University of Tripoli at AraSentEval: Fine-Tuning MARBERTv2 and CAMELBERT for Multi-Dialect Arabic Sentiment Analysis

Amani Bahlul Sharif, Sarah Farag S. Hmeid, Abdusalam F.A. Nwesri
University of Tripoli
Am.Sharif@uot.edu.ly, Sa.Hamid@uot.edu.ly, a.nwesri@uot.edu.ly

Abstract

This paper presents our contribution to the AraSentEval 2026 shared task, specifically for Subtask 1: Arabic Dialect Sentiment Analysis, hosted at the OSACT7 workshop during LREC 2026. The task focuses on classifying the sentiment (positive, negative, neutral) of text written in four major Arabic dialects: Moroccan, Egyptian, Jordanian, and Saudi. We addressed this by fine-tuning several pre-trained language models, including MARBERTv2 and CAMELBERT, on the provided Multi-Dialect-Sent (MDS-3) dataset. Our best-performing system MARBERTv2, achieved a Macro F1-score of 84.29% on the official test set, securing fourth place among 13 participating teams. Our findings underscore the value of leveraging large pre-trained models tailored to dialectal Arabic for improved sentiment classification in this under-resourced domain.

Keywords: Natural Language Processing (NLP), Sentiment Analysis(SA), Arabic Dialects, Fine-tuning, MARBERTv2, CAMELBERT.

1. Intoduction

The rapid growth of user-generated content on social media, review platforms, and online forums has led to an explosion of opinionated Arabic text, making sentiment analysis a key task for understanding public attitudes toward products, services, and social issues Abo et al. (2019). While sentiment analysis is now a mature field for English and other high-resource languages, work on Arabic remains relatively limited despite Arabic's status as one of the most widely used languages on the web Oueslati et al. (2020).

Arabic has its own challenges in sentiment analysis, given its complex morphology, word-formation processes, the fact that it's a right-to-left script, as well as the presence of Modern Standard Arabic (MSA) and its many dialects, which vary in vocabulary, spelling, and grammar Katat et al. (2024). These dialects have become very common in social media, which has its own set of complexities, such as the use of informal spelling, code-switching, the use of emojis, etc., as mentioned in Shamsi and Abdallah (2021). As a result, many techniques that work well for English or even MSA often degrade significantly when applied to noisy, dialectal Arabic text.

Over the past decade, research has explored a variety of approaches to Arabic sentiment analysis, ranging from lexicon-based and traditional machine learning methods to deep learning and transformer-based architectures Katat et al. (2024). Surveys consistently report that deep learning models and, more recently, pre-trained language models (e.g., AraBERT-like architectures) tend to outperform classic classifiers, especially when combined with suitable text preprocessing and rich embeddings Katat et al. (2024). Nevertheless, progress is constrained by persistent obstacles such as the scarcity of high-quality labeled datasets, inconsistent evaluation protocols, and limited coverage of dialects and domains Aliyu et al. (2024).

Shared tasks and evaluation campaigns have a significant role to play in filling these gaps by providing standardized benchmarks and common metrics and through competitive leaderboards that allow for fair comparison and methodological innovation Muaad et al. (2025). The AraSentEval benchmark follows this line of work by offering a unified evaluation framework for Arabic sentiment analysis across multiple datasets and domains. It encourages researchers to tackle core challenges—including dialectal variation, domain shift, and noisy user-generated text—and to systematically compare traditional machine learning pipelines with modern transformer-based models.

This paper describes the system that was submitted to the AraSentEval competition, and it is hoped that the reader will find the information provided to be informative and helpful in understanding the effectiveness of transformer-based models in Arabic sentiment analysis, which is demonstrated through the fourth position achieved on the leaderboard.

2. Related Work

Previous research on Arabic sentiment analysis has largely focused on Modern Standard Arabic (MSA) , while Arabic dialects have received comparatively limited attention Matrane et al. (2023). Recent surveys confirm that most available datasets and sentiment lexicons were designed for MSA, whereas dialectal resources remain relatively

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 296–301

small, domain-specific, or still under development Aladeemy et al. (2024); Alharbi et al. (2025b).

Early approaches to Arabic sentiment analysis relied heavily on lexicon-based methods , where sentiment polarity is inferred using predefined lists of annotated words. Notable resources such as the Arabic Senti-Lexicon and the Multi-domain Arabic Sentiment Corpus demonstrated that sentiment lexicons can significantly improve classification performance when combined with traditional classifiers Al-Moslmi et al. (2018); Badaro et al. (2014); Assiri et al. (2018); Alsemaree et al. (2024).

To address the limitations of lexicon-based approaches—particularly their reliance on manual resources and limited contextual understanding—researchers increasingly adopted supervised machine learning techniques . Over the past decade, algorithms such as Support Vector Machines (SVM) and Naive Bayes have consistently demonstrated strong performance in Arabic sentiment analysis, commonly using bag-of-words or TF–IDF features Qader et al. (2019); enc (2010); Aladeemy et al. (2024). Hybrid approaches that combine lexicon-based features with n-gram or embedding representations have shown further improvements in classification accuracy Matrane et al. (2023); Jaber et al. (2025). Additionally, semi-supervised and ensemble methods have been explored for low-resource dialectal datasets, achieving F1 scores approaching 89% on Algerian and mixed MSA/dialect Facebook data Jaber et al. (2025).

More recently, the field has shifted toward large pre-trained transformer models , which enable end-to-end fine-tuning rather than manual feature engineering. Models such as AraBERT and ArabicBERT have demonstrated significant improvements over traditional machine learning and earlier deep learning approaches on benchmark datasets including SemEval, ASTD, and ArSAS Alduailej and Alothaim (2022). Subsequent transformer models specifically designed for dialectal Arabic—such as MARBERT, CAMeLBERT and QARiB —have further improved performance, particularly for noisy Twitter data and multi-dialect corpora , establishing strong state-of-the-art baselines for dialect sentiment analysis Bourahouat et al. (2024); Alotaibi and Nadeem (2025); Mutawa and Sruthi (2025).

Another important development is the CAMeLBERT family of models developed by CAMeL Lab. These BERT-based encoders are pre-trained on large and diverse Arabic corpora covering MSA, dialectal Arabic (DA), Classical Arabic (CA), and mixed text varieties Alotaibi and Nadeem (2025); Mutawa and Sruthi (2025). Specialized variants such as CAMeLBERT-DA-sentiment, which are fine-tuned on major Twitter datasets including ASTD, ArSAS, and SemEval, have achieved strong F1 scores and often outperform multilingual models while remaining competitive with MARBERT on dialectal sentiment tasks Alotaibi and Nadeem (2025); Alahmadi (2025).

Several approaches were also evaluated in the AHaSIS 2025 shared task on Arabic dialect sentiment analysis (Alharbi et al., 2025b). The top-ranked system employed an ensemble of MARBERTv2, SaudiBERT, and DarijaBERT, achieving a micro-F1 score of 0.81 Al-Khalifa (2025). The second-ranked system used fine-tuning strategies for AraBERTv02, achieved a micro-F1 score of 0.7916 Jaballah et al. (2025). The third-ranked LBY team evaluated several transformer models—including MARBERTv2, AraBERT variants, QARiB, and DarijaBERT—and found that MARBERTv2 achieved the best performance, relying primarily on careful hyperparameter tuning rather than ensembling, with an F1 score of 0.79 (Nwesri et al., 2025).

Overall, current research indicates that Arabic-specific transformer models pre-trained on multi-dialect corpora , particularly MARBERT, AraBERT, QARiB, and CAMeLBERT , represent the most effective family of models for Arabic and dialectal sentiment analysis and frequently serve as the foundation for state-of-the-art single or ensemble systems Mansour et al. (2025); Mutawa and Sruthi (2025).

3. Methodology

3.1. Task

Subtask 1 of AraSentEval 2026 focuses on Arabic Dialect Sentiment Analysis, a multi-class classification problem. The goal is to predict the sentiment polarity of text written in various Arabic dialects. Participants were provided with a training dataset containing sentences from four major dialects: Moroccan, Egyptian, Jordanian, and Saudi. Each sentence is labeled with one of three sentiment classes: positive, negative, or neutral. During the evaluation phase, participants received an unlabeled test set and were required to submit predicted labels for each instance. The official evaluation metric for this subtask is the Macro F1-score, which calculates the F1-score for each class independently and then takes their average. This metric was chosen by the organizers to ensure fair evaluation across all sentiment categories, regardless of any potential class imbalance in the test data Ezzini et al. (2026). The task challenges participants to develop models that can generalize across multiple dialects, handling the lexical, morphological, and syntactic variations that distinguish each dialect from both Modern Standard Arabic and from one another.

3.2. Dataset

The dataset consists of 1,731 hotel reviews written in four Arabic dialects: Saudi, Maghribi (Darija), Egyptian, and Jordanian. Text reviews are annotated as either "Positive", "Negative", or "Neutral", and the specific dialect is included for each entry. The training set is well-distributed across these dialects; the Saudi, Darija, and Jordanian subsets each contain 152 positive, 166 negative, and 115 neutral reviews, while the Egyptian subset contains 153 positive, 159 negative, and 120 neutral reviews (Alharbi et al., 2025c).

The test set, released later during the evaluation phase, comprises 312 reviews (Alharbi et al., 2025a). Participants are required to predict the sentiment polarity of these reviews using their developed models. An additional column containing the predicted labels must be added to the test set prior to submission on the shared task platform, where automatic evaluation is conducted. Each participant is allowed a maximum of 10 submissions during the evaluation phase. Each team is allowed to show its best result on a public leaderboard, showcasing the performance of all participating teams.

3.3. Models

The benchmark for this task is the pre-trained BERT-based model, which was fine-tuned on Modern Standard Arabic (MSA) and various Arabic dialects. In our proposed method, the emphasis is on using and fine-tuning state-of-the-art models that have been developed to address the language complexities of both Modern Standard Arabic and the various Arabic dialects. For this purpose, the MARBERTv2 model, which is a strong language model developed to work on dialectal Arabic, and the models developed by the CAMeL-Lab particularly the "CAMeL-Lab/bert-base-arabic-camelbert-mix-sentiment", which have gained recognition for their strong performance on Natural Language Processing tasks in the Arabic language, have been used. They were chosen based on their effectiveness in identifying the sentiment expressed in the Saudi, Maghribi, Egyptian, and Jordanian dialects used in the dataset.

3.4. Evaluation Measure

The models were primarily evaluated using the F1-score and accuracy across the entire dataset, encompassing all four dialects simultaneously. This approach allowed for a direct comparison between MARBERTv2 and CAMELBERT to determine their overall performance on mixed dialectal Arabic.

4. Experiments: Models training

In our research, we decided to fine-tune two of the most advanced Arabic pre-trained models, namely MARBERTv2 and CAMELBERT, for the multi-dialectal sentiment analysis task. The fine-tuning was done using the Hugging Face's Trainer API, while all the experiments were carried out on the Google Colab platform Bisong (2019).

It should be noted that the dataset was divided into 80% for training and 20% for evaluation. During the preprocessing stage, the dedicated tokenizer associated with each model was employed. The text was processed through tokenization, truncation, and padding, with the maximum sequence length set to 128 tokens. In addition, several hyperparameters were optimized to achieve the best performance, including the learning rate, batch size, weight decay, and warmup steps. The models were trained for a predefined number of epochs, with the evaluation conducted at the end of each epoch. The macro-averaged F1 score (F1-macro) was used as the primary evaluation metric throughout the experiments. The best performance was obtained using the following hyperparameter configuration: a learning rate of 1e-4, batch size of 16, weight decay of 1e-4, and warmup steps of 0.1. Under these settings, the models were trained for 8 epochs, with evaluation performed after each epoch.

4.1. Fine-tuning on the Original Dataset

Both MARBERTv2 and CAMELBERT models were fine-tuned using the original training dataset. Table 1 presents the best results achieved during the internal validation phase. Both models demonstrated strong performance, with improvements resulting from careful hyperparameter tuning. Among the two models, MARBERTv2 outperformed CAMELBERT, achieving an F1-macro score of 97% on the validation set. The CAMELBERT model followed closely, obtaining a competitive F1-macro score of 96

4.2. Fine-tuning on the Preprocessed Dataset

Social media text often contains various forms of noise, including emojis, symbols, and elongated characters used to express emotions. In addition, Arabic dialectal text is characterized by inconsistent writing styles, where the same word may appear in multiple orthographic forms due to informal spelling. This issue is commonly addressed through text normalization, where different letter variants are unified into a single standardized form.

In this experiment, the training dataset was preprocessed prior to model fine-tuning by removing all hyperlinks, strings beginning with the @ symbol, diacritics, repeated characters, normalizing specific Arabic letters by replacing أ or إ with ا ; ة with ه ; and ى with ي , and finally removing any extra spaces within the text.

Table 1 presents the results obtained by fine-tuning the models using both the original dataset and the preprocessed dataset.

The results show that training on the original dataset consistently outperformed training on the preprocessed dataset across all evaluation metrics. For MARBERTv2, training on the original data achieved slightly higher performance, with an accuracy of 0.964 and F1 score of 0.965, compared with 0.960 accuracy and 0.958 F1 on the preprocessed data, along with similar improvements in precision (0.963 vs. 0.953) and recall (0.964 vs. 0.955). A larger performance drop was observed for CAMeLBERT, where the model achieved 0.955 accuracy and 0.960 F1 on the original dataset but only 0.936 accuracy and 0.932 F1 after preprocessing, with precision and recall also decreasing notably. While MARBERTv2 consistently outperformed CAMeLBERT on the preprocessed dataset, the performance gap between the two models became smaller when trained on the original dataset. Overall, these findings suggest that aggressive preprocessing and text normalization may remove useful linguistic cues from Arabic social media text, whereas transformer-based models can effectively handle noisy and informal text, benefiting more from the original unprocessed data.

5. Results on the testset

Both trained models were run on the released unlabeled dataset. Table 2 shows results obtained using the trained models. Out of the 8 submissions we made during the test phase, the MARBERTv2 result was submitted to the leaderboard under the name of "University of Tripoli". The result placed us at the forth position among 13 participating teams.

6. Discussion

The experimental results demonstrate the effectiveness of fine-tuning large pre-trained Arabic language models for the task of multi-dialectal sentiment analysis. Both MARBERTv2 and CAMELBERT achieved strong performance during the validation phase, confirming the capability of transformer-based architectures to capture sentiment information across diverse Arabic dialects. The results also highlight the advantage of domain-specific pre-trained models that have been trained on large-scale Arabic corpora containing dialectal and social media content.

A key observation from the experiments is that training on the original dataset consistently produced better results than training on the preprocessed dataset. While preprocessing techniques such as removing diacritics, normalizing letters, and eliminating repeated characters are commonly used in traditional natural language processing pipelines, the results indicate that these operations may inadvertently remove useful contextual or stylistic information embedded in social media text. Elements such as character elongation and informal orthographic variations often carry implicit sentiment or emphasis, which transformer-based models can learn to interpret effectively. Consequently, aggressive normalization may reduce the richness of the input data and negatively affect model performance.

The impact of preprocessing was particularly evident in the CAMELBERT model, which experienced a more substantial performance drop when trained on the preprocessed dataset compared with MARBERTv2. This behavior may be attributed to differences in the pre-training corpora and tokenization strategies used by the two models. MARBERTv2, which is specifically optimized for Arabic social media text, appears to be more robust to textual variations and noise, allowing it to maintain relatively stable performance even after preprocessing. In contrast, the CAMELBERT model appears to rely more heavily on the original lexical patterns present in the raw data.

The evaluation on the unseen test set further confirms the effectiveness of the trained models. MARBERTv2 achieved an F1 score of 0.843, outperforming CAMELBERT, which obtained an F1 score of 0.824. These results are consistent with the validation experiments and demonstrate that MARBERTv2 provides better generalization for the multi-dialectal sentiment analysis task. The submitted MARBERTv2 model ranked fourth among 13 participating teams on the official leaderboard, indicating competitive performance within the shared task.

Overall, the findings suggest that modern transformer-based language models can effectively handle the inherent noise and variability present in Arabic social media text, reducing the need for extensive preprocessing. Instead, preserving the original textual structure may allow the models to leverage subtle linguistic cues that contribute to sentiment detection.

7. Conclusion

This study evaluated the effectiveness of fine-tuning two state-of-the-art Arabic pre-trained language models, MARBERTv2 and CAMeLBERT, for multi-dialectal sentiment analysis using the Hugging

	Original Dataset				Preprocessed Dataset			
Model	Acc.	F1	P	R	Acc.	F1	P	R
MARBERTv2	0.964	0.965	0.963	0.964	0.960	0.958	0.953	0.955
CAMELBERT	0.955	0.960	0.958	0.953	0.936	0.932	0.934	0.934

Table 1: Results of fine-tuning models on the preprocessed and the original training dataset.

Model	Acc.	F1	P	R
MARBERTv2	0.843	0.843	0.843	0.843
CAMELBERT	0.824	0.824	0.824	0.824

Table 2: Results of running trained models on the unlabeled testset.

Face Trainer framework and the macro-averaged F1 score as the main evaluation metric. Experimental results showed that both models achieved strong performance, with MARBERTv2 consistently outperforming CAMeLBERT, reaching the best validation performance with an F1-macro score of 0.965 when trained on the original dataset. The study also compared training on the original dataset and a normalized preprocessed version, finding that the original dataset yielded better results, suggesting that aggressive preprocessing may remove useful linguistic cues from Arabic social media text. Evaluation on the official test set further confirmed these findings, where MARBERTv2 achieved an F1 score of 0.843, ranking fourth among 13 participating teams in the shared task leaderboard. These results demonstrate the effectiveness of transformer-based Arabic language models for handling dialectal variation in sentiment analysis, while future work may explore ensemble approaches, data augmentation, dialect-aware preprocessing, and larger language models to further improve performance.

8. References

2010. Encyclopedia of machine learning.

Mohamed Elhag Mohamed Abo, Ram Gopal Raj, and Atika Qazi. 2019. A review on arabic sentiment analysis: State-of-the-art, taxonomy and open research challenges. *IEEE Access*, 7:162008–162024.

Hend Al-Khalifa. 2025. iwan-nlp at ahasis 2025: A stacked ensemble of arabic transformers for sentiment analysis on arabic dialects in the hospitality domain. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 7–13, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tareq Al-Moslmi, M. Albared, Adel Al-Shabi, N. Omar, and S. Abdullah. 2018. Arabic senti-lexicon: Constructing publicly available language resources for arabic sentiment analysis. *Journal of Information Science*, 44:345 – 362.

Amani A. Aladeemy, Ali Alzahrani, Mohammad H. Algarni, S. N. Alsubari, Theyazn H. H. Aldhyani, Sachin N. Deshmukh, O. I. Khalaf, Wing-Keung Wong, and Sameer Aqburi. 2024. Advancements and challenges in arabic sentiment analysis: A decade of methodologies, applications, and resource development. *Heliyon*, 10.

D. Alahmadi. 2025. Human versus ai: A comparative study of zero-shot llms and transformer models against human annotations for arabic sentiment analysis. *International Journal of Advanced Computer Science and Applications*.

Alhanouf Alduailej and Abdulrahman Alothaim. 2022. Araxlnet: pre-trained language model for sentiment analysis of arabic. *Journal of Big Data*, 9.

Maram Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025a. Evaluating large language models on arabic dialect sentiment analysis. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Maram I. Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025b. AHaSIS: Shared task on sentiment analysis for Arabic dialects. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 1–6, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Maram I Alharbi, Saad Ezzini, Hansi Hettiarachchi, Tharindu Ranasinghe, and Ruslan Mitkov. 2025c. Ador: Dataset for arabic dialects in hotel reviews: A human benchmark for sentiment analysis. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 187–191.

Yusuf Aliyu, A. Sarlan, Kamaluddeen Usman Danyaro, A. S. B. A. Rahman, and Mujaheed Abdullahi. 2024. Sentiment analysis in low-resource settings: A comprehensive review of approaches, languages, and data sources. *IEEE Access*, 12:66883–66909.

Alanoud Alotaibi and Farrukh Nadeem. 2025. An unsupervised integrated framework for arabic aspect-based sentiment analysis and abstractive text summarization of traffic services using transformer models. *Smart Cities*.

Ohud Alsemaree, A. Alam, S. Gill, and Steve Uhlig. 2024. An analysis of customer perception using lexicon-based sentiment analysis of arabic texts framework. *Heliyon*, 10.

Adel M. A. Assiri, Ahmed Z. Emam, and Hmood Al-Dossari. 2018. Towards enhancement of a lexicon-based approach for saudi dialect sentiment analysis. *Journal of Information Science*, 44:184 – 202.

Gilbert Badaro, Ramy Baly, Hazem Hajj, Nizar Habash, and Wassim El-Hajj. 2014. A large scale arabic sentiment lexicon for arabic opinion mining. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 165–173. Association for Computational Linguistics.

Ekaba Bisong. 2019. *Building Machine Learning and Deep Learning Models on Google Cloud Platform*. Apress, Berkeley, CA.

Ghizlane Bourahouat, Manar Abourezq, and N. Daoudi. 2024. Improvement of moroccan dialect sentiment analysis using arabic bert-based models. *Journal of Computer Science*.

Saad Ezzini, Shadi Abudalfa, Maram Alharbi, Salmane Chafik, Samaher Alghamdi, Reem Alotaibi, Hamzah Luqman, Mo El-Haj, and Paul Rayson. 2026. AraSentEval: A shared task on sentiment analysis and swapping in arabic. In *Proceedings of the 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7), co-located with the Language Resources and Evaluation Conference (LREC 2026)*, Palma, Mallorca, Spain.

Mustapha Jaballah, Dhaou Ghoul, and Ammar Mars. 2025. Fine-tuning arabert model for arabic sentiment detection. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 14–23, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Areej Jaber, Israa Bahati, and Paloma Martínez. 2025. Leveraging pre-trained embeddings in an ensemble machine learning approach for arabic sentiment analysis. *Frontiers in Artificial Intelligence*, 8.

Souha Al Katat, Chamseddine Zaki, Hussein Hazimeh, Ibrahim El Bitar, Rafael Angarita, and Lionel Trojman. 2024. Natural language processing for arabic sentiment analysis: A systematic literature review. *IEEE Transactions on Big Data*, 10:576–594.

Omar Mansour, Eman Aboelela, Remon Talaat, and Mahmoud Bustami. 2025. Transformer-based ensemble model for dialectal arabic sentiment classification. *PeerJ Computer Science*, 11.

Yassir Matrane, F. Benabbou, and N. Sael. 2023. A systematic literature review of arabic dialect sentiment analysis. *J. King Saud Univ. Comput. Inf. Sci.*, 35:101570.

Abdullah Y. Muaad, Md Belal Bin Heyat, F. Akhtar, Usman Naseem, Wadeea R Naji, Suresha Mallappa, and H. J. 2025. Artificial intelligence for text analysis in the arabic and related middle eastern languages: Progress, trends, and future recommendations. *Int. J. Intell. Syst.*, 2025.

A. Mutawa and S. Sruthi. 2025. A comparative evaluation of transformers and deep learning models for arabic meter classification. *Applied Sciences*.

Abdusalam F. Ahmad Nwesri, Nabila Almabrouk S. Shinbir, and Amani Bahlul Sharif. 2025. Sentiment analysis on arabic dialects: A multi-dialect benchmark. In *Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects*, pages 86–91, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Oumaima Oueslati, E. Cambria, M. Hajhmida, and Habib Ounelli. 2020. A review of sentiment analysis research in arabic language. *ArXiv*, abs/2005.12240.

Raheel Qader, François Portet, and Cyril Labbé. 2019. Semi-supervised neural text generation by joint learning of natural language generation and natural language understanding models. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 552–562.

Arwa A. Al Shamsi and Sherief Abdallah. 2021. Text mining techniques for sentiment analysis of arabic dialects: Literature review. *Advances in Science, Technology and Engineering Systems Journal*, 6:1012–1023.