Arabic Intent Classification: From Data Creation to Prediction*

First Author[1][0000−1111−2222−3333], Second Author[2,3][1111−2222−3333−4444], and Third Author[3][2222−−3333−4444−5555]

[1] Princeton University, Princeton NJ 08544, USA
[2] Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany
lncs@springer.com
http://www.springer.com/gp/computer-science/lncs
[3] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. The growing volume of Arabic digital content has intensified the demand for intelligent systems capable of processing natural language. Intent classification, a fundamental task in Natural Language Processing (NLP), is essential for enabling effective human-computer interaction by identifying the underlying purpose of user input. However, Arabic intent classification remains challenging due to the language's morphological richness, dialectal diversity, and the scarcity of labeled data. In this paper, we propose a deep learning-based framework for Arabic intent classification, designed to enhance customer support services in social media and e-commerce platforms. A high-quality Arabic dataset was constructed, encompassing three primary intent categories: questions, requests, and complaints. We developed and fine-tuned two deep learning models: a Bidirectional Long Short-Term Memory (BiLSTM) network and a BERT-based transformer model. Both were evaluated using standard metrics, including accuracy and F1-score. Experimental results demonstrate the superior contextual understanding of BERT, while BiLSTM provides a lightweight and efficient alternative for resource-constrained environments. This work contributes to the advancement of Arabic NLP by addressing data scarcity, improving classification performance, and delivering practical tools for real-world applications.

Keywords: Arabic Intent Classification · Natural Language Processing · AraBERT · BiLSTM · Deep Learning · Customer Support Automation.

1 Introduction

The exponential growth of digital data in recent years, driven by widespread internet access and the proliferation of connected devices, has introduced significant challenges in the management and analysis of textual information. These challenges are particularly acute in applications that require real-time understanding of user intent, such as customer service systems, virtual assistants,

* Supported by organization x.

and conversational agents. Intent classification, a fundamental task in Natural Language Processing (NLP), addresses this need by identifying the underlying purpose behind user utterances, thereby enabling more natural and effective human-computer interaction.

In the context of Arabic NLP, intent classification poses unique challenges. Arabic is a morphologically rich and highly inflected language, characterized by diverse dialects, regional variations, and limited availability of labeled datasets. Despite having over 420 million speakers [1], Arabic remains underrepresented in NLP research compared to high-resource languages like English. These factors hinder the development of robust intent classification systems capable of handling real-world Arabic text, especially in informal domains such as social media.

Recent advances in deep learning have demonstrated strong capabilities in overcoming language-specific challenges. Architectures such as Bidirectional Long Short-Term Memory (BiLSTM) networks and transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) have shown remarkable performance in various NLP tasks, including intent classification. These models are capable of capturing contextual and semantic nuances in language, making them suitable for processing complex and noisy Arabic text.

In this paper, we present an Arabic intent classification framework tailored for customer support scenarios in social media and e-commerce platforms. Our contributions are twofold: (i) we construct a high-quality Arabic intent dataset encompassing three primary categories: questions, requests, and complaints; and (ii) we develop and fine-tune two deep learning models, BiLSTM and BERT, for intent classification.

Experimental evaluations demonstrate the superior contextual understanding of the BERT-based model, while the BiLSTM model offers a computationally efficient alternative with acceptable performance. The proposed system contributes to the advancement of Arabic NLP by improving classification accuracy, addressing the issue of data scarcity, and providing scalable tools for real-world deployment.

The remainder of this paper is organized as follows: Section 2 reviews related work in Arabic intent classification and deep learning approaches. Section 3 describes the proposed classification system, including its architecture, dataset creation process, and underlying models. Section 4 presents a comparative study of model performance. Finally, Section 5 concludes the paper and outlines directions for future research.

2 Related work

Arabic intent classification has garnered increasing attention in recent years, with researchers exploring a range of techniques, from traditional machine learning to modern deep learning and transformer-based models. Among the most notable advancements are Arabic-specific transformer models such as AraBERT, CAMeLBERT, and MARBERT, which have achieved state-of-the-art perfor-

mance in various Arabic NLP tasks [2]. These models are pretrained on large-scale Arabic corpora and are capable of capturing both contextual and semantic nuances. However, their effectiveness still relies heavily on the availability and quality of annotated datasets.

Several recent studies have tackled Arabic intent classification in specific domains. Skiredj et al. [3] introduced DarijaBanking, a dataset focused on Moroccan Arabic for banking-related intent detection, demonstrating the utility of AraBERT in handling dialectal variation. Galal et al. [4] applied CNN-BiLSTM and transformer models, including AraBERT and MARBERT, to detect sarcasm in Arabic tweets, an indirect but informative signal of user intent. Al Maruf et al. [5] conducted a comparative study of AraBERT and MARBERT for emotion and intent detection in social media content, concluding that MARBERT performed better on dialectal Arabic.

Similarly, Alqulaity et al. [6] investigated dialect identification in Arabic social media using CNN and transformer models, with intent classification contributing to chatbot enhancement. Mezahem [7] utilized CNN and MARBERT on movie reviews from Twitter, underlining MARBERT's effectiveness on noisy, informal text. In the e-commerce domain, Laraissia [8] used BiLSTM and CNN for sentiment analysis of Arabic product reviews, an approach extendable to intent classification. Berrimi [9] evaluated MARBERT and AraBERT in conversational systems, reporting a 71.1% F1-score for dialectal intent recognition using MARBERT. Additionally, Jefry et al. [10] provided a broader review of transformer-based Arabic NLP models, emphasizing their application in chatbots and cross-lingual tasks.

Table 1 summarizes these studies with respect to their domains, classification techniques, datasets, results, and notable observations.

Based on a thorough analysis of existing research, we identify several key gaps that motivate our proposed solution:

- **Lack of Arabic datasets for customer service scenarios:** Most Arabic intent classification datasets are domain-specific (e.g., banking, health, education) or general-purpose. There is a noticeable absence of annotated datasets tailored to customer support contexts, particularly on informal platforms like social media and e-commerce websites.
- **No smart intent classification system for customer support:** Current systems do not provide an integrated, real-time, AI-driven framework for automatically classifying customer intents in Arabic. A comprehensive system capable of routing queries and prioritizing responses could significantly enhance service efficiency.
- **Underrepresentation of dialectal Arabic:** Many models are trained predominantly on Modern Standard Arabic (MSA), limiting their performance in real-world applications where dialectal Arabic, often mixed and informal, is dominant, especially in user-generated content.
- **Absence of structured intent taxonomies for Arabic customer interactions:** There is no standardized taxonomy reflecting practical Arabic intents in business contexts (e.g., product inquiry, delivery issue, complaint,

Table 1. Summary of Arabic intent classification studies by domain, technique, and results.

Study	Domain	Technique	Dataset	Results	Notes
[3]	Banking (DA)	AraBERT	DarijaBanking	F1-score ≈ 0.90	focused on Moroccan Arabic intent detection in banking systems
[4]	Social Media	CNN-BiLSTM, AraBERT, MARBERT	Twitter	Accuracy = 87%	used sarcasm detection as a proxy for implicit intent
[5]	Social Media	AraBERT, MARBERT	Twitter	MARBERT better	MARBERT outperformed AraBERT in dialectal emotion/intent tasks
[6]	Social Media	CNN, MARBERT	Twitter	F1-score > 0.88	used for improving multi-dialect chatbot understanding
[7]	Reviews	CNN, MARBERT	Twitter (Movie Reviews)	Accuracy = 85%	MARBERT effective on noisy, user-generated text
[8]	E-commerce	BiLSTM, CNN	Arabic e-commerce reviews	Accuracy = 86%	demonstrated potential for use in commercial intent classification
[9]	Chatbots	AraBERT, MARBERT	Arabic dialogue corpus	MARBERT ≈ 71.1%	MARBERT performed better in dialogue understanding
[10]	General NLP	AraBERT, MARBERT, QARIB	Multi-domain	Comparative	Reviewed transformer use in Arabic chatbot and cross-lingual tasks

refund, general feedback). This lack of structure complicates both training and evaluation.
– **Insufficient handling of imbalanced and noisy real-world data:** Real-world Arabic text data, particularly from customer interactions, is often imbalanced across intent types and includes slang, typos, emojis, and non-standard grammar. Existing datasets do not adequately reflect these complexities, limiting the generalizability of trained models.

Addressing these gaps, our work contributes a novel dataset and a practical classification framework focused on Arabic customer support across informal digital channels.

3 Arabic intent classification model

This section presents an innovative system for Arabic intent classification, designed to support customer service on social media and e-commerce platforms. Addressing challenges such as the scarcity of domain-specific datasets and the lack of real-time intelligent systems, the proposed solution integrates a custom-built annotation tool with deep learning models, Bidirectional Long Short-Term Memory (BiLSTM) and AraBERT, to accurately classify customer messages into predefined intents such as Questions, Requests, and Complaints. By tackling Arabic's morphological complexity and dialectal diversity, this system aims to automate intent recognition and enhance customer support efficiency in Arabic-speaking markets.

3.1 Overall architecture

As illustrated in Fig. 1, the proposed system comprises three main components: the Input module, Deep Learning module, and Output module.

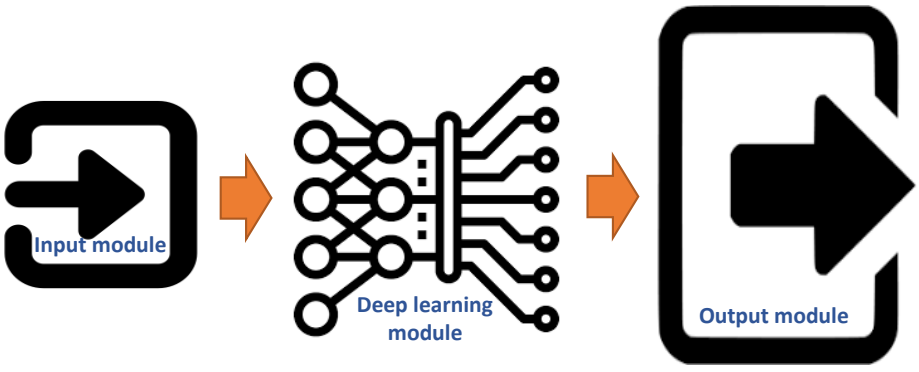

Fig. 1. Arabic intent classification system architecture.

The system operates in three distinct phases: Training, Evaluation, and Prediction. Each phase interacts with the architecture differently:

- **Training phase:** Uses 70% of the dataset for model training. No direct output is generated during this phase.
- **Evaluation (Testing) phase:** Applies the remaining 30% of the dataset to assess model performance, producing metrics such as accuracy and F1-score.
- **Prediction (Inference) phase:** Accepts new user input and classifies it into an intent, which is then displayed as output.

The deep learning module is implemented using either the BERT-based model or the BiLSTM-based model. The input/output modules vary according to the operational phase. The following subsections detail the dataset and the architecture of each model.

3.2 Dataset

A major limitation in Arabic NLP is the lack of publicly available datasets tailored to specific domains. While several domain-specific corpora exist in fields such as healthcare and education, datasets explicitly targeting customer support on informal platforms, like social media and e-commerce, remain scarce [11]. These domains require real-time understanding of informal, often dialectal, user input.

To address this gap, we constructed a new dataset by manually collecting Arabic customer messages from platforms such as Twitter, Instagram, and e-commerce websites like Amazon.ae and Noon. The messages, written in both Modern Standard Arabic (MSA) and various dialects, were manually labeled into one of three intent categories: Complaint, Question, or Request.

As shown in Fig. 2, the final dataset contains 2598 labeled instances, distributed nearly equally across the three categories: 865 questions, 866 requests, and 867 complaints. This balanced distribution supports robust training and evaluation.

3.3 BERT-based model

BERT (Bidirectional Encoder Representations from Transformers) is a deep learning model introduced by Google that uses a transformer architecture to generate context-aware representations of language [12]. Unlike traditional models, BERT processes words by considering both preceding and succeeding tokens, allowing it to better capture meaning in context.

For Arabic, AraBERT [13] is a pretrained language model fine-tuned on large-scale Arabic corpora, enabling precise classification of diverse intent types across formal and informal Arabic. Integrated with our custom dataset, AraBERT can effectively process varied language styles, supporting intent recognition in noisy, real-world settings.

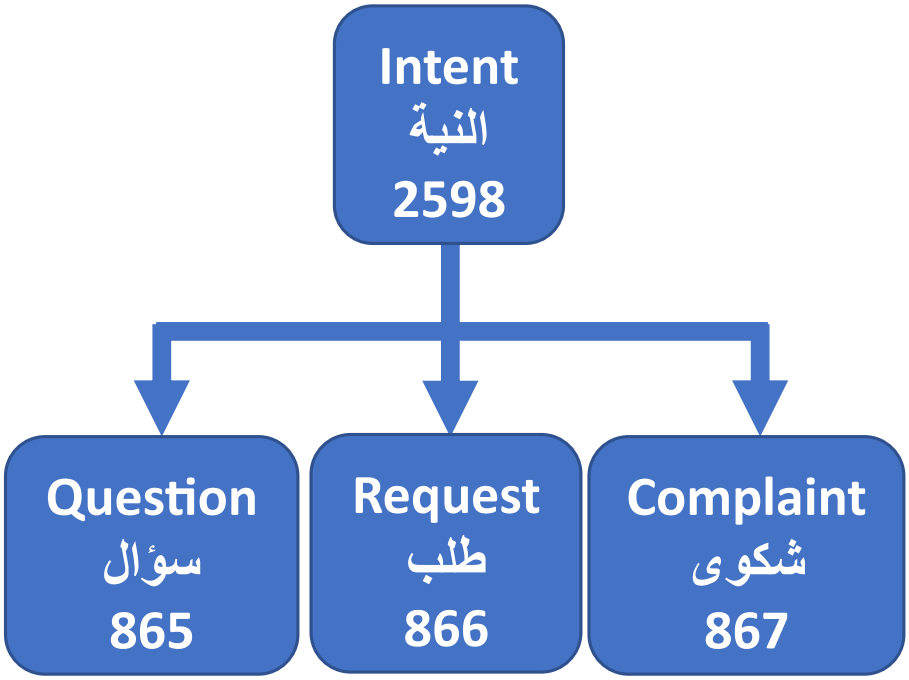

Fig. 2. Dataset taxonomy.

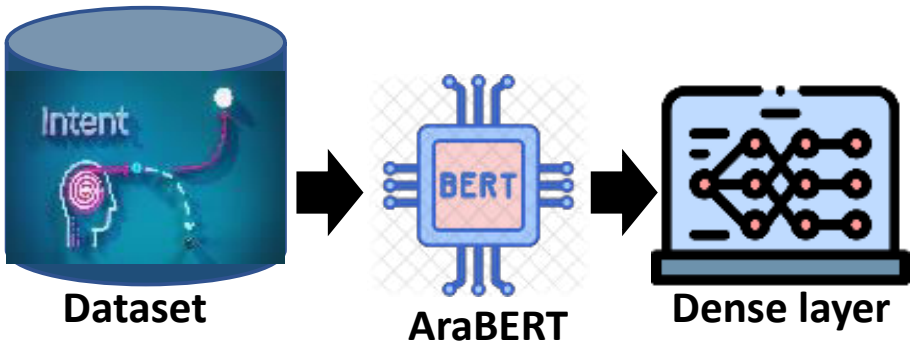

Fig. 3. BERT-based Arabic intent classification model.

As illustrated in Fig. 3, the BERT-based classifier leverages AraBERT for nuanced intent classification. It splits the dataset: 70% for training and 30% for evaluation. Arabic text is tokenized with BERT's tokenizer, adding special tokens ([CLS], [SEP]). The tokenized input is encoded by BERT's 12-layer transformer architecture (12 attention heads, 768 hidden size), generating contextualized representations. These are fed into a classification head (Dense layer with Softmax) to predict intents. The model is fine-tuned on the dataset to optimize performance for formal and dialectal Arabic. During Prediction, a new message is tokenized, encoded by the trained AraBERT model, and classified. This architecture excels at capturing nuanced intent from complex customer messages.

3.4 BiLSTM-based model

BiLSTM (Bidirectional Long Short-Term Memory) networks extend traditional LSTMs by processing input sequences in both forward and backward directions, improving the model's ability to understand context in sequence data [14, 15]. This capability is particularly useful for languages like Arabic, which have flexible word order and rich morphology.

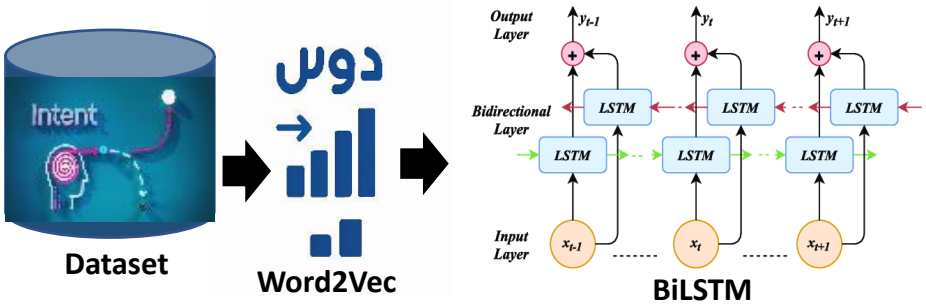

Fig. 4. BiLSTM-based Arabic intent classification model.

As depicted in Fig. 4, the BiLSTM classifier architecture is tailored for efficient intent classification of Arabic customer messages. It integrates an Arabic dataset, word embeddings, and a bidirectional LSTM network. The dataset is split into 70% for training and 30% for evaluation. Arabic text is tokenized and converted into word embeddings using a pretrained embedding layer, called Word2Vec [16, 17]. These embeddings are fed into a BiLSTM network with two layers (128 units each direction) to capture contextual dependencies in Arabic's complex syntax. A dense layer with Softmax activation produces intent probabilities. For real-time Prediction, new messages are tokenized, embedded, and classified using the trained BiLSTM model. This architecture offers a lightweight yet effective alternative for intent classification, particularly suitable for resource-constrained environments.

4 Comparative study

This section presents a comparative analysis of the BERT-based and BiLSTM-based models for Arabic intent classification, based on standard evaluation metrics: Precision, Recall, F1-Score, and Accuracy. These metrics provide a comprehensive understanding of each model's classification performance.

The evaluation metrics are defined as follows:

- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-Score = 2 × (Precision × Recall) / (Precision + Recall)
- Accuracy = (TP + TN) / (TP + FP + FN + TN)

where:

- TP (True Positives): The model correctly predicted a message as belonging to a specific intent category.
- FP (False Positives): The model incorrectly predicted a message as belonging to a specific intent, when it actually belongs to another.
- FN (False Negatives): The model missed a message that truly belongs to a specific intent; it predicted another class instead.
- TN (True Negatives): The model correctly rejected a message from a specific intent; it correctly classified it as a different class.

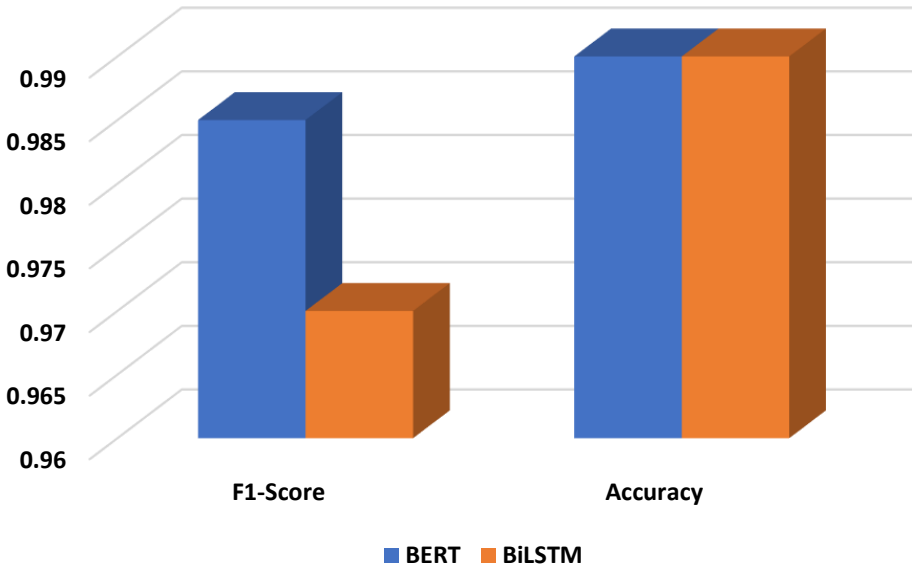

Fig. 5. Performance comparison of BERT vs. BiLSTM.

As shown in Fig. 5, both models achieved strong results, with accuracy scores of 0.99, indicating their high reliability in predicting the correct intent. However, a more nuanced comparison using the F1-score reveals the performance difference. While both models maintain excellent accuracy, BERT slightly outperforms BiLSTM in terms of F1-score. This suggests that BERT is more effective at balancing precision and recall, particularly in handling ambiguous or morphologically complex input. Its transformer architecture allows it to capture deep contextual dependencies and better generalize across formal and dialectal Arabic.

In contrast, the BiLSTM model, while effective, is limited by its sequential processing and lacks the global context modeling capability of transformers. This affects its ability to disambiguate similar intents when contextual clues are subtle. Nevertheless, BiLSTM demonstrated notable efficiency in terms of training speed and computational resource usage, making it a valuable option for low-resource or time-constrained deployment scenarios.

Overall, the comparative analysis suggests the following:

- BERT is preferable in applications where precision, nuanced understanding, and robustness are essential, such as intelligent customer support and automated triaging.
- BiLSTM remains a strong baseline, suitable for environments where speed, simplicity, and resource efficiency are prioritized over contextual depth.

The final choice between the models may depend on deployment constraints, model interpretability needs, and infrastructure availability. In our case study, BERT proved more effective for building a high-performance Arabic intent classification system tailored for customer service automation.

5 Conclusion and perspectives

This paper has presented a comprehensive study on Arabic intent classification using deep learning techniques, aimed at enhancing customer support services in social media and e-commerce platforms. By tackling key challenges in Arabic Natural Language Processing (NLP), including the scarcity of labeled data, dialectal diversity, and morphological complexity, this work contributes valuable resources and insights to the field of Arabic language understanding.

Several notable contributions have been made. First, a manually constructed Arabic intent classification dataset was introduced, encompassing three practical and business-relevant intent categories: Question, Request, and Complaint. This dataset is specifically tailored to real-world customer service scenarios, reflecting both formal and dialectal language usage. Second, two deep learning models, BiLSTM and BERT (AraBERT variant), were developed and evaluated. The comparative analysis demonstrated that while BiLSTM offers efficiency and simplicity, the BERT-based model achieves superior contextual understanding and overall performance, particularly in complex or ambiguous cases.

To further improve and extend this research, several future directions are proposed:

- Dataset expansion: Increasing the dataset's size and diversity by incorporating more intent categories, dialects, and domain-specific messages (e.g., healthcare, logistics) will improve model generalization and robustness.
- Multi-label classification: Supporting multi-intent classification would enable the system to better reflect the complexity of real user input, where a message may express multiple intents simultaneously.
- Model optimization: Investigating lighter transformer-based models such as DistilBERT, AraELECTRA, or ALBERT may offer better trade-offs between accuracy and computational efficiency, especially for deployment on low-resource devices.
- System deployment: Integrating the trained models into real-time applications, such as chatbots, helpdesk systems, or web APIs, would facilitate practical adoption and enable end-to-end customer support automation in Arabic.
- Human-in-the-loop feedback: Introducing feedback loops from customer support agents could help refine model predictions and continuously improve classification quality over time.

In conclusion, this work not only advances the state of Arabic intent classification but also lays the foundation for intelligent, scalable, and language-aware customer service systems in the Arabic-speaking world.

References

1. Al-dihaymawee, D.T.M., Merzah, A.A., Abdul Ridha, H.M.: The Story of Arabic Language: Historical Linguistics Study. Tasnim International Journal for Human, Social and Legal Sciences 3, 572-582 (2024)
2. Azzeh, M., Qusef, A., Alabboushi, O.: Arabic Fake News Detection in Social Media Context Using Word Embeddings and Pre-trained Transformers. Arabian Journal for Science and Engineering 50, 923-936 (2025)
3. Skiredj, A., Azhari, F., Berrada, I., Ezzini, S.: DarijaBanking: A new resource for overcoming language barriers in banking intent detection for Moroccan Arabic speakers. Natural Language Processing 1-31 (2024)
4. A. Galal, M., Hassan Yousef, A., H. Zayed, H., Medhat, W.: Arabic sarcasm detection: An enhanced fine-tuned language model approach. Ain Shams Engineering Journal 15, 102736 (2024)
5. Maruf, A.A., Khanam, F., Haque, M.M., Jiyad, Z.M., Mridha, M.F., Aung, Z.: Challenges and Opportunities of Text-Based Emotion Detection: A Survey. IEEE Access 12, 18416-18450 (2024)
6. Alqulaity, E.-Y., Yafooz, W.-M.S., Alourani, A., Jaradat, A.: Arabic Dialect Identification in Social Media: A Comparative Study of Deep Learning and Transformer Approaches. Intelligent Automation & Soft Computing 39, 907–928 (2024)
7. Mezahem, F.H.: Sentiment Analysis for Arabic Social media Movie Reviews Using Deep Learning. The British University in Dubai (2023)

8. LARAISSIA, L.Y.: Sentiment Analysis for Arabic Language using Advanced Deep Learning Techniques. (2024)
9. Berrimi, M.: Deep models for understanding and generating textual arabic data. (2024)
10. Jefry, W.G., Al-Doghman, F., Hussain, F.K.: A Review of Trends and Challenges in Adopting AI Models Through Cross-Lingual Transfer Learning Via Sentiment Analysis. KeAi: International Journal of Intelligent Networks (2023)
11. Alzamzami, F.: Towards Domain-Independent Multi-Lingual-Dialectal Online Social Behavior Modeling. Université d'Ottawa/University of Ottawa (2024)
12. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171-4186. Association for Computational Linguistics, (Year)
13. Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based Model for Arabic Language Understanding. pp. 9-15. European Language Resource Association, (Year)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9, 1735-1780 (1997)
15. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE transactions on Signal Processing 45, 2673-2681 (1997)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2, vol. 2, pp. 3111–3119. Curran Associates Inc., Lake Tahoe, Nevada (2013)
17. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32, pp. II–1188–II–1196. JMLR.org, Beijing, China (2014)