Date of publication JUN-30, 2025, date of current version MAR-30, 2025.
www.computingonline.net / computing@computingonline.net

Print ISSN 1727-6209
Online ISSN 2312-5381
DOI 10.47839/ijc.24.2.4011

Harnessing Pretrained Models for Arabic Idiomatic Expression Identification: LLMs

SALMA TACE, MOSSAB BATAL, SOUMAYA OUNACER, SANAA EL FILALI, MOHAMED AZOUAZI

Department of Information Technology and Modeling Laboratory,
Faculty of Sciences Ben M'Sik, Hassan II University Casablanca, Morocco

Corresponding author: Salma Tace (salma.tace-etu@etu.univh2c.ma).

ABSTRACT Researchers have increasingly focused on idiomatic expressions in recent years, particularly Arabic idiomatic expressions. These phrases, often derived from ancient stories, are characterized by deeply idiomatic and non-compositional meanings. In this study, we explore the capabilities of large language models (LLMs) to understand and identify these expressions. After collecting data on Arabic idiomatic expressions, we carried out a preprocessing phase. We conducted a comprehensive set of experiments comparing two models, ChatGPT 4 and Arabic Bidirectional Encoder Representations from Transformers (AraBERT). Using 80% of the data for training and 20% for testing, our results reveal the strong ability of LLMs to identify idiomatic expressions, with performance reaching up to 95% in terms of F1 score and accuracy. In the second part of our study, we evaluate the efficacy of the pretrained AraBERT model in detecting idiomatic expressions, comparing it to baseline models, namely Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM) and Bidirectional Long Short-Term Memory (BiLSTM). The analyses show that the pretrained AraBERT model outperforms the conventional CNN-LSTM method by 14% in accuracy and F1 score, and also outperforms the BiLSTM model by 22%.

KEYWORDS Arabic Idiomatic Expressions; AraBERT; Multilingual BERT (mBERT); LLMS; Arabic Natural Language Processing; Deep learning; Accuracy; F1 score.

I. INTRODUCTION

The Arabic language is the official language in over 20 countries across the Middle East and North Africa, making it the fourth most used language on the internet. As of 2018, Arabic had over 164 million users in the Middle East and 121 million in North Africa. The complexity of Arabic, including its intricate morphology, syntax, and semantics, poses significant challenges for Natural Language Processing (NLP) [1, 2]. Creating universal NLP models that cater to all Arabic speakers is complicated by the language's diversity, unique grammatical structures, and regional variations [3].

The focus of Arabic Natural Language Processing (ANLP) in our research is on fixed expressions within the Arabic language. Our commitment is driven by the goals of transferring knowledge and technology to the Arab region, modernizing and enriching the Arabic language, and enhancing Arabic linguistics, particularly in semantic analysis. ANLP aims to provide Arab users with advanced capabilities for research, extraction, synthesis, and translation of information. Idiomatic expressions, increasingly studied by researchers in various languages, are a key area of focus in our research.

Arabic, enriched by idiomatic expressions from the pre-Islamic era and early Islam, still relies on these expressions in everyday communication, literature, and poetry. Widely found in Arabic literature, including the Quran, these expressions are part of a rich linguistic heritage that includes literary works and proverbs, prompting researchers to gather, categorize, and clarify them. Various classification systems have been developed to meet different needs, categorizing expressions as either continuous or discontinuous, and either rigid or adaptable.

The main challenge is that the meanings of some Arabic idioms cannot be directly understood from their literal components, especially those derived from historical tales or based on ancient Arabic grammar. Yet, these idioms are constructed within the framework of classical Arabic grammar and vocabulary. Identifying Arabic idiomatic expressions is a critical task in understanding the Arabic language due to their widespread use.

The primary goal of this research is to develop an advanced model for identifying idiomatic expressions in contemporary Arabic. The initial phase involves collecting extensive data on

Arabic idiomatic expressions to build a robust database. During the preprocessing stage, this database undergoes careful cleaning, including the removal of unnecessary punctuation, determiners, and prepositions. For example, the phrase 'رجعت بخفي حنين' is adjusted to 'رجع خف حنين'.

Ultimately, the objective is to develop a reliable and efficient model that can accurately identify these expressions across different texts and extensive corpora. This modeling effort is expected to significantly enhance the understanding and analysis of idiomatic expressions in modern Arabic.

This paper is structured as follows. Section 1 outlines the problem and research questions of this study. Section 2 reviews the literature to underscore the various challenges associated with processing the Arabic language and, specifically, idiomatic expressions. Section 3 describes our model in three parts: dataset construction, data cleaning, and evaluation. Section 4 details the results from testing our dataset with LLM models. Finally, Section 5 concludes the paper, summarizing the effectiveness of our model and outlining future research directions.

II. RELATED WORK

Idiomatic expressions, also known as fixed expressions or idioms, play a significant role in the linguistic and cultural tapestry of the Arabic language. These expressions, characterized by their fixed form and meaning, are utilized in specific contexts or situations and are not meant to be altered or modified. The study of idiomatic expressions has garnered considerable attention from researchers of various languages. Minko et al. (2003) conducted a comprehensive theoretical study on Arabic idiomatic expressions, classifying their morphosyntactic structures and analyzing the degree of syntactic and semantic idiomaticness to build a robust lexical database [4]. Hijab et al. (2014) focused on the structure and context of these expressions in the Saudi press, providing valuable insights into their usage and significance [5]. Kourtin et al. (2021) developed lexico-grammar tables for modern Arabic idiomatic expressions and implemented these expressions in NooJ dictionaries, enhancing their accessibility for natural language processing applications [6]. Jorge Baptista and colleagues (2004) explored the formal variation of idiomatic expressions in European Portuguese, building an electronic dictionary that addresses the challenges in natural language processing [7].

In the context of the Arabic language, significant work has been done on the translation of idiomatic expressions, particularly those found in the Quran. Mohamed Ali et al. (2016) identified and analyzed the problems posed by translating these expressions into French [8]. Mennat-Allah Abdelmaksoud (2018) further examined the translation of Quranic idiomatic expressions into French, shedding light on the idiomatic processes that characterize them [9]. Maurice Gross (1993) provided a foundational classification of idiomatic expressions in French, categorizing them based on the presence of fixed nominal groups in various syntactic positions [10]. Other notable studies include A. El-Mahdi et al. (2016), who developed a comprehensive framework for translating Arabic idiomatic expressions into English, highlighting the cultural nuances and contextual dependencies involved [11]. Alotaibi (2017) investigated the pedagogical implications of teaching Arabic idiomatic expressions to non-native speakers, emphasizing the importance of context and usage in language acquisition [12].

In addition to these foundational studies, recent advancements in natural language processing have introduced novel approaches for summarizing Arabic texts. Nada et al. (2020) proposed a versatile architecture for Natural Language Understanding (NLU) and Natural Language Generation (NLG) [13]. Alami and his team (2020) developed an effective method for addressing offensive language in Arabic on Twitter using AraBERT embeddings, achieving remarkable performance [14]. Faraj and Abdullah (2021) employed an ensemble technique with pretrained language models such as ChatGPT and AraBERT to tackle sentiment and sarcasm detection in Arabic, demonstrating the potential of modern NLP techniques in handling complex linguistic tasks [15]. Furthermore, Al-Shargi and Rambow (2015) explored the computational challenges of parsing Arabic idiomatic expressions, proposing a hybrid model that combines rule-based and statistical methods to enhance accuracy [16]. Mohammed and Haji (2019) conducted a diachronic study on the evolution of Arabic idiomatic expressions, tracing their historical roots and transformation across different periods of the Arabic language [17]. These studies collectively underscore the importance and complexity of idiomatic expressions in Arabic, highlighting the ongoing efforts to understand and process these unique linguistic elements.

The integration of idiomatic expressions in NLP systems has also been a focal point for enhancing machine translation and text generation. El-Haji et al. (2014) developed an annotated corpus of Arabic idiomatic expressions to support machine learning models, improving the accuracy of translations and contextual understanding [18]. Similarly, Alharbi et al. (2017) presented a method for identifying and translating Arabic idioms within text corpora, leveraging statistical techniques to differentiate between literal and idiomatic meanings [19]. This method has proven to be effective in reducing errors in automated translations, thereby enhancing the quality of machine-generated Arabic texts.

In educational settings, Bani-Khaled (2019) examined the role of idiomatic expressions in Arabic language learning, emphasizing their importance in achieving fluency and cultural competence among non-native speakers [20]. The study proposed instructional strategies that integrate idioms into language curricula, demonstrating improvements in learners' comprehension and communicative skills. Additionally, computational models like those developed by Al-Ghamdi et al. (2018) have been instrumental in creating educational tools that utilize idiomatic expressions to enhance interactive learning experiences for students [21].

Moreover, the impact of social media on the evolution and usage of Arabic idiomatic expressions was investigated by researchers like Saad et al. (2020). Their study analyzed the prevalence of idiomatic expressions in Arabic social media posts, revealing usage patterns that reflect contemporary linguistic trends and cultural shifts [22]. This research highlights the dynamic nature of language and the need for NLP systems to adapt to evolving linguistic phenomena. Furthermore, Al-Sulaiti and Atwell (2006) provided a comprehensive overview of the challenges in creating and maintaining a balanced Arabic corpus, stressing the importance of including idiomatic expressions to ensure the corpus's representativeness and utility in linguistic research [23].

III. PROPOSED MODELS

Our proposed model is designed with three key components, as illustrated in Figure 1. The initial stage involves collecting two distinct categories of expressions: one set of idiomatic expressions and another of nonidiomatic expressions. Training the model on these specific sets is crucial to effectively distinguish between these types of expressions.

The next stage is the preprocessing phase. To ensure the data meets high standards for model training, it is essential to clean and preprocess all expressions. This process involves removing unwanted characters, normalizing the text, and implementing other steps to enhance data quality.

In the final stage, we employ language models to make predictions. The inputs to the LLM (including ChatGPT 4 and AraBERT) consist of BERT-based and LLM models, both specially adapted for the Arabic preprocessed expressions categorized as either idiomatic or nonidiomatic. This categorization leads to the final outputs of our modeling pipeline.

Our system's approach to identifying static expressions involves systematic data collection and meticulous preprocessing, combined with natural language processing models. This approach enables us to achieve accurate results without errors across various languages, particularly in Arabic.

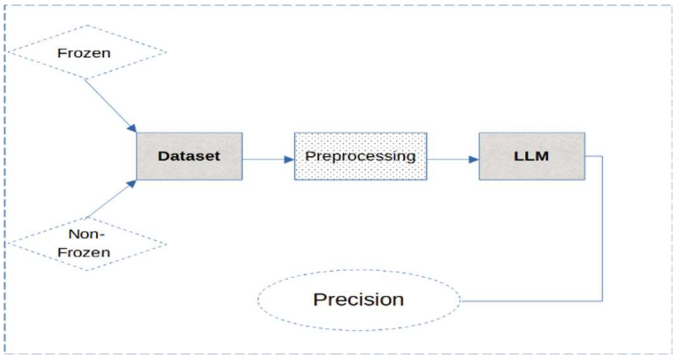


Figure 1. Data preprocessing pipeline

Our research delved deeper by consulting various Arabic dictionaries of idiomatic expressions, including "المعاصرة العربية في الإصطلاحي التعبير معجم," translated as "the dictionary of idioms in Modern Arabic" [24]. This crucial step allowed us to build a substantial dataset comprising approximately 1000 expressions—500 idiomatic and 500 nonidiomatic. After the manual data collection, we removed duplicates to ensure consistency, resulting in a dataset of 4,628 words: 1,800 words for idiomatic expressions and 2,828 for nonidiomatic. Additionally, punctuation marks appeared 200 times within idiomatic expressions and 130 times in nonidiomatic ones. Verbs were distributed with 450 in idiomatic expressions and 600 in nonidiomatic expressions. The detailed composition of this data is presented in Table 1. This distribution not only enriches our understanding of idiomatic structures but also lays a solid foundation for the subsequent phases of our research. The comprehensive database generated through this process serves as a valuable resource for further exploration and in-depth analysis in the broader context of Arabic linguistics.

Table 1. Dataset statistics

	Idiomatic expressions	Nonidiomatic expressions
Sentences	500	500
Words	1800	2828
Punctuations	200	130
Verbs	450	600
Lines	500	500

Preprocessing is a crucial step that prepares textual data for use by algorithms or machine learning models. This step enhances data quality, reduces noise, and facilitates more accurate analysis. Key preprocessing tasks performed by us include punctuation and stopword removal.

Punctuation removal: Texts undergo a process where punctuation marks are identified and eliminated, typically using predefined lists of punctuation symbols or regular expressions. For example, the phrase مِن شدّة حبه قلد من عَبس؛ becomes من شدة حبه قلد مِن عبس after punctuation is removed.

Stopword removal: Stopwords are commonly used words, such as "and," "the," "is," and "in", which are necessary for sentence structure but do not significantly contribute to meaning. This step removes such words from our data to focus on the core content.

For example, ما قل ودل؛ becomes قل دل. And هذا الشبل من ذاك الأسد reduces to الشبل الأسد after stopword removal.

In summary, the Punctuation and Stopword Removal process is a key preprocessing step aimed at cleaning and preparing textual data for more effective tasks in natural language processing (NLP) and text analysis.

Tokenization: Tokenization is the process of converting a text into a sequence of "tokens," which are meaningful units such as words or subwords. Tokenization divides a text into smaller elements called tokens, facilitating the analysis and further processing of the text. Tokens can be words, phrases, paragraphs, or even characters, depending on the desired level of granularity. Tokenization is fundamental in NLP and computational linguistics [25].

Stemming and lemmatization: These techniques aim to reduce words to their base forms. Stemming truncates affixes, while lemmatization brings words back to their canonical (lemma) form [26].

For example, the phrase "رجعت بخفي حنين" is reduced to رجع خف حني. We also normalize variations in Alef إ, أ, آ by converting them to ا and replaceTeh Marbuta ة with ه.

Data transformation for model compatibility: For input into the fine-tuned AraBERT model, we follow these steps:

1) Tokenize the input texts.

2) Convert each text to a BERT format by adding the [CLS] token at the start and [SEP] tokens between sentences and at the end.

3) Map each token to an index based on AraBERT's pretrained vocabulary.

The overall goal of preprocessing is to simplify and organize textual data, making it more accessible for NLP algorithms to understand and utilize [27].

The pretrained AraBERT model for Arabic represents a major advancement in natural language processing (NLP) for the Arabic language. Developed by the research community, AraBERT is based on the BERT (Bidirectional Encoder Representations from Transformers) architecture [28] and is specifically pretrained on a substantial amount of Arabic textual data. In this study, we fine-tune the pretrained AraBERT for Arabic text categorization. To accomplish this, we connect the AraBERT outputs, post fine-tuning, to an additional layer with a Softmax classifier to predict the text category.

First, each text is tokenized into N tokens, with the [CLS]

added at the beginning. An input representation Vi is then created for each token i, constructed by summing the vector embeddings corresponding to the token, its segment (i.e., sentence it belongs to), and its position in the text. These Vi vectors are fed into AraBERT, and its parameters are fine-tuned using labeled data from the corpus.

As illustrated in Figure 2, the final hidden vector for the special [CLS] token is denoted as k and the final hidden vector for each input token as Tn. The final hidden state h is taken as the representation of the entire text and is used as an input to the feed-forward layer with SoftMax classifier to obtain a probability distribution over the predicted output category (Sun et al., 2019) [29]:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d^k}}\right) \times V, \qquad (1)$$

where:

- Q is the vector representation of the query token;
- K^T is the transpose of the matrix K, which encodes the key token;
- V is the vector encoding of the value token;
- d^k denotes the dimensionality of the vector representations.

In all experiments, 80% of the data is allocated for training, and 20% for testing. We employ the AraBERT model (Antoun et al., 2020), pretrained on 70 million Arabic sentences. Our annalysis includes two versions of the AraBERT model: AraBERTv1 and AraBERTv2. Additionally, we compare AraBERT with the multilingual BERT model (Devlin et al., 2018), which is pretrained on 104 languages using data from the entire Wikipedia dump for each language, excluding user and talk pages. This multilingual BERT model has 12 layers, 768 hidden units, 12 attention heads, and 110 million parameters.

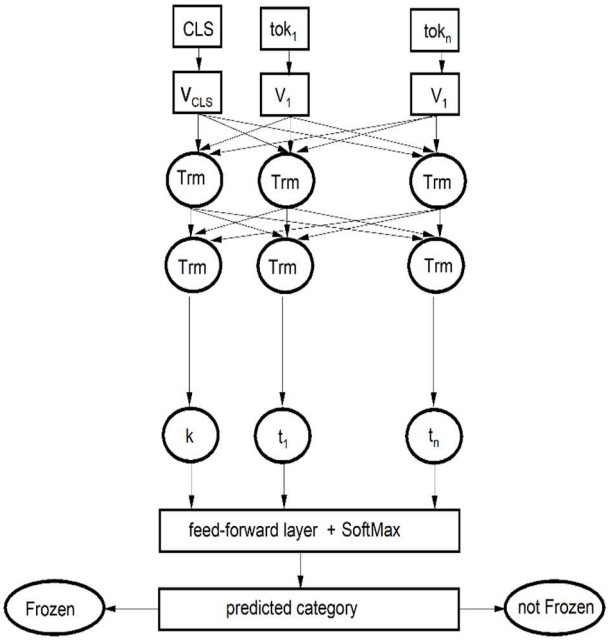


Figure 2. Pretraining model of AraBERT

Table 2 provides detailed descriptions of each version of AraBERT, highlighting their differences in the pretraining process [30].

Table 2. AraBERT v1 vs AraBERT v2.

Model	Size		Dataset		
	MB	Param	Sentences	Size	Words
AraBERT v2	543	136	200 M	77 Gb	8.6 B
AraBERT v1	543	136	77 M	23 Gb	2.7 B

IV. RESULTS

In this section, we explore two versions of the pretrained AraBERT model, including AraBERTv1 and AraBERTv2, and compare their performance with the mBERT model. Classification accuracy is used as a primary metric to evaluate model performance, representing the ratio of correctly predicted labels to the total number of samples in the testing dataset. Formally:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \qquad (2)$$

TP: True Positives (Samples the classifier has correctly classified as positives);

TN: True Negatives (Samples the classifier has correctly classified as negatives);

FP: False Positives (Samples the classifier has incorrectly classified as positives);

FN: False Negatives (Samples the classifier has incorrectly classified as nega-tives).

Since the dataset classes are highly unbalanced, we also evaluate performance using the F1-score, Recall and Precision, which offer a more balanced assessment in this context. These metrics are calculated as follows:

$$\boldsymbol{F1 - score} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \frac{1}{2}(\boldsymbol{FP} + \boldsymbol{FN})}, \qquad (3)$$

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}}, \qquad (4)$$

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}}. \qquad (5)$$

Table 3 provides the results for mBERT and AraBERT models in terms of Precision, Recall, Accuracy, and F1-score. These results are also visually represented in Figure 3.

Table 3. Performance results of AraBERTv2 and AraBERTv1 versus mBERT

Model		Accuracy (%)	Prescision (%)	Recall (%)	F1-score (%)
AraBERT v1	Nonidiomatic (0)	92	88	96	92
	Idiomatic (1)	92	96	88	92
AraBERT v2	Nonidiomatic (0)	95	97	93	95
	Idiomatic (1)	95	93	97	95
MBert	Nonidiomatic (0)	91	95	86	91
	Idiomatic (1)	91	88	96	91

All models deliver comparable results, with the mBERT model achieving 91% accuracy and an F1-score of 91% for the Idiomatic expressions (1), and 91% accuracy with an F1-score of 92% for non-idiomatic expressions (0). In contrast, the AraBERT model reaches up to 95% accuracy, indicating an approximate 4% improvement over the multilingual model. This difference can be attributed to the AraBERT model's exclusive pretraining on the Arabic language, which includes a larger dataset and vocabulary compared to mBERT, thereby enhancing word diversity and model performance.

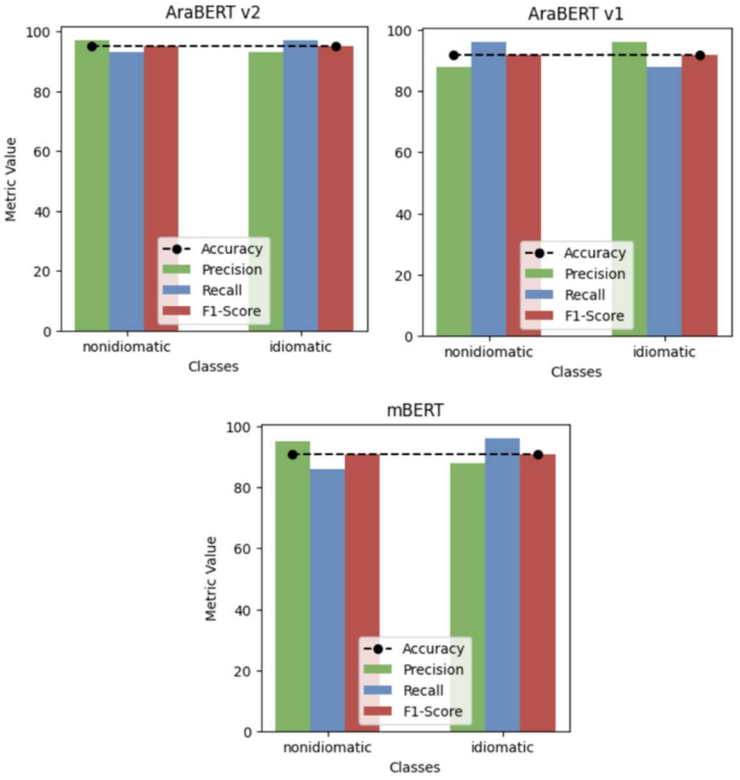


Figure 3. Accuracy, precision recall and F1-score of fixed and non-fixed expressions.

In the second part of our study, we examine the performance of pretrained AraBERT models in detecting idiomatic expressions, compared to baseline models, namely CNN-LSTM and BiLSTM. The BiLSTM model is a bidirectional recurrent neural network that employs Long Short-Term Memory (LSTM) mechanisms, allowing it to process sequential data by accounting for both past and future contextual information. On the other hand, the CNN-LSTM model combines Convolutional Neural Network (CNN) layers with LSTM layers, effectively integrating spatial and temporal features. Table 4 presents the results in terms of accuracy, precision, recall and F1-score of CNN-LSTM, LSTM and BiLSTM, and these results are also illustrated in Figure 4 (a, b). The CNN-LSTM model demonstrates high performance in detecting both idiomatic (1) and nonidiomatic expressions (0), achieving an accuracy of 81% compared to the BiLSTM model's accuracy of 73%.

Table 4. Performance results of CNN-LSTM and BiLSTM

Model		Accuracy (%)	Prescision (%)	Recall (%)	F1-score (%)
BiLSTM	Nonidiomatic (0)	73	80	67	72
	Idiomatic (1)	73	71	81	78
CNN-LSTM	Nonidiomatic (0)	81	79	84	82
	Idiomatic (1)	81	83	78	80

The performance of the pretrained CNN-LSTM and BiLSTM models is shown in Figure 4 (a, b). The classical CNN-LSTM model achieves 81% accuracy and F1-score, outperforming BiLSTM model, which reaches 73% in these metrics.

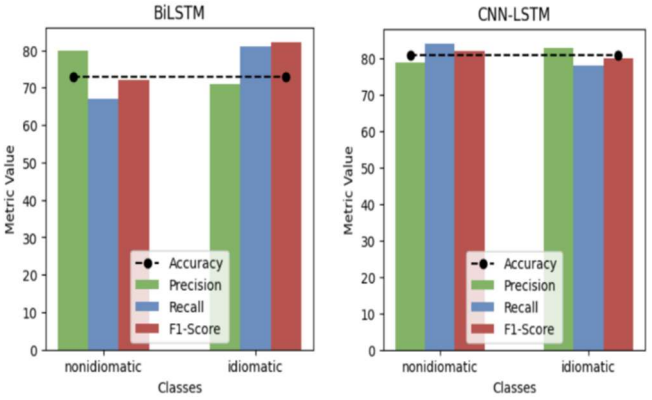


Figure 4. (a,b) Accuracy, precision recall and F1-score of idiomatic and nonidiomatic expressions using CNN-LSTM and BiLSTM

Table 5. Performance results of LLM

Model		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
LLM	Nonidiomatic (0)	91	93	89	91
	Idiomatic (1)	91	90	93	91

In this section, we utilized OpenAI's GPT-4, a prominent variant of large language models (LLMs) used in artificial intelligence and natural language processing, to evaluate our dataset. These models are trained on extensive textual datasets, enabling them to understand and generate complex text. As shown in Table 5, which displays the results for the original embeddings applied to idiomatic and nonidiomatic expressions, our models achieved an F1-score and accuracy of 91%. These results are illustrated in Figure 5.

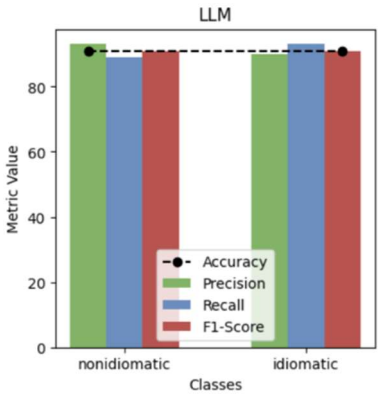


Figure 5. Accuracy, precision, recall and F1-score for detecting idiomatic and nonidiomatic expressions using Large Language Models (LLMs).

The obtained performance of the pretrained AraBERT model and LLM compared to CNN-LSTM and BiLSTM models is presented in Figure 6. The AraBERT models and LLM outperform the classical CNN-LSTM method by 14% and 10% in terms of accuracy and F1-score, respectively. They also exceed BiLSTM by 22% and 18%. It is unsurprising that the AraBERT model and LLM achieve the best performance since they incorporate context from both directions.

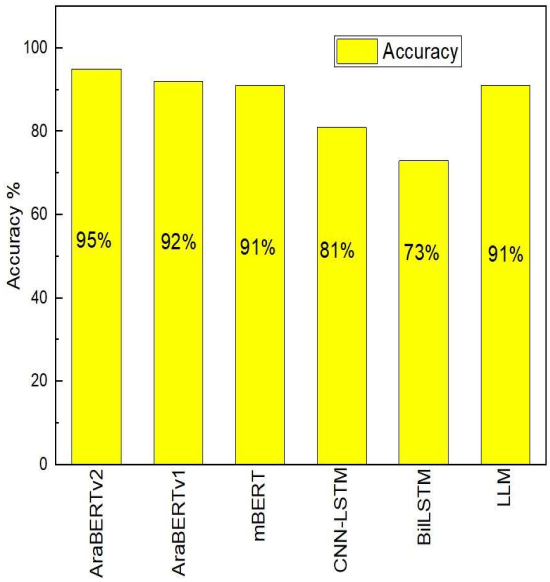


Figure 6. Performance of the pretrained LLM; AraBERT model versus CNN-LSTM, mBERT and BiLSTM models

V. CONCLUSION AND FUTURE WORK

The challenge of automatically identifying idiomatic expressions continues to be a complex issue that engages NLP researchers. This paper introduces the use of Large Language Models (LLMs) to enhance the prediction capabilities for Arabic idiomatic expressions. Initially, we gathered data and subjected it to rigorous cleaning and preprocessing, then formatted it to meet the requirements for fine-tuning with the AraBERT and ChatGPT-4 models. The performance of these models, evidenced by the results, demonstrates high precision and robustness, achieving F1-scores and accuracy rates of up to 95% in recognizing Arabic idiomatic expressions. This significantly surpasses the performance of mBERT and ChatGPT-4, which achieved 91%, and outperforms the CNN-LSTM and BiLSTM models, with scores of 81% and 73%, respectively.

Looking forward, we aim to expand the application of AraBERT to a broader range of NLP tasks, such as Arabic named entity recognition. We also plan to enhance our dataset validation by incorporating additional LLMs, allowing us to benchmark our findings against recent advancements in the field. This ongoing work will further refine our understanding and capabilities in handling complex linguistic phenomena unique to the Arabic language.

References

[1] I. A. Al-Sughaiyer, I. A. Al-Kharashi, "Arabic morphological analysis techniques: A comprehensive survey," *J. Am. Soc. Inf. Sci.*, vol. 55, issue 3, pp. 189-213, 2004. https://doi.org/10.1002/asi.10368.

[2] M. Altantawy, N. Habash, O. Rambow, S. Ibrahim, "Morphological analysis and generation of Arabic nouns: A morphemic functional approach," *Proceedings of the Language Resource and Evaluation Conference*, Malta, 2010. [Online]. Available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/442_Paper.pdf.

[3] A. N. De Roeck, W. Al-Fares, "A morphologically sensitive clustering algorithm for identifying Arabic roots," *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics - ACL'00*, 2000, pp. 199-206. https://doi.org/10.3115/1075218.1075244.

[4] S. A. Minko-Mi-Nseme, Modélisation des Expressions Figées en Arabe en Vue de la Constitution d'une Base de Données Lexicale, Ph.D. Thesis, Lyon 2, 2003. (in French).

[5] H. M. Alqahtni, *The Structure and Context of Idiomatic Expressions in the Saudi Press*, Ph.D. Thesis, University of Leeds, 2014.

[6] A. Kourtin, et al., "Lexicon-grammar tables for modern Arabic idiomatic expressions," *In NooJ Conference Proceedings*, 2021. https://doi.org/10.1007/978-3-030-92861-2_3.

[7] J. Baptista, "Compositional vs. idiomatic sequences," *J. Appl. Linguist. Spec. Issue Lexicon-Grammar*, pp. 81-92, 2004.

[8] M. S. Ali, "La traduction des expressions figées: Langue et culture," *Traduire*, vol. 235, pp. 103-123, 2016. (in French). https://doi.org/10.4000/traduire.865.

[9] M. Abdelmaksoud, *Les Expressions Idiomatiques dans le Coran et Leur Traduction Française: Étude Analytique Contrastive de l'Arabe vers le Français dans Trois Interprétations Françaises du Sens du Coran*, Ph.D. Thesis, Mansoura University, Egypt, 2018. (in French).

[10] M. Gross, "Les phrases figées en français," *L'information Grammaticale*, vol. 59, issue 1, pp. 36-41, 1993. https://doi.org/10.3406/igram.1993.3139.

[11] A. El-Mahdi, F. Bensalem & S. Khalil, "A framework for translating Arabic idiomatic expressions into English: Cultural and contextual insights," *Journal of King Saud University – Computer and Information Sciences*, vol. 29, issue 3, pp. 245–262, 2017.

[12] M. Alotaibi, "Pedagogical implications of teaching Arabic idiomatic expressions to non-native speakers," *Arab World English Journal*, vol. 8, issue 1, pp. 85-98, 2017.

[13] A. M. A. Nada, et al., "Arabic text summarization using AraBERT model using extractive text summarization approach," *International Journal of Academic Information Systems Research (IJAISR)*, vol. 4, issue 8, pp. 6-9, 2020.

[14] F. Alami, et al., "Multilingual offensive language detection method based on transfer learning from transformer fine-tuning model," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, issue 8, part. B, pp. 6048-6056. 2022. https://doi.org/10.1016/j.jksuci.2021.07.013.

[15] D. Faraj, M. Abdullah, "SarcasmDet at SemEval-2021 Task 7: Detect humor and offensive based on demographic factors using RoBERTa pretrained model," *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 527-533, https://doi.org/10.18653/v1/2021.semeval-1.64.

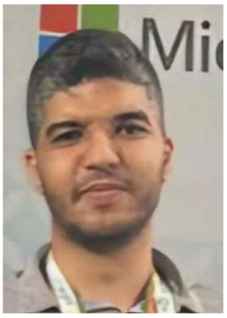
[16] M. Al-Shargi, O. Rambow, "Computational challenges of parsing Arabic idiomatic expressions," *Proceedings of the Computational Linguistics Conference*, 2015.

[17] F. Mohammed, A. Haji, "A diachronic study on the evolution of Arabic idiomatic expressions," *Journal of Historical Linguistics*, 2019.

[18] M. El-Haj, U. Kruschwitz, C. Fox, "Creating an Arabic diacritized corpus for statistical machine translation," *Proceedings of the Ninth*

International Conference on Language Resources and Evaluation (LREC'14), 2014.
[19] M. Alharbi, M. Aziz, "Statistical machine translation for Arabic idioms," *Procedia Computer Science*, vol. 117, pp. 112-119, 2017.
[20] T. Bani-Khaled, "The role of idiomatic expressions in Arabic language learning," *Arab World English Journal*, vol. 10, issue 2, pp. 35-48, 2019.
[21] D. Al-Ghamdi, H. Altalhi, "Integrating idiomatic expressions in interactive language learning tools," *International Journal of Computer-Assisted Language Learning and Teaching*, vol. 8, issue 3, pp. 58-72, 2018.
[22] M. Saad, W. Ashour, "Arabic idiomatic expressions in social media: A computational approach," *Journal of Social Media Studies*, vol. 5, issue 1, pp. 98-112, 2020.
[23] L. Al-Sulaiti, E. Atwell, "The design of a corpus of contemporary Arabic," *International Journal of Corpus Linguistics*, vol. 11, issue 2, pp. 135-171, 2006. https://doi.org/10.1075/ijcl.11.2.02als.
[24] M. M. Daoud, *Dictionary of Idiomatic Expressions in Contemporary Arabic*, Dar Gharib for Printing, Publishing, and Distribution, Cairo, Egypt, 2003.
[25] A. El Mahdaouy, E. Gaussier, S. O. El Alaoui, "Arabic text classification based on word and document embeddings," In: Hassanien, A., Shaalan, K., Gaber, T., Azar, A., Tolba, M. (eds) *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016. AISI 2016, Advances in Intelligent Systems and Computing*, 2017, vol. 533, pp. 32–41. Springer, Cham. https://doi.org/10.1007/978-3-319-48308-5_4.
[26] A. Kourtin, A. Amzali, M. Mourchid, A. Mouloudi, S. Mbarki, "Lexicon-grammar tables standardization and implementation," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 33, no. 2, pp. 1243-1251, 2024. https://doi.org/10.11591/ijeecs.v33.i2.pp1243-1251.
[27] F. Z. El-Alami, S. O. El Alaoui, "Word sense representation based-method for arabic text categorization," *Proceedings of the 9th IEEE International Symposium on Signal, Image, Video and Communications*, Rabat, Morocco, 2018, pp 141-146. https://doi.org/10.1109/ISIVC.2018.8709234.
[28] "BERT (modèle de langage)," Wikipédia, [Online]. Available at: https://fr.wikipedia.org/w/index.php?title=BERT_(mod%C3%A8le_de_langage)&oldid=176328149.
[29] Q. H. Sun, R. M. Horton, D. A. Bader, B. Jones, L. Zhou, T. T. Li, "Projections of temperature-related non-accidental mortality in Nanjing, China. *Biomed. Environ. Sci.*, vol. 32, issue 2, pp. 134-139, 2019. https://doi.org/10.3967/bes2019.019.
[30] W. Antoun, F. Baly, H. Hajj, *AraBERT: Transformer-Based Model for Arabic Language Understanding*, arXiv preprint arXiv:2003.00104, 2020.

SALMA TACE, *Bachelor in Mathematics and Computer Sciences, Mathematic and Informatique, Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco, 2014. Master in Engineering and technologies of education and information Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco, 2016. Doctorate Student, natural language processing, Faculty of sciences Ben M'sik - FSBM, Hassan II University, Casablanca, Morrocco. The Last Scientific Position: Assoc. Prof., FSBM, Hassan II University, since 2000. Research Interests: Data Sciences, NLP Education, Mathematics, Deep Learning.*



MOSSAB BATAL, *is a PhD student at Ben M'sik Faculty of Sciences in Casablanca, Morocco, focusing on the "smart campus" initiative. He is an instructor at EMSI - Honoris United Universities and a backend engineer. His research leverages technology to enhance the educational experience and create ai based sustainable campus environments. He holds a Master's degree in data science. His expertise lies in backend development and data science.*

SOUMAYA OUNACER, *Bachelor in Mathematics and Computer Sciences, Mathematic and Informatique, Faculty of Sciences Ben M'sik, Hassan II University, Casablanca, Morocco, 2008. Master in Computer Engineering, National School for Computer Science and Systems Analysis - ENSIAS, University Mohammed V, Rabat, Morocco, 2010. Doctorate in Big Data and Machine learning, Faculty of sciences Ben M'sik - FSBM, Hassan II University, Casablanca, Morrocco, 2020. The Last Scientific Position: Assoc. Prof., FSBM, Hassan II University, Casablanca, Morocco, Since 2021. Research Interests: Machine Learning, Data Sciences, Education, Mathematics.*

SANAA EL FILALI *is currently a full professor of computer science in the Department of Mathematics and Computer Science at Faculty of Science Ben M'Sik, Hassan II University of Casablanca. She received her Ph.D. in computer science from the Faculty of Science Ben M'sik in2006. Her research interests include computer training, the Internet of things, and information processing. She can be contacted at elfilalis@gmail.com*

MOHAMED AZZOUAZI, *Bachelor in Mathematics and Applications, Mathematic and Informatique, Faculty of Sciences Ben M'sik - FSBM, Hassan II University, Casablanca, Morocco, 1990. Master in Information Systems, National School for Computer Science and Systems Analysis - ENSIAS, Mohammed V University, Rabat, Morocco, 1994. Doctorate in Big Data Analytics, Faculty of Sciences Ben M'sik - FSBM, Hassan II University, Casablanca, Morocco, 2017. The Last Scientific Position: Assoc. Prof., FSBM, Hassan II University, Since 2000. Research Interests: Artificial Intelligence, Machine Learning, Deep Learning, Big Data Solutions.*