The Journal of Engineering and Exact Sciences – jCEC, Vol. 10 N. 07 (2024)
journal homepage: https://periodicos.ufv.br/jcec
eISSN: 2527-1075
ISSN: 2446-9416

Sentiment analysis of Algerian Arabic dialect on social media Using Bi-LSTM recurrent neural networks

Article Info:
Article history: Received 2024-09-09/ Accepted 2024-10-27 / Available online 2024-10-27
doi: 10.18540/jcecvl10iss7pp20058

Abdelghani BOUZIANE
ORCID: https://orcid.org/0000-0002-8583-3858
Department of Computer Science, Institute of Sciences, University centre of Naama, Algeria
E-mail: bouziane@cuniv-naama.dz
Benamar BOUOUGADA
ORCID: https://orcid.org/0009-0006-4020-4093
Department of Computer Science, Institute of Sciences, University centre of Naama, Algeria
E-mail: bouougada@cuniv-naama.dz
Djelloul BOUCHIHA
ORCID: https://orcid.org/0000-0001-7314-4329
Department of Computer Science, Institute of Sciences, University centre of Naama, Algeria
E-mail: bouchiha@cuniv-naama.dz
Noureddine DOUMI
ORCID: https://orcid.org/0000-0002-3108-4732
Department of Computer Science, University of Saida, Algeria
E-mail: noureddine.doumi@univ-saida.dz

Abstract

This paper presents a sentiment analysis approach using Bidirectional Long Short-Term Memory (Bi-LSTM) Recurrent Neural Networks to train predictive models for sentiment analysis on social media, particularly focusing on Algerian Arabic Dialect. The method leverages word-to-vector embedding for word representation and incorporates natural language understanding of emojis to improve semantic interpretation. The model achieves a high accuracy of 94%, demonstrating its effectiveness in analyzing sentiments in online discussions. The originality lies in applying Bi-LSTM to handle multilingual challenges on social platforms. The findings have practical implications for business, policymaking, and public sentiment evaluation, while also contributing positively to fostering informed online discourse.

Keywords: Sentiment analysis, Artificial intelligence, Social Web evolution, Deep learning solutions, Bi-LSTM.

1. Introduction

The Arabic language, with its rich historical and cultural significance, holds a prominent position in the global linguistic landscape. Spoken by over 450 million people across more than 22 countries, primarily in the Middle East and North Africa, Arabic plays a vital role in communication, literature, science, and philosophy. Among the varieties of Arabic, the Algerian dialect, influenced by French and Berber, is widely spoken in informal settings and on social media platforms such as Facebook, Twitter, and Instagram.

Algeria, the largest country in Africa, had a population of 45.26 million in January 2023, of which 32.09 million were internet users, representing 70.9% of the total population (Kemp, 2023). Social networks like Facebook, Twitter, Instagram, and YouTube are increasingly popular, with users frequently posting updates, comments, and messages in the Algerian dialect (See Figure 1).

This highlights the importance of local languages in online communication, with users leveraging the dialect to express themselves authentically and connect with others.

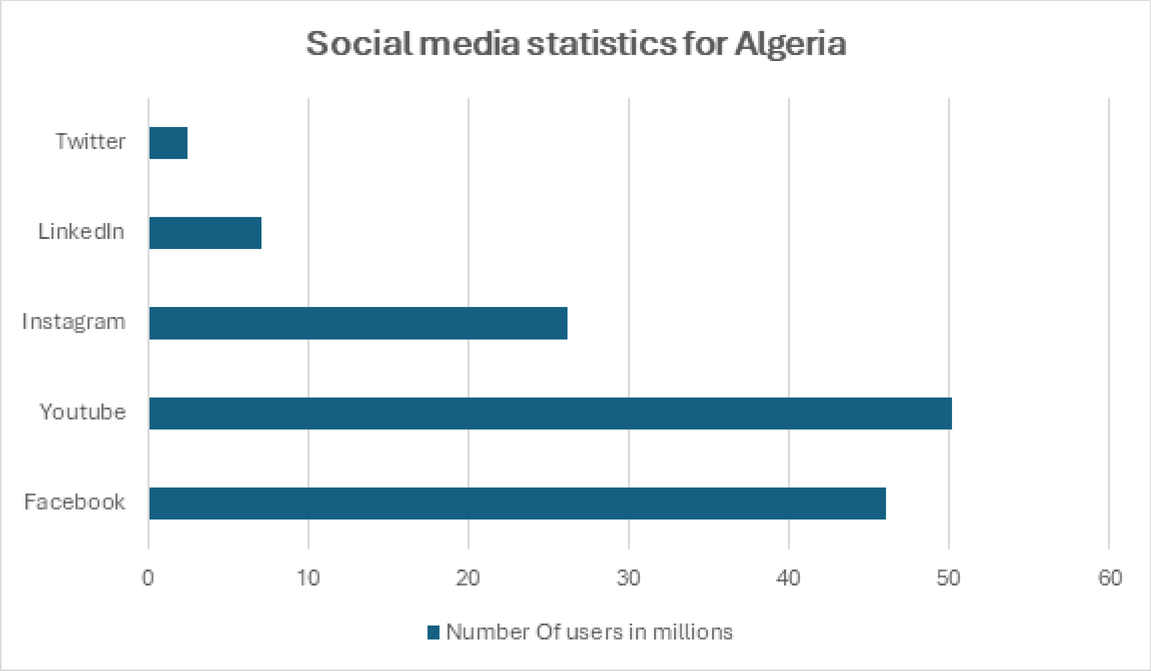


Figure 1 – Number of Algerian users on social networks in January 2023

Sentiment analysis (SA) has emerged as a crucial task in understanding and analyzing the opinions, attitudes, and perceptions shared on social media. SA is widely used by companies, organizations, and governments to gauge public sentiment on specific products, services, or topics (Medhat et al., 2014). Various approaches to sentiment analysis exist, ranging from rule-based methods to machine learning models such as Support Vector Machines (SVMs), Naive Bayes, and Recurrent Neural Networks (RNNs) (Birjali, Kasri, & Beni-Hssane, 2021). In Natural Language Processing (NLP), sentiment analysis using Bidirectional Long Short-Term Memory (Bi-LSTM), a variant of RNN, has proven effective in capturing the context of a given text, which is essential for accurately determining sentiment (Xu, Meng, Qiu, Yu, & Wu, 2019).

The objective of this research is to develop a Bi-LSTM-based model for sentiment analysis, focusing on the Algerian Arabic dialect within a COVID-19 dataset. The following sections of this paper include a survey of relevant literature, a description of the proposed methodology, an evaluation of the model's performance, and a conclusion with perspectives for future work.

2. Literature review

Among the notable contributions to sentiment analysis (SA) utilizing Bidirectional Long Short-Term Memory (Bi-LSTM) networks is the study by Omara et al. (2022), which introduced a novel approach that strategically leverages character-level information in textual content for enhanced sentiment analysis (Omara, Mousa, & Ismail, 2022). By combining word-level and character-level features, they developed a neural network model capable of accurately predicting the sentiment of Arabic text. The results indicated that their model outperformed other prevalent models in terms of precision and resilience, highlighting the effectiveness of the character-gated recurrent neural network approach.

Similarly, Wankhade and Rao (2022) employed a combination of Bi-LSTM and Bidirectional Encoder Representations from Transformers (BERT) to analyze large volumes of text and identify key aspects of public opinion (Wankhade & Rao, 2022). Their findings revealed that the BERT-Bi-LSTM ensemble method exceeded other techniques in both accuracy and speed, establishing it as a valuable tool for understanding public sentiment on significant issues such as COVID-19.

In a broader context, Al-Ayyoub et al. (2019) provided an exhaustive survey of SA methodologies employed in Arabic. They examined existing methods, categorizing them into three main groups: lexicon-driven approaches, rule-based methods, and machine learning (ML) techniques (Al-Ayyoub, Khamaiseh, Jararweh, & Al-Kabi, 2019). The authors highlighted the challenges and limitations associated with each approach and evaluated their performance. Additionally, they offered a comprehensive overview of available resources and tools for Arabic sentiment analysis, including datasets, corpora, and software tools, providing valuable insights for researchers and practitioners in the field.

In a different vein, Alayba et al. (2018) focused on developing and implementing a machine learning model for SA in Arabic (Alayba, Palade, England, & Iqbal, 2018). By combining two widely adopted deep learning methodologies, Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), they constructed a model that was tested on a dataset of Arabic tweets. Their comparative analysis demonstrated that the combined model outperformed traditional models in both accuracy and efficiency.

Furthermore, Shanmugavadivel et al. (2022) investigated the application of deep learning algorithms for SA and the identification of offensive language in multilingual code-mixed data (Shanmugavadivel et al., 2022). Utilizing pre-trained models such as BERT, RoBERTa, and adapter-BERT, they analyzed sentiment and offensive content in social media posts and online forums written in multiple languages. The outcomes of their experiments underscored the efficacy of deep learning algorithms in detecting sentiments and identifying offensive content within code-mixed data, concluding that these algorithms can provide insights into online users' opinions, emotions, and social attitudes.

Finally, Du et al. (2024) provide a comprehensive review of sentiment analysis in the financial sector, focusing on two key research areas: improving Financial Sentiment Analysis (FSA) techniques and applying FSA to financial markets (Du, Xing, Mao, & Cambria, 2024). The authors explore various methods, datasets, and applications used in FSA, and introduce frameworks for understanding the relationships between financial sentiment, investor sentiment, and market sentiment. They also highlight challenges and suggest future directions for enhancing FSA in financial forecasting and decision-making.

3. Methodology and proposed approach

In this section, we first present a dataset analysis to adjust the dataset for compatibility with our model. Next, we introduce our Bi-LSTM-based model for sentiment analysis in the context of Algerian Arabic Dialect.

3.1. Dataset analysis

To train our model, we utilized the "AraCOVID19-SSD" dataset, a manually annotated Arabic dataset consisting of 5162 tweets related to the COVID-19 pandemic (Hadj Ameur & Hassina, 2023). Social media tweets are predominantly written in Arabic dialects. Tweets were labeled according to two key tasks: sarcasm detection (labeling tweets as sarcastic or not) and sentiment analysis (classifying tweets as positive, negative, or neutral). The dataset, which was meticulously curated and analyzed, comprises tweets collected from December 2019 to December 2020.

For dataset analysis, we employed Stanza, a Python package that provides highly accurate neural network architecture for natural language processing (Qi, Zhang, Zhang, Bolton, & Manning, 2020). Stanza supports multiple languages, including Modern Standard Arabic (MSA), and offers functionalities such as sentence splitting, tokenization, lemmatization, part-of-speech (POS) tagging, morphological tagging, syntactic and dependency parsing, and named entity recognition.

While Stanza primarily supports MSA, our focus was on Algerian Arabic Dialect. At this stage, a key question arose: *"How similar is Algerian Arabic to MSA?"*. To address this, we conducted a linguistic analysis of the dataset using Stanza. A tweet, in this case, is treated as raw text segmented into sentences and tokens, with each Arabic word labeled using Universal POS (UPOS) tags. Non-Arabic words are not recognized by the POS and morphological tagging module. Additionally, tweets often contain URLs, emojis, hashtags, and user tags.

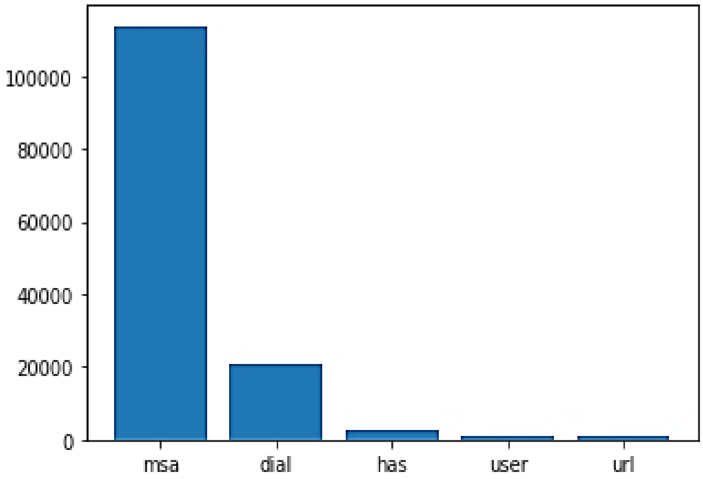


Figure 2 – Frequency of MSA words in the dataset

As illustrated in Figure 2, the majority of words in the dataset are in Modern Standard Arabic (MSA), reflecting the linguistic similarity between Algerian Arabic dialect and MSA. The POS tagging module effectively recognizes and processes these MSA words.

Figure 3 illustrates that the most frequent words in the dataset are nouns, followed by adpositions, verbs, pronouns, punctuation, conjunctions, and adjectives. The lower-frequency words include particles, determiners, numerals, subordinating conjunctions, auxiliaries, interjections, adverbs, and proper nouns.

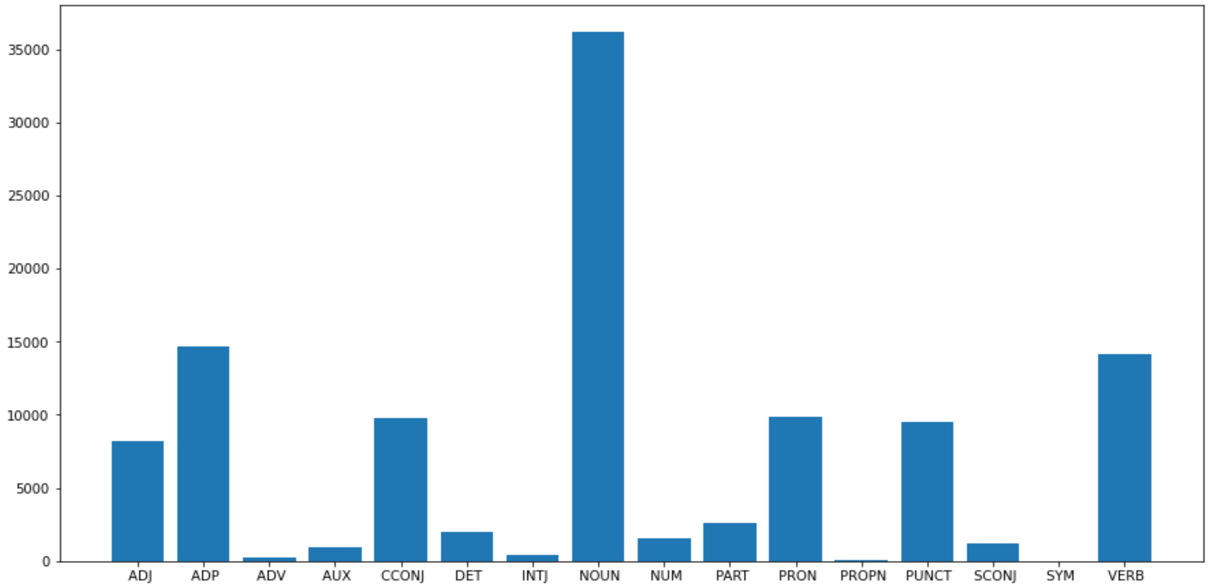


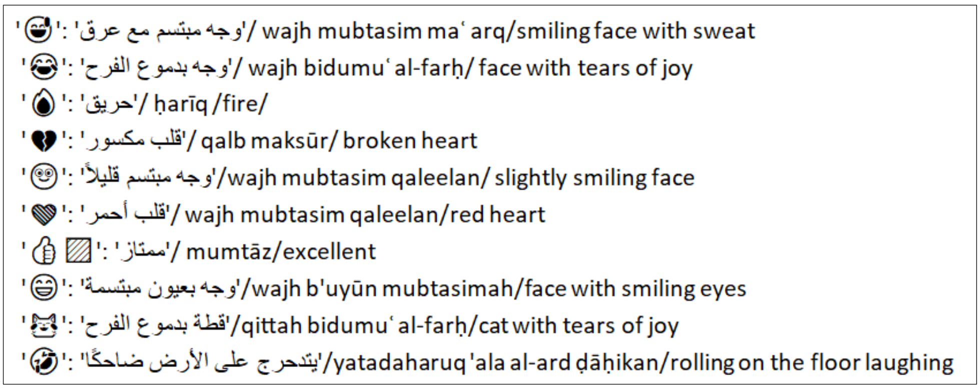
Figure 3 – POS tag statistics in the dataset

URLs appear in 24% of the tweets in the dataset. When users post URLs on Twitter, they are automatically shortened to an http://t.co link, obscuring the original text format of the URL. Consequently, we replace all URLs with the phrase “موقع واب/ mawqīʿ wāb” (website in Arabic) during processing. User tags are present in 26% of the tweets, and we replace these with “مستخدم/ mustakhdim” (user in Arabic).

Hashtags are natural language text, so we transform them into their corresponding text format. Over 54% of tweets contain hashtags, with 1356 total hashtags categorized into 489 positive, 299 negative, and 801 neutral tweets.

Emojis, which convey emotions or ideas within the text, appear in 43% of the tweets. We convert emojis into MSA text using the emoji library, which provides emoji codes defined by the Unicode Consortium along with natural language aliases in English.

Figure 4 shows the Arabic translations of the emoji text.

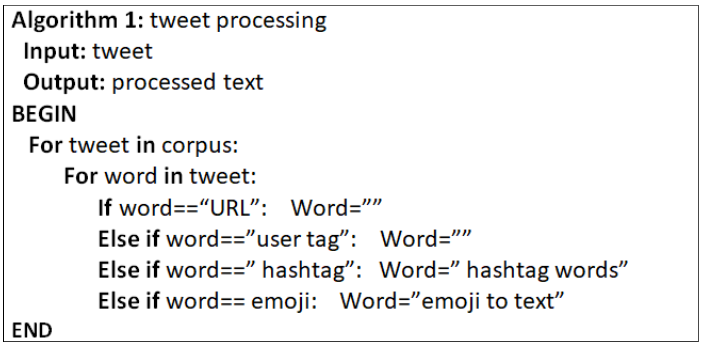


'😅': 'وجه مبتسم مع عرق'/ wajh mubtasim maʿ arq/smiling face with sweat
'😂': 'وجه بدموع الفرح'/ wajh bidumuʿ al-farḥ/ face with tears of joy
'🔥': 'حريق'/ ḥarīq /fire/
'💔': 'قلب مكسور'/ qalb maksūr/ broken heart
'🙂': 'وجه مبتسم قليلاً'/wajh mubtasim qaleelan/ slightly smiling face
'❤': 'قلب أحمر'/ wajh mubtasim qaleelan/red heart
'👍 ▨': 'ممتاز'/ mumtāz/excellent
'😄': 'وجه بعيون مبتسمة'/wajh b'uyūn mubtasimah/face with smiling eyes
'😹': 'قطة بدموع الفرح'/qittah biduмuʿ al-farḥ/cat with tears of joy
'🤣': 'يتدحرج على الأرض ضاحكًا'/yatadaharuq 'ala al-ard ḍāḥikan/rolling on the floor laughing

Figure 4 – Arabic, Buckwalter transliteration, and English translation of emojis

In summary, the input text (tweet) undergoes the following preprocessing steps:

- Tokenizing the words of the text.
- POS tagging the tweet using Stanza.
- Removing stop words from the tweets.
- Processing the URL, user tags, hashtags, and emojis (algorithm of Figure 5):

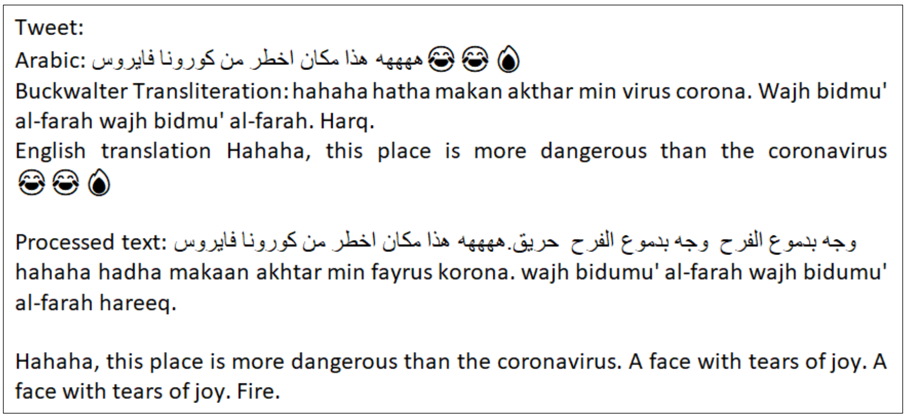


```
Algorithm 1: tweet processing
 Input: tweet
 Output: processed text
BEGIN
  For tweet in corpus:
      For word in tweet:
          If word==“URL”:   Word=””
          Else if word==”user tag”:   Word=””
          Else if word==” hashtag”:  Word=” hashtag words”
          Else if word== emoji:   Word=”emoji to text”
END
```

Figure 5 – Tweet processing Algorithm

- Replacing MSA words in the dataset with lemmatized forms.

Figure 6 illustrates an example of the preprocessing step.



Tweet:
Arabic: ههههه هذا مكان اخطر من كورونا فايروس 😂😂🔥
Buckwalter Transliteration: hahaha hatha makan akthar min virus corona. Wajh bidmu' al-farah wajh bidmu' al-farah. Harq.
English translation Hahaha, this place is more dangerous than the coronavirus 😂😂🔥

Processed text: وجه بدموع الفرح وجه بدموع الفرح حريق.ههههه هذا مكان اخطر من كورونا فايروس
hahaha hadha makaan akhtar min fayrus korona. wajh bidumu' al-farah wajh bidumu' al-farah hareeq.

Hahaha, this place is more dangerous than the coronavirus. A face with tears of joy. A face with tears of joy. Fire.

Figure 6 – Example of the processing step

Finally, at the conclusion of the dataset analysis, words are encoded using the Word2Vec method. Word2Vec is a widely utilized technique for generating distributed word representations through simple neural networks (Mikolov, Chen, Corrado, & Dean, 2013). It consists of two models: Continuous Bag-of-Words (CBOW), which predicts the current word based on its surrounding context, and Skip-Gram, which predicts surrounding words given the current word. For our implementation, we utilize AraVec, an open-source project that offers pre-trained distributed word representations or word embeddings (Soliman, Eissa, & El-Beltagy, 2017).

3.2. Building the Bi-LSTM model

The dataset prepared earlier is used to train our proposed model, which relies on a Bi-LSTM architecture to analyze sentiment expressed in Algerian Arabic social media tweets.

The key concept behind our model is that natural language is a chronological, sequential, and contextual process where the meaning of a word is shaped by both its preceding and following words. Recurrent Neural Networks (RNNs) are effective sequence-learning models, connecting nodes across hidden layers to share features across different positions in the text. However, in traditional RNN and Long Short-Term Memory (LSTM) models, information flows only in a forward direction. To capture contextual information from both the past and the future, we employ a Bidirectional LSTM (Bi-LSTM), which combines a bidirectional RNN with LSTM units, enabling the model to gather information from both directions (Schuster & Paliwal, 1997).

The architectural design of a neural network depends on the input and output dimensions of each layer. Recurrent neural networks like Bi-LSTM can be adapted to various architectures, including many-to-many, many-to-one, and one-to-many configurations. They can also be integrated with different artificial neural network approaches, such as Convolutional Neural Networks (CNN), deep neural networks, and shallow neural networks.

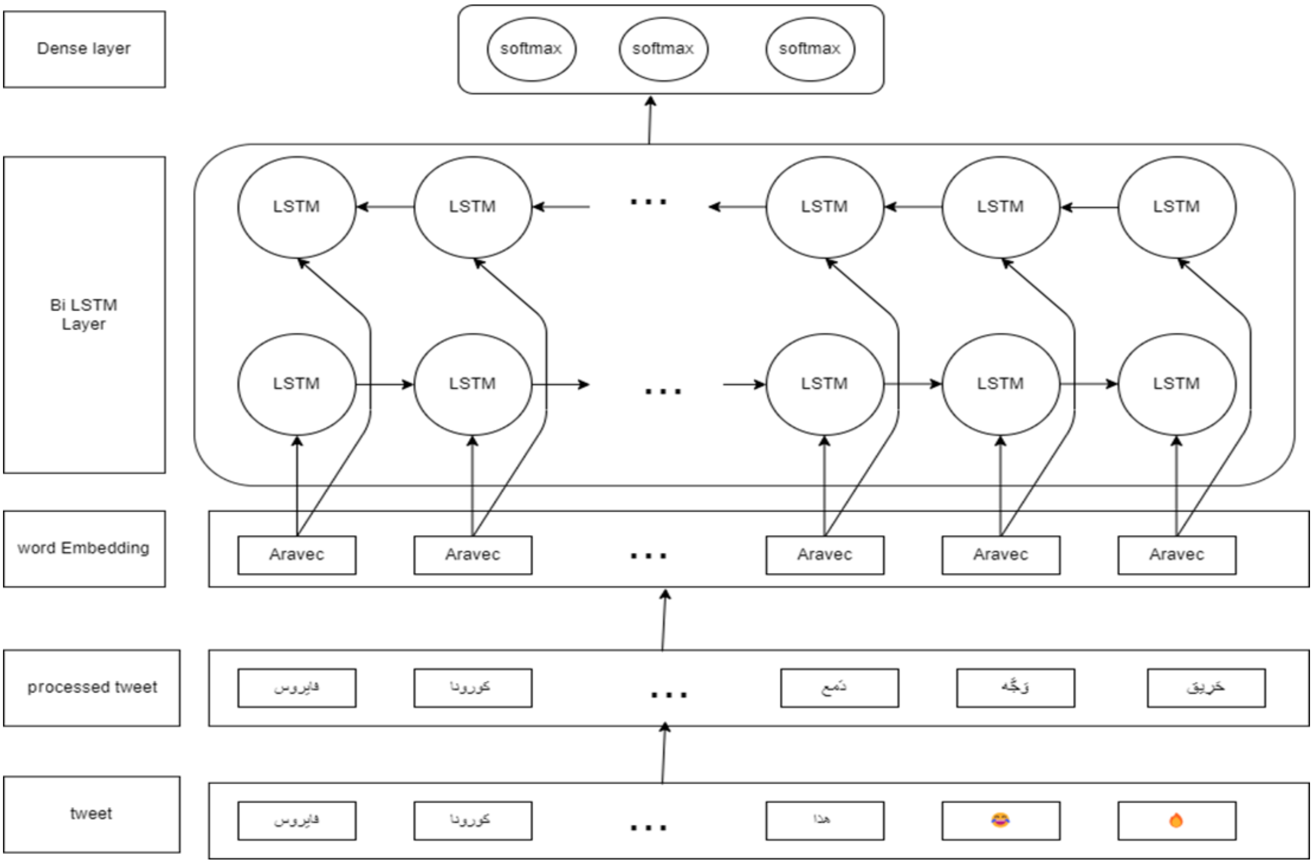


Figure 7 – Bi-LSTM recurrent neural network architecture for sentiment analysis

As shown in Figure 7, the initial input to the Bi-LSTM model is the embedding layer, which functions as a lookup table with keys representing the indices of the enumerated dataset words and values as their corresponding word embedding representations. The Bi-LSTM layer processes the word embeddings and the input sequence in both forward and backward directions. The Bi-LSTM units allow the model to capture context and dependencies between words in the sentence. Additionally, a dense layer with Softmax activation is applied for output classification.

To combat overfitting, we incorporate a dropout rate of 0.2. The model is compiled using the categorical cross-entropy loss function, the Adam optimizer, and accuracy as the evaluation metric.

4. Evaluation findings

The evaluation results of the proposed Bi-LSTM neural network, utilizing AraVec for word embeddings in sentiment analysis (SA) of the Algerian Arabic dialect, demonstrate impressive accuracy, as illustrated in Figure 8. Our comprehensive testing revealed that the model consistently achieved high performance metrics, including an accuracy of 0.94, precision of 0.93, recall of 0.91, and an F1-score of 0.92. These results indicate the model's remarkable capability in accurately classifying sentiment in textual data, as evidenced by the confusion matrix.

The combination of the Bi-LSTM architecture and word embeddings has proven to be a powerful approach for capturing complex contextual relationships and semantic meanings within the text, resulting in robust SA performance. The Algerian dialect, prevalent in social media, includes a rich mixture of non-Arabic words, which account for 24% of the vocabulary in our dataset. These findings highlight the model's potential for real-world applications, such as sentiment tracking in social media, customer feedback analysis, and opinion mining across various domains.

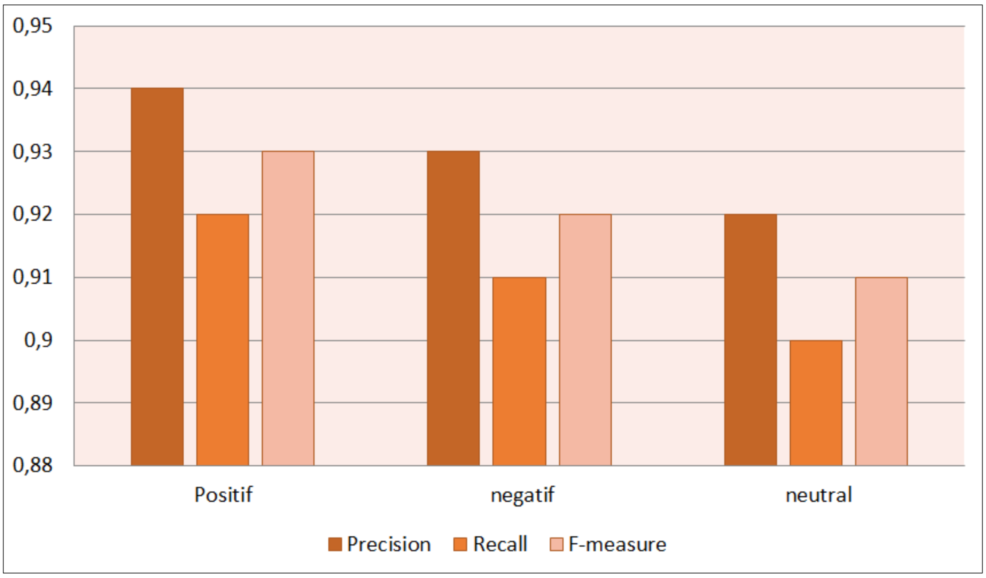


Figure 8 – Performance evaluation of the Bi-LSTM Neural Network with AraVec word embeddings for SA in the Algerian Arabic dialect

The future of this study holds exciting possibilities and potential advancements in several directions:

- **Model refinement**: Further enhancement of the proposed sentiment analysis model could involve integrating cutting-edge methodologies from deep learning and advanced language processing. This includes exploring innovative architectures, fine-tuning hyperparameters, and experimenting with novel word embedding strategies to boost model performance.
- **Dataset expansion**: Expanding the dataset to include more diverse and representative samples from the Algerian social network will be crucial. Incorporating a broader range of linguistic

expressions, such as colloquialisms, slang, and various writing styles, will enhance the model's adaptability and generalizability.

- **Multimodal analysis**: Integrating multimodal analysis, which involves processing not only textual data but also images, videos, and other media types, could provide a more comprehensive understanding of sentiment within the Algerian social network. This approach would enable the capture of nuanced emotions conveyed through visual content, contributing to a holistic sentiment analysis system.
- **Real-time adaptation**: Adapting the model for real-time sentiment analysis will be essential to keep pace with the dynamic nature of social media conversations. Optimizing the model for speed and efficiency will facilitate timely insights into evolving sentiments within the Algerian social network.
- **Cross-linguistic application**: Exploring the model's adaptability to other Arabic dialects or languages with similar linguistic characteristics can extend the impact of this study beyond the Algerian context. This cross-linguistic application would contribute to the development of sentiment analysis tools with broader applicability.
- **User feedback integration**: Incorporating user feedback and iterative model updates based on user interactions can enhance the model's performance over time. This feedback loop will foster a more user-centric and adaptive sentiment analysis system tailored to the evolving needs of the Algerian social network community.
- **Ethical considerations**: As sentiment analysis models become increasingly prevalent across various domains, addressing ethical considerations such as bias, fairness, and privacy is paramount. Future research should focus on ensuring that the developed model is ethical, fair, and respects user privacy, particularly in the context of the diverse linguistic expressions found in the Algerian social network.

By pursuing these avenues, this study can significantly contribute to the ongoing development of sentiment analysis technologies tailored to the specific characteristics and dynamics of the Algerian social network while also providing insights and innovations with broader applications in the field of computational linguistics.

5. Conclusion and perspectives

The Algerian social network is experiencing rapid growth, with its digital discourse continually evolving. To keep pace with this dynamic landscape, there is a pressing need for more data to be made available by and for the community. Advances in deep learning technologies offer promising opportunities to develop robust approaches, tools, and software tailored to the nuances of the Algerian Arabic dialect.

While recurrent neural networks and word embeddings have demonstrated efficiency in natural language classification tasks, their effectiveness often hinges on intensive computations and access to large datasets. Overcoming these challenges is crucial to unlocking the full potential of these technologies for the Algerian dialect.

This study presents a novel Bi-LSTM-based model specifically designed for sentiment analysis of the Algerian Arabic dialect. Leveraging word embeddings, the model successfully captures both the morphological and semantic subtleties of the text. The results underscore the model's high proficiency, achieving impressive metrics such as 0.94 accuracy, 0.93 precision, 0.91 recall, and 0.92 F1-score, reflecting its effectiveness in accurately classifying sentiment within user-generated content.

These findings contribute significantly to the development of sentiment analysis tools adapted to the specific characteristics of the Algerian social network, enabling deeper insights into digital discourse. Moving forward, this work sets the foundation for further research and application, with promising opportunities to extend sentiment analysis techniques to other dialects and languages, enhance real-time processing, and address ethical concerns surrounding AI in linguistic contexts.

References

Al-Ayyoub, M., Khamaiseh, A. A., Jararweh, Y., & Al-Kabi, M. N. (2019). A comprehensive survey of arabic sentiment analysis. *Information Processing & Management*, 56(2), 320-342. doi: https://doi.org/10.1016/j.ipm.2018.07.006

Alayba, A. M., Palade, V., England, M., & Iqbal, R. (2018). A Combined CNN and LSTM Model for Arabic Sentiment Analysis, Cham.

Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226, 107134. doi: https://doi.org/10.1016/j.knosys.2021.107134

Du, K., Xing, F., Mao, R., & Cambria, E. (2024). Financial Sentiment Analysis: Techniques and Applications. *ACM Comput. Surv.*, 56(9), Article 220. doi: https://doi.org/10.1145/3649451

Hadj Ameur, M. S., & Hassina, a. (2023). ARACOVID19-SSD: ARABIC COVID-19 SENTIMENT AND SARCASM DETECTION DATASET. *Revue de l'Information Scientifique et Technique*, 27(1), 8–15.

Kemp, S. (2023). Digital 2023: Algeria Retrieved September 2024, 2024, from https://datareportal.com/reports/digital-2023-algeria

Mikolov, T., Chen, K., Corrado, G. S., & Dean, J. (2013, May 2-4 2013). Efficient Estimation of Word Representations in Vector Space. Paper presented at the International Conference on Learning Representations, Scottsdale, Arizona.

Omara, E., Mousa, M., & Ismail, N. (2022). Character gated recurrent neural networks for Arabic sentiment analysis. *Scientific Reports*, 12(1), 9779. doi: https://doi.org/10.1038/s41598-022-13153-w

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, July). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. Paper presented at the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673-2681. doi: https://doi.org/10.1109/78.650093

Shanmugavadivel, K., Sathishkumar, V. E., Raja, S., Lingaiah, T. B., Neelakandan, S., & Subramanian, M. (2022). Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. *Scientific Reports*, 12(1), 21557. doi: https://doi.org/10.1038/s41598-022-26092-3

Soliman, A. B., Eissa, K., & El-Beltagy, S. R. (2017). AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP. *Procedia Computer Science*, 117, 256-265. doi: https://doi.org/10.1016/j.procs.2017.10.117

Wankhade, M., & Rao, A. C. S. (2022). Opinion analysis and aspect understanding during covid-19 pandemic using BERT-Bi-LSTM ensemble method. *Scientific Reports*, 12(1), 17095. doi: https://doi.org/10.1038/s41598-022-21604-7

Xu, G., Meng, Y., Qiu, X., Yu, Z., & Wu, X. (2019). Sentiment Analysis of Comment Texts Based on BiLSTM. *IEEE Access*, 7, 51522-51532. doi: https://doi.org/10.1109/access.2019.2909919