 1<sup>st</sup> University and an Engineering degree in Computer Engineering from Cadi Ayyad University. He is also certified in project management (PMP) and in Agile methods (PMI-ACP) since 2013. After 10 years of professional experience in the field of Software Engineering with International IT companies, he joined the Sultan Moulay Slimane University in 2019 as Assistant Professor. His research focuses are artificial intelligence, arabic natural language processing and data science. He can be contacted at email: ismail.elbazi@umsba.ac.ma.



**Sara Arezki**    is a Professor on computer science at the faculty of science and technologies (Hassan First University of Settat Morocco). She holds a Ph.D. in Computer Science (2013) from Hassan 2nd University, Casablanca, Morocco and he graduated in Computer Science (2009) in ENSIAS, Rabat, Morocco where she got her engineering's degree in computer science. Her main research interests Information system, digital transformation, and blockchain. She can be contacted at email: sara.arezki@uhp.ac.ma.



**Nabil Laachfoubi**    is a computer science professor at Hassan 1<sup>st</sup> University of Settat, Morocco. He defended his doctoral thesis in 2000 and continues research in various areas notably machine learning and computer vision. He published several papers in reputed journals. He can be contacted at email: nabil.laachfoubi@uhp.ac.ma.