Brahim CHEROUATI, Youcef FEKIR, Mohamed SENOUCI

DOI: 10.15199/48.2025.05.37

OPEN ACCESS

Automatic identification of Algiers dialect based on word-level machine learning and deep learning methods

Automatyczna identyfikacja dialektu algierskiego na podstawie metod uczenia maszynowego i głębokiego uczenia na poziomie słowa

Abstract. *Algiers dialect identification is a specific field of natural language processing (NLP) aimed at automatically predicting the Algiers dialect of a given text. This paper presents different methods for identifying the Algiers dialect based on NLP, machine learning, and deep learning methods. The results show that the long short-term memory (LSTM) algorithm achieves 97% sensitivity in discriminating the Algiers dialect from Modern Standard Arabic (MSA) and an average of 94% sensitivity in discriminating this dialect from all other dialects presented in the studied dataset.*

Streszczenie. *Identyfikacja dialektu algierskiego jest specyficzną dziedziną przetwarzania języka naturalnego (NLP) mającą na celu automatyczne przewidywanie dialektu algierskiego danego tekstu. W tym artykule przedstawiamy różne metody identyfikacji dialektu Algieru w oparciu o NLP, uczenie maszynowe i metody głębokiego uczenia się. Wyniki pokazują, że algorytm pamięci długo-krótkoterminowej (LSTM) osiąga 97% czułości w odróżnianiu dialektu Algieru od współczesnego standardowego języka arabskiego (MSA) i średnio 94% czułości w odróżnianiu tego dialektu od wszystkich innych dialektów przedstawionych w badanym zbiorze danych.*

Keywords: Algiers dialect, NLP, machine learning, deep learning.
Słowa kluczowe: dialekt algierski, NLP, uczenie maszynowe, uczenie głębokie.

Introduction

Dialect is a Key area of sociolinguistics, which explores how linguistic variations within a language reflect a community'svarioussocial,historical,andcultural dimensions.Sociolinguisticsexaminesthelinguistic structure of dialects and their geographical distribution, evolution, and role in speakers' social and cultural identity [1]. In Algeria, dialects vary considerably from region to region, incorporating Berber, Arabic, and French influences, and are important markers of regional and social identity [2]. Compared to other dialects, Arabic dialects are spoken, not written, and therefore have no predefined rules for writing them. However, the same word can have several orthographic forms, all acceptable since there are no reference writing rules. What's more, because these dialects are different from the Arabic language, they are also different from each other; for example, the dialects of the Maghreb differ from those of the Middle East [3].

The Algiers dialect is widely used as a conversational language in Algeria but it does not escape the linguistic problems mentioned above, as there are no predefined or stable rules for representing it formally. In recent years, research has focused on creating corpora of Arabic languages including Algiers dialects in order to analyze and process this type of language and automatically identify its linguistic characteristics.

To discriminate between Algiers dialects and the other dialects, sentences are analyzed and processed in the following steps: for machine learning methods, sentences are vectorized using the term frequency-inverse document frequency (TF-IDF) technique to extract relevant features. In contrast, for deep learning approaches, sentences are segmented into distinct words and then converted into dense vectors using an embedding layer [4].

In [5], The authors employed Linear Support Vector Machine (L-SVM), Bernoulli Naive Bayes (BNB), and Multinomial Naive Bayes (MNB) classifiers for Arabic Dialect Identification, focusing on word-level and sentence- level approaches. They enhanced their results by combining these classifiers' output using a voting procedure. They used two datasets: Parallel Arabic Dialect Corpus (PADIC), and a manually constructed dataset of Algiers dialects. Their approach achieved an average accuracy of 76%.

The authors in [6] investigated the impact of parallel corpora on Arabic dialect identification. They compared statistical and neural methods using both parallel and non- parallel corpora, extending the PADIC dataset with data from the Kabyle dialect. Using classifiers such as Naive Bayes, KNN, Logistic Regression (LR), and Convolution Neural Network (CNN), they achieved an average accuracy of 92%.

In this paper, two approaches were conducted to discriminate between the Algiers dialect and the other dialects present in the dataset, the first approach involved sentence-level analysis using TF-IDF combined with three machine learning classifiers: SVM, LR, and MNB, and the second approach focused on word-level analysis, which will be converted to a word embedding vector combined with the deep learning algorithms CNN and LSTM.

Methodology

We address the identification of Algiers dialects as a supervised learning task. The primary goal of the proposed methodology is to automatically determine the Algiers dialect of a given text using NLP techniques combined with machine learning or deep learning approaches. This framework is designed to aid linguistic research and support applications requiring accurate dialect classification. We approach the identification of Algiers dialects as a supervised learning problem.

The main objective of the proposed methodology is to automatically predict the Algiers dialect from a given text using NLP techniques combined with machine learning or deep learning methods. This framework aims to support linguistic research and applications that require precise dialect identification.

The methodology starts with loading the dataset, involving data preparation, the data is organized by dialect and

labeled for supervised learning. For preprocessing, text inputs are segmented and tokenized at the sentence level for machine learning approaches and at the word level for deep learning models. NLP techniques, such as TF-IDF, are used to extract features for machine learning models, while word embeddings are generated to provide input for deep learning models.

For classification, we test a variety of machine learning algorithms, including SVM, LR, and, MNB trained on sentence--level TF-IDF features. In parallel, we employ a deep learning approach combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks using word--level embeddings. Finally, the performance of each model is evaluated using standard metrics such as accuracy, sensitivity, and specificity to assess their effectiveness in discriminating Algiers dialects from other Arabic dialects presented in the PADIC dataset.

Fig. 1 shows the flowchart of the described methodology given above.

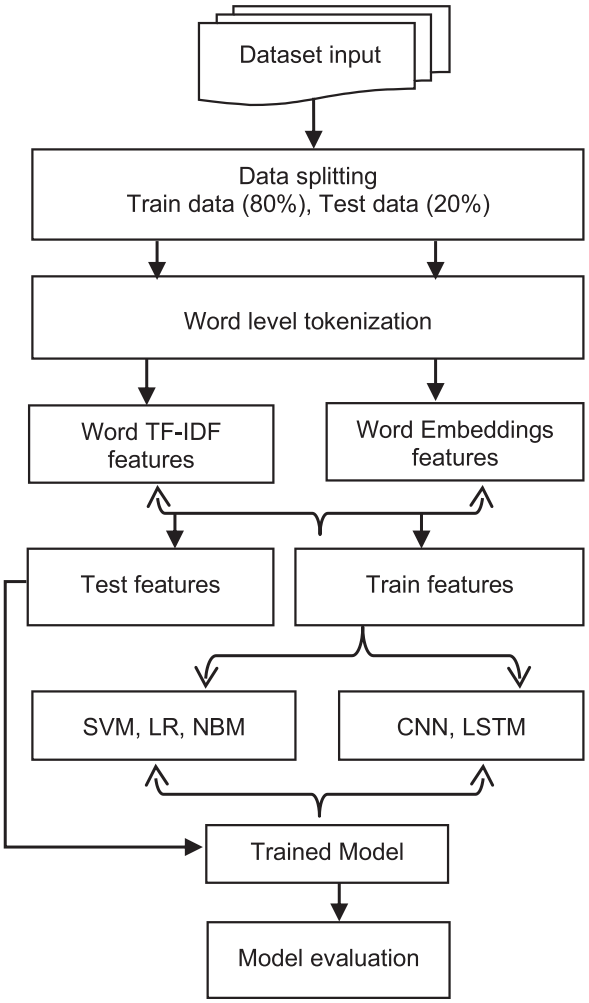

Fig. 1. Flowchart diagram of Algiers dialect identification

Dataset

This study focuses on five parallel dialects in addition to MSA (standard Arabic language) from PADIC (Parallel Arabic Dialectal Corpus) [7], this dataset is composed of the Annaba dialect (ANB), spoken in eastern Algeria; the Algiers dialect (ALG), used in the Algerian capital; the Sfax dialect (TUN), spoken in southern Tunisia; and the Syrian (SYR) and Palestinian (PAL) dialects, spoken in Damascus and Gaza, respectively. Table 1 presents some sentences from the PADIC dataset.

Table 1. Short text example from PADIC

Dialect	Text
MSA	أعجبتني كثيرا فكرة الرسائل الالكترونية بالعامية، لم نفكر فيها من قبل
ALG	عجبتني بزاف الفكرة تاع لي امايل بالدارجة ماخممناش فيها من قبل
ANB	عجبتني ياسر الفكرة تاع لي امايل بالدارجة ماخممناش فيها من قبل
TUN	عجبتني برشا فكرة لي امايل بالدارجة مفكرناش فيها من قبل
PAL	عجبتني كتير فكرة الرسائل الالكترونية بالعامية ، ما فكرناش فيها من قبل
SYR	عجبتني كتير فكرة الإيميلات بالعامية ما خطرتلنا من قبل
ENG	I liked the idea of emails in dialect, something we had never considered before.

Sentence and word processing

In this step, texts of the different dialects, as well as the texts of the standard Arabic language MSA, are segmented into distinct sentence and word vectors for further feature extraction. Sentence feature vectors are generated using TF-IDF, while word feature vectors are generated using the word-level embedding technique. These vectors are the input for the machine and deep learning algorithms.

a) Term frequency — inverse document frequency

Distinct sentences are extracted from the raw text dataset, and then distinct words are extracted from these sentences. TF-IDF is applied to the words to generate feature vectors. The formula for calculating TF-IDF is given by equations eq. (1), (2), and (3) [9]. The formula for calculating TF-IDF is given by the equations eq. (1), (2) and (3) [8].

$$TF - IDF = TF \times IDF \tag{1}$$

$$TF = \frac{Number\ of\ times\ term\ t\ appear\ in\ document}{Total\ number\ of\ terms\ in\ document\ d} \tag{2}$$

$$IDF = log(\frac{Number\ of\ the\ documents\ in\ the\ dataset}{Number\ of\ documents\ in\ the\ dataset\ contain\ the\ term\ t}) \tag{3}$$

Where the term t represents the distinct sentence generated by NLP text processing tools, table 2 shows an example of TF-IDF applied to MSA's sentence from table 1.

b) Word embedding

Unlike TF-IDF which represents words as sparse vectors based on their frequency, embeddings are dense vectors that capture semantic and contextual relationships between words [9]. Table 3 shows an example of the use of embeddings applied to the ALG's sentence from table 1.

Machine learning

Machine Learning is a branch of artificial intelligence that focuses on developing algorithms and statistical models to enable computers to learn patterns from data and make decisions or predictions without being explicitly pro-

Table 2. Word TF-IDF example from MSA language

Word	TF	IDF	TF-IDF
أعجبتني	0.0909	6.687	0.608
كثيرا	0.0909	4.639	0.422
فكرة	0.0909	**5.155**	**0.469**
الرسائل	0.0909	**7.157**	0.651
الالكترونية	0.0909	7.667	0.697
الالكترونية	0.0909	8.073	0.734
لم	0.0909	2.803	0.255
نفكر	0.0909	7.380	0.671
فيها	0.0909	5.103	0.464
فيها	0.0909	2.325	0.211
قبل	0.0909	4.895	0.445

Table 3. Word Embeddings Example from ALG dialect

Word	Dense embeddings vectors				
عجبتني	0,0428	0,0023	0,0232	0,0209	–0,0277
عجبتني	0,0458	0,0354	–0,0257	–0,0020	–0,0486
عجبتني	0,0425	–0,0087	–0,0342	0,0219	0,0034
تاع	0,0236	0,0292	0,0432	0,0485	0,0182
لي	0,0333	0,0378	0,0245	–0,0006	0,0081
امايل	0,0160	0,0290	–0,0285	–0,0136	–0,0043
امايل	0,0250	–0,0340	0,0488	0,0415	0,0234
ماخممناش	0,0098	0,0109	–0,0314	0,0466	–0,0245
فيها	0,0050	0,0419	–0,0096	–0,0064	–0,0480
من	–0,0164	0,0467	0,0074	0,0227	0,0426
قبل	–0,0128	0,0124	0,0396	0,0114	–0,0240

grammed. In the context of text classification or discrimination tasks, the following machine learning methods are commonly used

a) Support Vector Machines (SVM)

SVM is a supervised learning algorithm that finds the optimal hyperplane for separating data points of different classes. The objective is to find the hyperplane that maximizes the margin between two classes while minimizing the classification error. The predicted class in an SVM is determined using the following linear decision function [10].

$$f(X) = W^T X + b \tag{1}$$

Where W is the weight vector, X is the features vector, and b is the bias that defines the hyperplane position. The predicted class is given by the sign of f(x)

b) Logistic Regression (LR)

LR is a linear model that predicts the probability of a data point belonging to a specific class. It's widely used for binary and multiclass classification tasks and performs well with TF--IDF features by modeling the relationship between the input features and the target labels. LR used the sigmoid activation function as given by the formula (1) [11].

$$\sigma(a) = \frac{1}{1+e^{-a}} \tag{2}$$

Where $a=W^T X+b$, a linear combination of a feature vector X with the weights W and the bias b.

c) Multinomial Naive Bayes (MNB)

MNB is a probabilistic classifier based on Bayes' theorem, assuming independence between features. It is computationally efficient and performs well for text classification, especially with TF-IDF features, as it captures word frequency information in documents. The MNB P (xi|Ck) is computed using the formula given by [12] as follow:

$$P(x_i|C_k) = \frac{F(x_i,C_k)+\alpha}{\sum_{j=1}^{n}(F(x_j,C_k)+\alpha)} \tag{3}$$

where F (xi, Ck) represents the frequency of feature xi in all documents of class Ck, and α is a Laplacian smoothing parameter, used to avoid zero probabilities.

Deep learning

Deep learning is a subset of machine learning that uses multi-layered artificial neural networks to model and solve complex problems, such as text classification. CNN and LSTM are widely used in this field, CNN is trained to learn patterns across space while LSTM is trained to learn patterns across time [13].

Convolution Neural Network (CNN)

CNN network begins with an embedding layer that converts words into dense vector representations. Followed by a 1D convolutional layer, using filters and a kernel size to learn spatial patterns in the text. The max pooling layer reduces dimensionality and extracts prominent features. A dropout layer is added to prevent overfitting, and the feature map is flattened into a vector. The dense layers then process these features, with the final layer using a sigmoid activation function to output probabilities for binary classification. The model is compiled using binary cross- entropy as the loss function and the Adam optimizer for efficient training. The structure of the used CNN architecture is given in table 4.

Table 4. Structure of the proposed CNN

Layer	Feature Map	Total Parameters
Embedding	(None, 50, 100)	500,000
Convolution	(None, 46, 128)	64,512
Max Pooling	(None, 23, 128)	0
Dropout	(None, 23, 128)	0
Flatten	(None, 2944)	0
Dense	(None, 64)	188,16
Dense	(None, 1)	65

Long Short-Term Memory (LSTM)

In this network, we used an embedding layer to convert words into dense vectors, a dropout function to prevent overfitting, and an LSTM layer to capture temporal dependencies. A ReLU activation layer was added to learn complex patterns,

and finally, a sigmoid function was used for binary classification. The model was compiled with binary cross-entropy as the loss function and the Adam optimizer, making it effective for text classification tasks. The structure of the used LSTM is given in table 5.

Table 5. Structure of the proposed LSTM

Layer	Feature Map	Total Parameters
Embedding	(None, 50, 100)	500,000
Dropout	(None, 50, 100)	0
LSTM	(None, 100)	80400
Dense	(None, 64)	6464
Dense	(None, 1)	65

Results and discussions

Three metrics were selected to evaluate our model: accuracy, sensitivity, and specificity, which are defined as follows:

$$\text{Accuracy} = ((TP+TN)/(TP+FN+FP+FN))*100 \quad (1)$$

$$\text{Sensitivity} = (TP/(TP+FN))*100 \quad (2)$$

$$\text{Specificity} = (TN/(TN+FP))*100 \quad (3)$$

Where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative respectively.

Tables 6, 7, and 8 used word-level TF-IDF feature extraction, and an average sensitivity of 88% is reported for the identification between the Algiers dialect and the other language and dialects for the SVM, LR, and MNB classifier. On the other hand, a low sensitivity is reported for the Annaba dialect (ANB), which is due to its proximity to the Algiers dialect, as it is from the east of Algiers. Tables 9 and 10 used word-level embeddings, and an average sensitivity of 90% and 94% is reported for CNN and LSTM respectively.

In accordance with of these results, the use of embedding techniques is particularly promising compared to the TF-IDF approach. Whereas TF-IDF relies on a statistical representation of words based on their frequency of occurrence and importance in a given corpus, embeddings offer a dense, continuous representation of words, capturing semantic and contextual relationships. This comprehensive modeling allows embeddings to better understand linguistic nuances, particularly in the case of dialect identification, where lexical and syntactic variations play a crucial role. As a result, the integration of embeddings significantly improves model performance, particularly in terms of sensitivity and ability to generalize on complex data.

Table 6. SVM model for Algiers dialects identification (%)

ALG Vs	Accuracy	Sensitivity	Specificity
MSA	95	98	93
ANB	72	71	73
TUN	85	86	84
PAL	93	92	93
SYR	92	93	91
Average	87,4	88	86,8

Table 7. LR model for Algiers dialects identification (%)

ALG Vs	Accuracy	Sensitivity	Specificity
MSA	95	94	95
ANB	73	73	73
TUN	86	87	85
PAL	93	92	93
SYR	92	94	90
Average	87,8	88	87,2

Table 8. MNB model for Algiers dialects identification (%)

ALG Vs	Accuracy	Sensitivity	Specificity
MSA	96	96	97
ANB	73	71	75
TUN	87	85	88
PAL	94	94	94
SYR	93	93	94
Average	88,6	87,8	89,6

Table 9. CNN model for Algiers dialects identification (%)

ALG Vs	Accuracy	Sensitivity	Specificity
MSA	95	98	93
ANB	69	73	67
TUN	85	88	83
PAL	93	95	91
SYR	93	96	90
Average	87	90	84,8

Table 10. LSTM model for Algiers dialects identification (%)

ALG Vs	Accuracy	Sensitivity	Specificity
MSA	95	97	94
ANB	84	90	80
TUN	86	89	83
PAL	90	97	85
SYR	90	97	85
Average	89	94	85

Conclusion

This paper proposed a comparative study for identifying Algiers dialects using NLP tools, exploring machine learning and deep learning approaches. Machine learning algorithms, such as SVM, LR, and MNB, use word-level TF- IDF for feature extraction, while deep learning models, represented by CNN and LSTM, use word-level embeddings. The results show that deep learning methods give better performance for sensitivity than machine learning in the case of the Algerian dialect identification problem. In future work, we plan to extend our experiments by incorporating a larger dataset of the Algerian dialect and comparing the performance of our results with various other methods.

Authors: *Brahim Cherouati, Computer Science Department, University Mustapha Stambouli of Mascara, 29000 Mascara, Algeria, E-mail: brahim.cherouati@univ-mascra.dz, Youcef FEKIR, Computer Science Department, University Mustapha Stambouli of Mascara, 29000 Mascara, Algeria, E-mail: youcef.fekir@univ- mascara.dz, Mohamed Senouci, Department of Computer Sciences, University Ahmed Ben Bella Oran1, E-mail: msenouci@yahoo.fr*

REFERENCES

[1] J. Blommaert, The Sociolinguistics of Globalization, 1st ed. CambridgeUniversityPress,2010.doi: 10.1017/CBO9780511845307.

[2] O. Azzoug, “Sociolinguistic Dimensions of Dialect use in the Algerian Education System, ” Trad. Lang., vol. 10, no. 2, pp. 18–29, Dec. 2011, doi: 10.52919/translang.v10i2.853.

[3] S. Harrat, K. Meftouhy, M. Abbasz, K.-W. Hidoucix, and K. Smaili, “An Algerian dialect: Study and Resources, ” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 3, 2016, doi: 10.14569/IJACSA.2016.070353.

[4] D. Wójcik, “Uncovering insights with kanguage modeling. Analyzing Przeglad Elektrotechniczny with GPT, ” PRZEGLĄD ELEKTROTECHNICZNY, vol. 1, no. 12, pp. 353–356, Dec.

[5] 2023, doi: 10.15199/48.2023.12.69.

[6] M. Lichouri, M. Abbas, A. A. Freihat, and D. E. H. Megtouf, “Word-Level vs Sentence-Level Language Identification: Application to Algerian and Arabic Dialects, ” Procedia Comput. Sci., vol. 142, pp. 246–253, 2018, doi: 10.1016/j.procs.2018.10.484.

[7] M. Lichouri, K. Lounnas, and M. Abbas, “Evaluating the Influence of Parallel Corpora on Arabic Dialect Identification: A Comparative Study, ” in 2024 2nd International Conference on Electrical Engineering and Automatic Control (ICEEAC), May 2024, pp. 1–6. doi: 10.1109/ICEEAC61226.2024.10576299.

[8] K. Meftouh, S. Harrat, S. Jamoussi, M. Abbas, and K. Smaïli, “Machine Translation Experiments on PADIC: A Parallel Arabic Dialect Corpus, ” presented at the Pacific Asia Conference on Language, Information and Computation, Oct. 2015.

[9] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, “The Impact of Features Extraction on the Sentiment Analysis, ” Procedia Comput. Sci., vol. 152, pp. 341–348, 2019, doi: 10.1016/j.procs.2019.05.008.

[10] S. Wang, W. Zhou, and C. Jiang, “A survey of word embeddings based on deep learning, ” Computing, vol. 102, no. 3, pp. 717–740, Mar. 2020, doi: 10.1007/s00607–019–00768–7.

[11] [10] C. Cortes and V. Vapnik, “Support-vector networks, ” Mach. Learn., vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[12] [11] C. M. Bishop, Pattern recognition and machine learning. in Information science and statistics. New York: Springer, 2006.

[13] [12] A. McCallum and K. Nigam, “A Comparison of Event Models for Naive Bayes Text Classification”.

[14] [13] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, “Deep Learning--based Text Classification: A Comprehensive Review, ” ACM Comput.

[15] Surv., vol. 54, no. 3, pp. 1–40, Apr. 2022, doi: 10.1145/3439726.