AraBERT transformer model for Arabic comments and reviews analysis

Hicham El Moubtahij[1], Hajar Abdelali[2], El Bachir Tazi[3]
[1]Systems and Technologies of Information Team, High School of Technology, University of Ibn Zohr, Agadir, Morocco
[2]LISAC Laboratory, Faculty of Sciences Dhar Mahraz, University of Sidi Mohamed Ben Abdellah, Fez, Morocco
[3]Computer Science department, Polydisciplinary Faculty, University of Sidi Mohamed Ben Abdellah, Taza, Morocco

Article Info

Article history:

Received Sep 27, 2021
Revised Dec 24, 2021
Accepted Jan 4, 2022

Keywords:

AraBERT
Arabic language understanding
Farasa segmentation
Natural language processing

ABSTRACT

Arabic language is rich and complex in terms of word morphology compared to other Latin languages. Recently, natural language processing (NLP) field emerges with many researches targeting Arabic language understanding (ALU). In this context, this work presents our developed approach based on the Arabic bidirectional encoder representations from transformers (AraBERT) model where the main required steps are presented in detail. We started by the input text pre-processing, which is, then, segmented using the Farasa segmentation technique. In the next step, the AraBERT model is implemented with the pertinent parameters. The performance of our approach has been evaluated using the ARev dataset which contains more than 40,000 comments-remarks records relate to the tourism sector such as hotel reviews, restaurant reviews and others. Moreover, the obtained results are deeply compared with other relevant states of the art methods, and it shows the competitiveness of our approach that gives important results that can serve as a guide for further improvements in this field.

Corresponding Author:

Hicham El Moubtahij
Systems and Technologies of Information Team, High School of Technology, University of Ibn Zohr
Agadir, Morocco
Email: h.elmoubtahij@uiz.ac.ma

1. INTRODUCTION

Arabic is an international language, spoken by more than 500 million speakers. It is considered as one of the important Semitic languages family. From the Arabian gulf to the atlantic ocean, Arabic language is administrative and official language of more the 21 countries [1]. Arabic is a rich and complex language in terms of word morphology compared to English, the presence of various dialects is some of the distinguishing prominent factors in the language. Moreover, the large differences between the modern standard Arabic (MSA) and the dialectical Arabic (DA) increase this complexity. It should be noted that MSA is employed for formal (administrative) writing and DA is employed for informal daily communication on social media for example [2]. From the work of Guellil *et al.* [3] published in 2021, the DA is divided into six collections: i) Maghrebi (MAGH), ii) Egyptian (EGY), iii) Iraqi (IRQ), iv) Levantine (LEV), v) Gulf (GLF), and vi) others remaining dialect. On the other hand, the Arabic language used on short messaging system (SMS), chat forums and on social media generally is called "Arabizi" [4]. Its written text is a mixture of Latin characters, numerals and some punctuation. For example, the sentence: "يا لاه نسافرو", that is translated into English as "let's travel", is written in Arabizi form as "yallah nsaaferou" [5].

Despite its spread usage, there is little research in the field of modern computational linguistics interested in the Arabic language compared to other language. However, in the last years, several research efforts has been made and many paper appear in various language processing tasks. Practically, the named entity recognition (NER) and the sentiment analysis (SA) are the most difficult tasks of Arabic natural language processing (ANLP) [6].

In order to obtain satisfactory results with tolerable performance for ANLP tasks, research works of the last years have focused on the application of transfer learning by the fine-tuning of large pre-trained language models with a relatively small number of samples. It should be mentioned that this approach is based on a self-supervised pre-trained language models. They allow us to represent the set of words as dense vectors in a vector space of minimum dimension and construct continuous distributed representations for texts. Despite the effectiveness of word embedding, it is unable to take into account the relationship between several words and the meaning of complete sentences in the text. Seeing the next two sentences, " نفسها المرأة هذه " .On the one hand, their word embedding representations are identical, and on the other hand, their meanings are entirely different. However, the high computational cost is a disadvantage in the training phase of the models (more than 500 TPU working for weeks). Moreover, a huge corpus is needed for the pre-training phase [7], [8].

In this work, we define and describe the important process and steps of our approach base on Arabic bidirectional encoder representations from transformers (AraBERT) transformer model for the Arabic language understanding (ALU). We can effectively classify the comments and the reviews into positive and negative categories. Hence, we evaluated our model on ARev dataset which contains more than 40,000 comments, hotel, restaurant, product, attraction and movie reviews written on a mixture of standard Arabic and Algerian dialect. The experiments show that our approach achieves very good results.

This reminder of this paper is structured as: in section 2, we present the most important techniques and approaches used in the natural language processing (NLP) field to deal with the ALU problem. Then, in section 3 we describe and clarify our model’s architecture where BERT represents its basic core. In section 4, we describe the ARev dataset on which we perform our experiments, then we compare our results with those of relevant methods. Finally, section 5 concludes the paper and outlines the main points of our future works.

2. RELATED WORKS

There are various techniques and approaches used in NLP to solve the problem of ALU. In this section, we briefly present some work in this field. The first work on the meaning of words began in 2013 with the word2vec model developed by Mikolov *et al.* [9], then researchers are oriented towards variants of word2vec like GloVe by Pennington *et al.* [10] in 2014 and fast-text by Mikolov *et al.* [11] in 2017. By the introduction of the concept of "contextual information" in 2018, the results were improved noticeably on different tasks [12], increasingly the structures became larger which had superior representations of words and sentences. From this date, the famous models of language comprehension have been developed, for example: i) bidirectional encoder representations from transformers (BERT) [13], ii) universal language model fine-tuning (ULMFiT) [14], iii) text-to-text transfer transformer (T5) [15], iv) A Lite BERT (ALBERT) [16]. These offered improved performance by exploring different pre-training methods, modified model architectures and larger learning corpora.

Concerning the AraBERT model, we note that there is little work done in relation to other languages. In the following we quote some in chronological order. In 2020, Nada *et al.* [17] proposed a new approach for Arabic text summarizer founded on a general-purpose architecture for natural language understanding (NLU), and natural language generation (NLG): generation and understanding of natural language to summarize the Arabic text by extracting and evaluating the most important sentences at this text.

Alami, a member of the LISAC FSDM-USMBA team at SemEval-2020 [18], proposed an effective method for dealing with the offensive Arabic language in Twitter by using AraBERT embeddings. In the First, they started with pre-processing tweets by handling emojis (containing their Arabic meanings), in the next, they substituted each detected emojis by the special token (MASK) into both fine-tuning and inference phases. Then, by applying the AraBERT model they represent tweets tokens. Finally, to decide whether a tweet is offensive or not, they feed the tweet representation into a sigmoid function. There proposed method achieved the best results, a score equal to 90.17% on OffensEval 2020.

In the next year, Faraj and Abdullah [19] published the best solution for the shared task on sentiment and sarcasm detection in the Arabic language. The objective global of the task is to identify whether a tweet is sarcastic or not. The proposed solution is based on the ensemble technique with AraBERT pre-trained model. In their paper, they started by defining the architecture of the model in the shared task. In the next, the hyperparameter and the experiment tuning that lead to this result are presented in detail. Their model is ranked 5th out of 27 teams with an F1 score of 0.5985.

In the recent work of 2021, Hussein *et al.* [20] worked on an effective approach for fighting Tweets COVID-19 Infodemic by using the AraBERT model. The organisation of their approach is: in the first step, the goal is to transform Twitter jargon, including emoticons and emojis, into plain text by involving a sequence of pre-processing procedures, and they exploited a version of AraBERT in the second step, which was pre-trained on plain text, to fine-tune and classify the tweets concerning their Label. Their approach can be predict 7 binary properties of an Arabic tweet about COVID-19. By using the dataset provided by NLP4IF 2021, they ranked 5th in the Fighting the COVID-19 Infodemic task results with an F1 of 0.664.

3. METHODOLOGY

The objective of this section is to describe and clarify the architecture of our model based on the AraBERT model, where BERT represents the basic core. Subsection 3.1 show BERT model. Our model based on AraBERT see in subsection 3.2.

3.1. BERT model

BERT stands for bidirectional encoder representations from transformers, it came out of Google AI labs in late 2018. We mention that it is: i) more powerful than its predecessors in terms of results; ii) more powerful than its predecessors in terms of learning speed; iii) once pre-trained, in an unsupervised way, it has its own linguistic “representation”. It can be trained in incremental mode (in a supervised way this time) to specialize the model quickly and with little data; and iv) finally, it can work in a multi-model way, taking as input data of different types such as images and/or text, with some manipulations. It has the advantage over its competitors OpenAI's generative pre-trained transformer (GPT) and embeddings from language models (ELMo) [12] of being bi-directional, it does not have to look only backwards like OpenAI GPT or concatenate the “back” view and the “front” view driven independently like for ELMo, as shown in Figure 1.

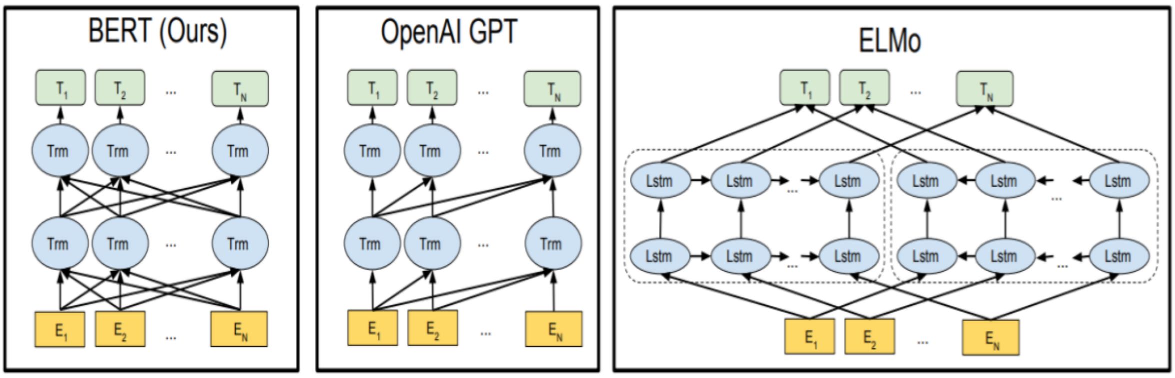

Figure 1. Differences in pre-training model architectures

Examples of what it can do: i) BERT can do the translation. He can even once pre-trained to translate [French/English-English/French] and then [English/German-German/English], translate from French to German without training; ii) BERT can compare the meaning of two sentences to see if they are equivalent; iii) BERT can generate text; iv) BERT can describe and categorize an image; and v) BERT can do logical sentence analysis, i.e. determine if a given element is a subject, a verb, and a direct object complement.

3.1.1. Bidirectional encoder representations from transformers (BERT) architecture

BERT reuses the architecture of transformers (hence the “T” in BERT). Indeed, BERT is nothing more than a superposition of encoders that all have the same structure but do not share the same weights. The “Base” version of BERT consists of 12 encoders. There is another larger version called “Large” which has 24 encoders. Certainly, the large version is more powerful but more demanding on machine resources. The above model has 512 entries, each corresponding to a token. The first entry corresponds to a special token the “[CLS]” for “classification” which allows BERT to be used for a text classification task. It also has 512 outputs of size 768 each (1024 for the base version). The first vector is the classification vector. The output of each of the 12 encoders can be considered as a vector representation of the input sequence. The relevance of this representation is ensured by the attention mechanism implemented by the encoders.

3.1.2. Training procedure

BERT differs from its predecessors (pre-trained NLP models) in the way it is pre-trained on a large dataset consisting of texts from English Wikipedia pages (2,500 million words) as well as a set of books (800 million words). This pre-training is done on two tasks. Fisrt, a masked language modelling (MLM) task. Second, a next sentence prediction (NSP) task.

a. Task 1: masked language modelling (MLM)

The objective of this task is to predict the hidden word. Therefore, because of the ability of the transformer architecture to simultaneously take into account the right and left contexts of the target word, this task allows the model to learn even more contextualised representations than one-way models such as ELMo [12]. In practice, target words are sometimes replaced with a special symbol [MASK], or replaced with another random word, or kept as they are as shown in Figure 2.

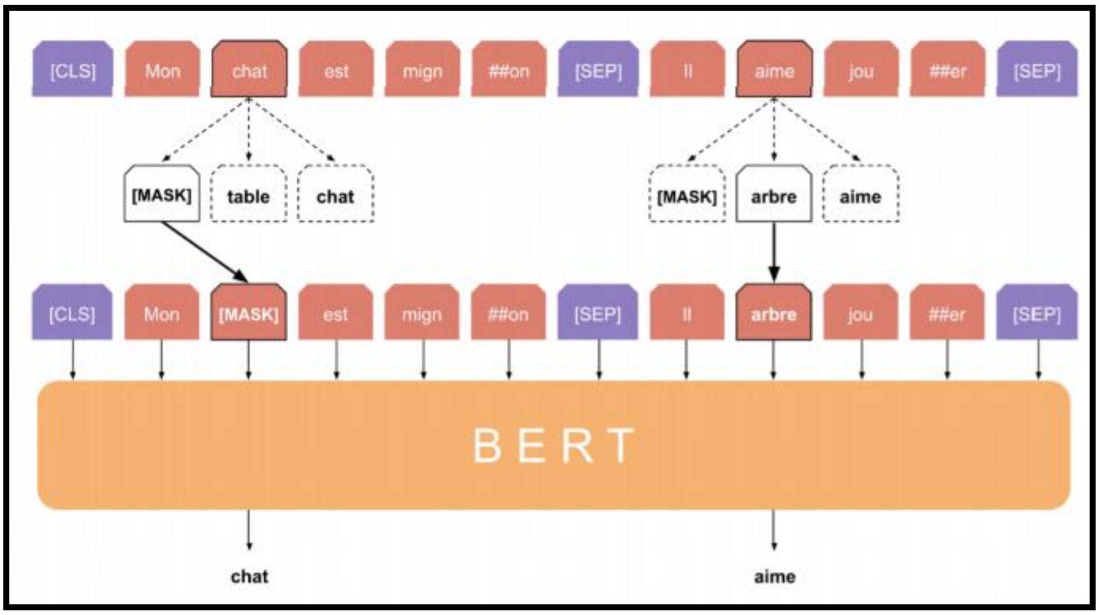

Figure 2. MLM

b. Task 2: next sentence prediction (NSP)

BERT is also trained on a next-sentence prediction task in which it must decide whether two input sentences are consecutive. The rationale for this task is to improve the performance of the model on tasks where the objective is to qualify the relationship between a pair of sentences. In practice, the special symbol representation [CLS] is used to classify each pair of input sentences as well as for any other classification task once the model has been trained.

3.1.3. BERT: fine-tuning

Fine-tuning consists of using a pre-trained version of BERT in a model architecture for a specific NLP task. Adding a basic neural network layer is enough to get very good results. For a text classification task, for example, and more precisely for the analysis of the sentiment of moviegoers' reviews, the architecture of the fitted model may look like this as shown in Figure 3. It is sufficient to add, downstream of BERT, a feed-forward followed by a softmax.

3.2. Our model based on AraBERT

In our approach, we used AraBERT based on the BERT model. It is a widely used model in various NLP tasks for several languages. AraBERT is a pre-trained model for the Arabic language, based on the Google BERT architecture [6] there are six versions of the model: AraBERTv0.1-base, AraBERTv0.2-base, AraBERTv0.2-large, AraBERTv1-base, AraBERTv2-base and AraBERTv2-large. In Table 1 we describe in detail the important information for each version in relation to the pre-training process. The overall view of our model is shown in Figure 4. We have been working on the customer/user review database for the sentiment analysis area, our dataset is titled ARev.

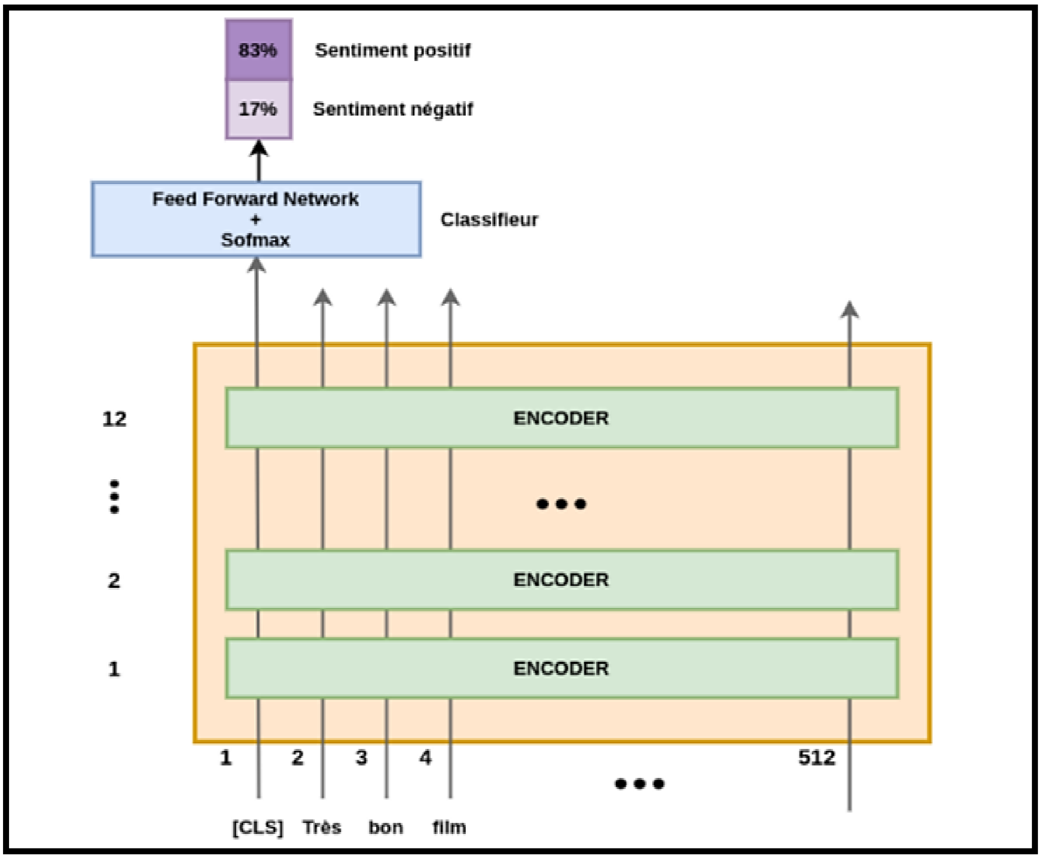

Figure 3. Architecture of the fine-tuning

Table 1. Model pre-training parameters

Model	Size		Pre-	Dataset		
	MB	Param.	segmentation	Sentences	Size	Words
AraBERTv0.2-base	543 M	136M	No	200 M	77 GB	8.6 B
AraBERTv0.2-large	1.38 G	371M	No	200 M	77 GB	8.6 B
AraBERTv2-base	543 MB	136M	Yes	200 M	77 GB	8.6 B
AraBERTv2-large	1.38 G	371M	Yes	200 M	77 GB	8.6 B
AraBERTv0.1-base	543 MB	136M	No	77 M	23 GB	2.7 B
AraBERTv1-base	543 MB	136M	Yes	77 M	23 GB	2.7 B

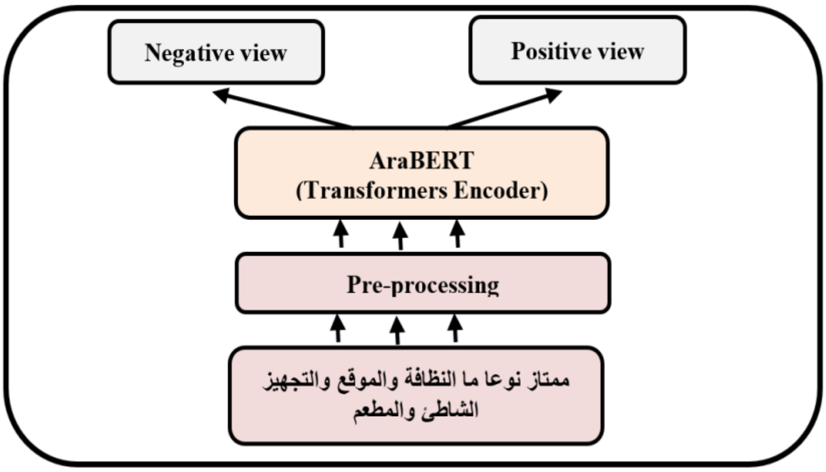

Figure 4. AraBERT architecture overview

At the input of our system, we go through the pre-processing stage where we clean the text of any unsentimental content, such as usernames, hashtags and URLs, and then proceed to segment the text by using the Farasa segmentation [21]. First, we segment the words into stems, prefixes and suffixes. Look for sentence, “ الكتاب – Alkittab “ becomes “ ال + كتا + ب" - " Al+kitta+b". Then, in unigram mode, we trained a Sentence Piece [22] on the segmented pre-training dataset to produce a subword vocabulary of more than 59K tokens. It must be noted that before the application of Farasa segmentation, the dataset that is used for pre-training has a size more of 70 GB, more than 8.5 billion words and more than 200 million sentences. To create a well pre-training dataset, we used several websites such as: i) OSIAN Corpus. ii) Arabic Wikipedia dump, iii) Assafir news articles, iv) 1.5 billion word Arabic Corpus, and v) OSCAR unfiltered and sorted
In our model based on AraBERT, we successively used two special tokens: Tok1: segment separation (“SEP’) and Tok2: classification (“CLS”). For any classifier, we used it as the first input token which we help us to derive an output vector. Then, in order to obtain the probability distribution on the predicted output classes, we add a simple layer composed of feed-forward and Softmax see (1):

$$P = \mathrm{softmaxCWT)} \tag{1}$$

where P is probability of each category, W is matrix of the classification layer, and C is output of the transformers.

4. EXPERIMENT AND RESULTS

4.1. ARev dataset

We evaluated our model on the sentiment analysis task. For this reason, we used the Arabic reviews (Arev) dataset [23]. Using the Facebook API, the ARev dataset is built by more than 100 K comments of the most popular Algerian Facebook pages. We needed tree input for our ARev dataset which are: the Facebook page identifier, the identifier of the Facebook page post and the access token as shown in Figure 5. To enrich our ARev dataset, three open-source datasets of modern standard Arabic and Algerian Arabic comments are used see Table 2. Finally, after pre-processing and deleting the duplicate elements, the dataset is saved in CSV format. The statistics of our dataset are presented in Table 3.

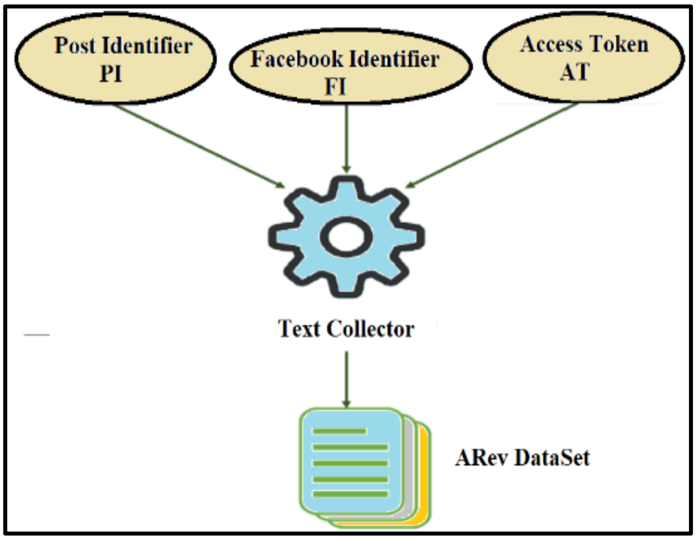

Figure 5. Inputs of dataset collection from Facebook

Table 2. Various datasets used

Datasets	Type of language	Description
LABR [24]	Standard Arabic	Book reviews
The dataset of Elsahar and El-Beltagy [25]		Hotel reviews, restaurant reviews, product reviews, attraction reviews, movie reviews.
The dataset of Mataoui *et al*. [26]	Algerian Dialect	Comments

Table 3. Statistics on the ARev dataset

	Positive	Negative
Total comments	24932	24932
Total words	1180663	1345029
Avg. words in each comment	47.36	53.95
Avg. characters in each comment	253.15	294.47

4.2. Experimental setup

We used the Google Colab tool to run our experiments where we can take good advantage of TensorFlow's performance. Note that we worked with a masking probability of 15%, a random seed of 34, and a duplication factor was set to 10. In our approach, we worked through the version of AraBERTv1 implemented in the work of [6] where our model was pre-trained on a TPUv2-8 pod. Table 4 resume the parameters used for fine-tuning in our models.

Table 4. Parameter values

Parameter	Value
Learning Rate	1e-4
Epsilon (Adam optimizer)	1e-8
Maximum Sequence Length	256
Epochs	27

4.3. Results and discussion

To show the importance of our module, we compared the result obtained by our approach with those existing in the state of the art for the domain of sentiment analysis. For this reason, we used the accuracy metric, as shown in Table 5. The previous results show that our approach gives an important result that is comparative to those of the state of the art. We obtained an accuracy value of 92.5% for a database containing more than 40,000 comments written by a mixture of standard Arabic and Algerian dialect. However, the approach of Alomari *et al.* [27] gives an accuracy value better than ours by +1.3%, which is a slight difference due to the two reasons following: Firstly, the number of tweets in [27] does not exceed 1800 tweets, secondly, the language mix used in our approach generates more linguistic specifications than the Jordanian dialect. The AraBERT v1 with the best parameters chosen for fine-tuning gives our approach this competitiveness over other models.

Table 5. Performance of our model implemented on AraBERTv1 compared by the previous state of the art systems

Dataset	Descriptions	Language	Accuracy
ASTD [28]	The dataset contains 10,000 tweets.	Egyptian dialect	92.6
Arsen TD lev [29]	The dataset contains 4,000 tweets.	Levantine dialect	59.4
AJGT [27]	The Arabic Jordanian General Tweets dataset contains more than 1,800 tweets.	Jordanian dialect	93.8
ArSarcasm-v2 [30]	Collection of 15,548 sarcasm and sentiment tweets.	Standard Arabic and dialectal Arabic	67.7
ARev Our dataset	The Dataset of a mixture of comments and Hotel reviews, restaurant reviews, product reviews, attraction reviews, movie reviews.	Standard Arabic and Algerian dialect	**92.5**

5. CONCLUSION AND FUTURE WORK

The automatic understanding of Arabic scripts is still a challenging process and an open issue for researchers in the NLP field. In this work, we have presented our approach based on the AraBERT language model. Also, we have described and detailed the main steps of the proposed architecture using diagrams and examples. The process starts with the input of our model into a pre-processed text from the ARev database, then version 1 of the AraBERT model was implemented by using Farasa segmentation. Moreover, our evaluation is based on the ARev dataset, which contains more than 40,000 comments and reviews. With well-tuned parameters of the AraBERT model, we obtained an accuracy value of 92.5%, which represents a very competitive result. In future work, we aim to address the problem of Arabic text segmentation, try to improve the farasa segmentation version.

REFERENCES

[1] N. Boudad, R. Faizi, R. Oulad Haj Thami, and R. Chiheb, “Sentiment analysis in Arabic: A review of the literature,” *Ain Shams Engineering Journal*, vol. 9, no. 4, pp. 2479–2490, Dec. 2018, doi: 10.1016/j.asej.2017.04.007.

[2] A. Wadhawan, “Dialect identification in nuanced arabic tweets using farasa segmentation and AraBERT,” *arXiv:2102.09749*, Feb. 2021.

[3] I. Guellil, H. Saâdane, F. Azouaou, B. Gueni, and D. Nouvel, “Arabic natural language processing: An overview,” *Journal of King Saud University-Computer and Information Sciences*, vol. 33, no. 5, pp. 497–507, Jun. 2021, doi: 10.1016/j.jksuci.2019.02.006.

[4] T. Tobaili, “Sentiment analysis for the low-resourced latinised Arabic ‘Arabizi’’,’” The Open University, 2020.

[5] I. Guellil, F. Azouaou, F. Benali, A. E. Hachani, and M. Mendoza, “The role of transliteration in the process of Arabizi translation/sentiment analysis,” in *Studies in Computational Intelligence*, Springer International Publishing, 2020, pp. 101–128.

[6] W. Antoun, F. Baly, and H. Hajj, “AraBERT: transformer-based model for Arabic language understanding,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 8440–8451.

[7] A. Conneau *et al.*, “Unsupervised cross-lingual representation learning at scale,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Nov. 2020, pp. 8440–8451.

[8] D. Adiwardana *et al.*, “Towards a human-like open-domain chatbot,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Jan. 2018, pp. 52–55.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in *Advances in neural information processing systems*, 2013, pp. 3111–3119.

[10] J. Pennington, R. Socher, and C. Manning, “Glove: global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543, doi: 10.3115/v1/D14-1162.

[11] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, “Advances in pre-training distributed word representations,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1712.09405.

[12] M. Peters *et al.*, “Deep contextualized word representations,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 2227–2237, doi: 10.18653/v1/N18-1202.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: pre-training of deep bidirectional transformers for language understanding,” *arXiv:1810.04805*, Oct. 2018.

[14] J. Howard and S. Ruder, “Universal language model fine-tuning for text classification,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Jan. 2018, pp. 328–339, doi: 10.18653/v1/P18-1031.

[15] C. Raffel *et al.*, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *arXiv:1910.10683*, Oct. 2019.

[16] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: a lite BERT for self-supervised learning of language representations,” *arXiv:1909.11942*, Sep. 2019.

[17] A. M. A. Nada, E. Alajrami, A. A. Al-Saqqa, and S. S. Abu-Naser, “Arabic text summarization using AraBERT model using extractive text summarization approach,” *International Journal of Academic Information Systems Research (IJAISR)*, vol. 4, no. 8, pp. 6–9, 2020.

[18] H. Alami, S. Ouatik El Alaoui, A. Benlahbib, and N. En-nahnahi, “LISAC FSDM-USMBA Team at SemEval-2020 Task 12: Overcoming AraBERT’s pretrain-finetune discrepancy for Arabic offensive language identification,” in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, 2020, pp. 2080–2085, doi: 10.18653/v1/2020.semeval-1.275.

[19] D. Faraj and M. Abdullah, “SarcasmDet at SemEval-2021 Task 7: detect humor and offensive based on demographic factors using RoBERTa pre-trained model,” in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, 2021, pp. 527–533, doi: 10.18653/v1/2021.semeval-1.64.

[20] A. Hussein, N. Ghneim, and A. Joukhadar, “DamascusTeam at NLP4IF2021: fighting the Arabic COVID-19 infodemic on Twitter using AraBERT,” in *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 2021, pp. 93–98, doi: 10.18653/v1/2021.nlp4if-1.13.

[21] A. Abdelali, K. Darwish, N. Durrani, and H. Mubarak, “Farasa: a fast and furious segmenter for Arabic,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016, pp. 11–16, doi: 10.18653/v1/N16-3003.

[22] T. Kudo, “Subword regularization: improving neural network translation models with multiple subword candidates,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 66–75, doi: 10.18653/v1/P18-1007.

[23] A. Abdelli, F. Guerrouf, O. Tibermacine, and B. Abdelli, “Sentiment analysis of Arabic Algerian dialect using a supervised method,” in *2019 International Conference on Intelligent Systems and Advanced Computing Sciences (ISACS)*, Dec. 2019, pp. 1–6, doi: 10.1109/ISACS48493.2019.9068897.

[24] M. Aly and A. Atiya, “Labr: A large scale Arabic book reviews dataset,” 2013.

[25] H. ElSahar and S. R. El-Beltagy, “Building large Arabic multi-domain resources for sentiment analysis,” in *Computational Linguistics and Intelligent Text Processing*, Springer International Publishing, 2015, pp. 23–34.

[26] M. Mataoui, O. Zelmati, and M. Boumechache, “A proposed lexicon-based sentiment analysis approach for the vernacular Algerian Arabic,” *Research in Computing Science*, vol. 110, no. 1, pp. 55–70, Dec. 2016, doi: 10.13053/rcs-110-1-5.

[27] K. M. Alomari, H. M. ElSherif, and K. Shaalan, “Arabic tweets sentimental analysis using machine learning,” in *Advances in Artificial Intelligence: From Theory to Practice*, Springer International Publishing, 2017, pp. 602–610.

[28] 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.

[29] R. Baly, A. Khaddaj, H. Hajj, W. El-Hajj, and K. B. Shaban, “ArSentD-LEV: a multi-topic corpus for target-based sentiment analysis in Arabic levantine tweets,” *The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools*, 2018.

[30] I. A. Farha and W. Magdy, “Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection,” in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 21–31.

BIOGRAPHIES OF AUTHORS

Prof. Hicham El Moubtahij is currently a Professor of Computer Science at the University of Ibn Zohr, Agadir, Morocco. He received his Ph.D. in Computer Science from the University of Sidi Mohamed Ben Abdellah, Fez, Morocco, in 2017. He is now a member of the Systems and Technologies of Information Team at the High School of Technology at the University of Ibn Zohr, Agadir. His current research interests include machine learning, deep learning, Arabic handwriting recognition, Text Mining, and medical imagery. Dr. El Moubtahij Hicham has published articles in indexed international journals and conferences, has been a reviewer for scientific journals, and has served on the program committee of several conferences. He can be contacted at email: h.elmoubtahij@uiz.ac.ma.

Dr. Hajar Abdelali holder of a bachelor’s degree in experimental sciences, a bachelor's degree in mathematics and computer science, a master's degree in information sciences, networks and multimedia from Sidi Mohammed Ben Abdellah, University of Fez, Morocco in 2013. She joined the laboratory XLIM of the University of Poitiers in France in collaboration with the scientific laboratory LIMS of the Faculty of Sciences Dhar Mahraz of Sidi Mohammed Ben Abdellah, University of Fez, Morocco where he obtained his Ph.D. degree in computer science in 2019. She can be contacted at email: abdelali.hajar@usmba.ac.ma.

Prof. El Bachir Tazi graduated in Electronic Engineering from ENSET Mohammedia Morocco in 1992. He obtained his DEA and DES in Automation and Signal Processing and his PhD in Computer Science from Sidi Mohammed Ben Abdellah University, Faculty of Sciences in Fez, Morocco respectively in 1995, 1999 and 2012. He is now a member of the engineering sciences laboratory and associate professor at Sidi Mohammed Ben Abdellah University, Polydisciplinary Faculty of Taza, Morocco. His areas of interest generally include all areas of automatic recognition based on artificial intelligence methods and applications related to automatic speaker. He can be contacted at email: elbachirtazi@yahoo.fr.