A Hybrid Transformer-Based Model for Sentiment Analysis of Arabic Dialect Hotel Reviews

Rawand Alfugaha
College of Information Technology
Lusail University
Doha, Qatar
ralfoqha@lu.edu.qa

Mohammad AL-Smadi
Digital Learning and Online Education Office
Qatar University
Doha, Qatar
malsmadi@qu.edu.qa

Abstract

This paper describes the AraNLP system developed for the "Ahasis" shared task on sentiment detection in Arabic dialects for hotel reviews. The task involved classifying the overall sentiment of hotel reviews (Positive, Negative, or Neutral) written in Arabic dialects, specifically Saudi and Darija. Our proposed model, AraNLP, is a hybrid deep learning classifier that leverages the strengths of a transformer-based Arabic model (AraELECTRA) augmented with classical bag-of-words style features (TF-IDF). Our system achieved an F1-score of 76%, securing the 5th rank in the shared task, significantly outperforming the baseline system's F1-score of 56%.

1 Introduction

Arabic dialect sentiment analysis presents unique challenges due to morphological complexity, diglossia, and regional variations (Abdul-Mageed et al., 2021). While Modern Standard Arabic (MSA) has been well-studied, dialects like Saudi and Darija remain under-resourced despite their prevalence in user-generated content (Salameh et al., 2018; Talafha et al., 2020). Recent advances in transformer models have shown promise for Arabic NLP (Antoun et al., 2020), but dialect-specific adaptations remain limited. Hotel reviews are particularly challenging due to domain-specific terminology mixed with dialectal variations (AL-Smadi et al., 2023).

The Ahasis shared task (Alharbi et al., 2025a) presented a significant challenge in the field of Arabic Natural Language Processing (NLP), focusing on sentiment analysis in the hospitality domain, specifically for diverse Arabic dialects. Sentiment analysis of user-generated content, such as hotel reviews, provides invaluable insights for both businesses and consumers (Al-Smadi et al., 2019). However, the Arabic language, with its rich morphology and wide range of dialects, poses unique difficulties for NLP tasks. Modern Standard Arabic (MSA) is the formal version of the language, while numerous regional dialects (e.g., Egyptian, Levantine, Gulf, Maghrebi) are predominantly used in informal online communication, including hotel reviews. These dialects often lack standardized orthography and can differ significantly from MSA and from each other in terms of lexicon, syntax, and morphology (Birjali et al., 2021).

In this paper, we present our system, AraNLP, which participated in the "Ahasis" shared task. Our approach is a hybrid deep learning model that combines the contextual understanding capabilities of a pre-trained transformer-based model for Arabic (AraELECTRA) with the statistical strength of TF-IDF features. This hybrid architecture aims to capture both semantic nuances and important lexical cues from the review texts.

The rest of this paper is organized as follows: Section 2 sheds the light on related work, Section 3 demonstrates the research methodology, Section 4 presents the model results, Section 5 discusses the model results, and Section 6 concludes the research paper and provides insights for future work.

2 Related Work

Sentiment analysis in Arabic, particularly for dialectal Arabic, has garnered increasing attention from the research community. Early approaches often relied on lexicon-based methods, which utilize predefined dictionaries of words tagged with sentiment polarities (Birjali et al., 2021). While straightforward, these methods struggle with the nuances of dialects, context-dependent sentiment, and the lack of comprehensive dialectal lexicons.

Machine learning techniques, including Support Vector Machines (SVM), Naive Bayes, and Logis-

Proceedings of the Shared Task on Sentiment Analysis for Arabic Dialects (AHaSIS) associated with RANLP 2025,
pages 62–68, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-109-7-010

tic Regression, have been widely applied to Arabic sentiment analysis, often outperforming lexicon-based approaches when sufficient labeled data is available (Al-Smadi et al., 2018). These models typically rely on features such as n-grams, TF-IDF, and word embeddings. For instance, Al-Smadi et al. (2019) explored the use of morphological, syntactic, and semantic features to enhance aspect-based sentiment analysis of Arabic hotel reviews, demonstrating the value of linguistic features.

Deep learning models, particularly those based on Recurrent Neural Networks (RNNs) like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs), have shown significant promise in capturing sequential information and contextual dependencies in text (Alyami et al., 2022). Elfaik and Nfaoui (2020) and Ombabi et al. (2020) employed LSTM networks for aspect-based sentiment analysis of Arabic text, highlighting their effectiveness. More recently, AL-Smadi et al. (2023) proposed a GRU model combined with a multilingual universal sentence encoder for Arabic aspect-based sentiment analysis, achieving strong results.

Transformer-based models, such as BERT and its variants, have revolutionized the field of NLP by achieving state-of-the-art performance on various tasks, including sentiment analysis. Several pre-trained transformer models have been developed specifically for the Arabic language, such as ALBERT (Lan et al., 2019), AraBERT (Antoun et al., 2020), AraELECTRA (Antoun et al., 2021), QARiB(QCRI Arabic and dialectal BERT) (Abdelali et al., 2021), and CAMeLBERT (Inoue et al., 2021). These models are pre-trained on large Arabic corpora and can be fine-tuned for specific downstream tasks like sentiment classification. The use of such models is becoming increasingly common due to their ability to understand complex linguistic patterns and contextual information. Recent work has also explored hybrid approaches combining transformers with other neural network architectures. For example, Bourahouat et al. (2024) proposed BERT-based models that are pre-trained on Arabic datasets, namely AraBERT, QARIB, ALBERT, AraELECTRA, and CAMeLBERT integrated with machine learning and deep learning models such as SVM and CNN for sentiment analysis of Darija (Moroccan dialect).

Hybrid models combining transformers with sequential or ensemble components have gained traction. (Alzahrani et al., 2024) achieved 97% accuracy by integrating AraBERT with LSTM to model long-term dependencies in Arabic text. For dialect detection, (Saleh et al., 2025) proposed a stacked transformer framework (AraBERT and XLM-R) with a meta-learner, achieving 93% F1-score on the IADD dataset. In sentiment analysis, (Mansour et al., 2025) demonstrated that transformer ensembles outperform single-model approaches by aggregating linguistic features across dialects.

The Ahasis shared task builds upon this body of work by focusing on the challenging aspects of dialectal Arabic (Saudi and Darija) in the specific domain of hotel reviews. Our work contributes to this line of research by proposing a hybrid model that combines the strengths of transformer architectures with traditional feature engineering to tackle the sentiment analysis task in diverse Arabic dialects.

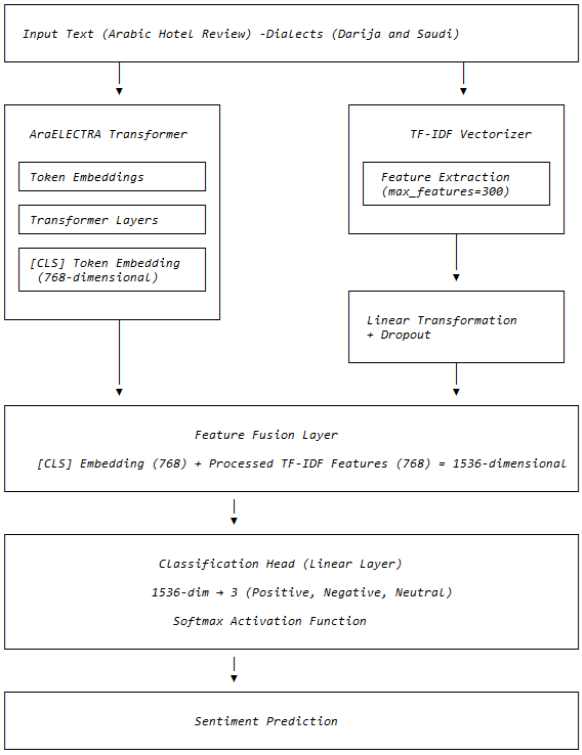

Figure 1: AraNLP Model Architecture: Hybrid integration of AraELECTRA and TF-IDF for sentiment classification of Arabic dialect hotel reviews.

3 Research Methodology

This section outlines the methodology employed in developing the AraNLP system. We first describe the shared task, followed by details of the dataset provided and our model architecture.

3.1 Task Definition

The Ahasis Shared Task focuses on sentiment classification in the hospitality domain for Arabic dialects. Specifically, given a hotel review written in either Moroccan Arabic (Darija) or Saudi dialect, the goal is to predict its overall sentiment as Positive, Neutral, or Negative (Alharbi et al., 2025a). Unlike aspect-based sentiment analysis which targets sentiment toward specific aspects (Alyami et al., 2022; AL-Smadi et al., 2023), this task concerns the general sentiment of the entire review. The official evaluation metric is Macro-averaged F1-score across the three sentiment classes, ensuring that performance on each class (including the often under-represented neutral class) is given equal importance. Participants were provided a labeled dataset (with a predefined train/test split) and a baseline model for reference. The baseline using AraBERT attained 56% Macro-F1, illustrating the difficulty of capturing sentiment in this domain and setting a performance bar for participants (Alharbi et al., 2025b).

3.2 Dataset

The shared task dataset consists of Arabic hotel reviews collected from online sources (e.g., booking websites or social media platforms). The training set contains 860 reviews and the test set contains 216 reviews. Both sets are evenly balanced across the two dialects and the three sentiment categories. In practice, this means the training data has roughly equal numbers of Moroccan Darija and Saudi reviews (approximately 430 each), and within each dialect the distribution of positive, neutral, and negative labels is also approximately equal. The reviews vary in length from short comments (a few words) to longer sentences. Some examples of typical review content include praise or complaints about the room, cleanliness, staff behavior, price, or location. Neutral reviews often describe the experience factually without strong emotion. Positive reviews might use enthusiastic phrases or adjectives (in dialect, e.g., "زوين بالزاف" meaning "very nice" in Darija), whereas negative reviews contain criticism or negative expressions (e.g., "ما عجبنيش الحال" meaning "I did not like the situation" in Darija, or "مو نظيف أبداً" meaning "not clean at all" in Saudi dialect).

3.3 Data Preprocessing

We performed only tokenization. We used an Arabic tokenizer (compatible with AraELECTRA's vocabulary) to segment each review into tokens. We did not apply stemming, lemmatization, dialect normalization, or remove stopwords. The rationale was to let the AraELECTRA model and TF-IDF vector to capture the presence of any word that might carry sentiment (including shifting words, dialect words or foreign terms). While more aggressive text normalization (e.g., unifying Arabic letter variants or removing diacritics) can sometimes help, we chose to keep the text intact to preserve dialectal cues (for instance, the difference between "جميل" (beautiful in MSA) and"زوين" (beautiful in Darija) is important to maintain). The dataset was used in the given train/test split; we did not use cross-validation or external data. A small portion of the training set was held out as a validation set for early stopping and model selection, as described below.

3.4 Model

Our proposed system, AraNLP, employs a hybrid deep learning architecture designed to effectively capture both semantic and lexical features from Arabic hotel reviews. The core components of our model are a pre-trained transformer model (AraELECTRA) and TF-IDF features, which are combined and passed through a classification head.

3.4.1 AraELECTRA Embeddings

We utilize AraELECTRA (Antoun et al., 2021), a transformer-based model pre-trained on a large corpus of Arabic text. AraELECTRA is an ELECTRA-style model, which is trained as a discriminator to distinguish between original input tokens and plausible but synthetically generated replacements produced by a small generator network. This pre-training scheme has been shown to be more sample-efficient than standard masked language modeling (MLM) approaches like BERT. For each input review, we feed the tokenized text into AraELECTRA to obtain contextualized embeddings for each token. We use the embedding of the special '[CLS]' token as the aggregate representation of the review's semantics.

3.4.2 TF-IDF Features

To complement the deep contextual features from AraELECTRA, we incorporate traditional bag-of-words style features using Term Frequency-Inverse

Document Frequency (TF-IDF). We use 'TfidfVectorizer' to convert the collection of review texts into a matrix of TF-IDF features. We set the maximum number of features (i.e., dimensionality of the TF-IDF vectors) to 300. We tried different dimensions (i.e. 100 and 500) but 300 achieved the best results. These features capture the importance of different words in distinguishing between sentiment classes based on their frequency in individual documents and across the entire corpus.

3.4.3 Feature Fusion and Classification

The AraELECTRA '[CLS]' token embedding and the 300-dimensional TF-IDF vector are first processed independently. The TF-IDF vector is passed through a linear transformation layer followed by a dropout layer to project it into a space that is compatible with the transformer embeddings and to add regularization. The resulting processed TF-IDF vector is then concatenated with the AraELECTRA '[CLS]' embedding. This combined feature vector, which now contains both rich semantic information from the transformer and salient lexical information from TF-IDF, is then fed into a final linear classification layer with a softmax activation function to predict the sentiment class (Positive, Negative, or Neutral).

3.4.4 Training Setup

We trained our AraNLP model using the *AdamW optimizer* with a learning rate of 2e-5. A linear warm-up scheduler was employed for the learning rate. The loss function used was *CrossEntropyLoss*, with equal weighting for all three sentiment classes to handle potential class imbalances. We implemented an early stopping mechanism based on the validation loss. We monitored the validation loss after each epoch. If the validation loss did not improve for 3 consecutive epochs, we stopped training. In practice, our model converged within 5 epochs. We found that validation loss typically plateaued or began to increase after the 4th epoch. Early stopping helped prevent overfitting on spurious patterns in the training set. The model parameters from the epoch with the lowest validation loss were retained for final evaluation on the test set.

4 Results

Our AraNLP system was evaluated on the official test set provided by the Ahasis shared task organizers. The primary evaluation metric was the macro F1-score, which considers the F1-score for each sentiment class (Positive, Negative, Neutral) and then averages them, providing a balanced measure of performance across all classes.

As depicted in Table 1, AraNLP, achieved a macro F1-score of 76%. This performance placed our system at the 5th rank among all participating teams in the shared task. For comparison, the baseline system provided by the Ahasis organizers (referred to as "BaseLine (Ahasis)") achieved a macro F1-score of 56% and was ranked 13th. This indicates that our hybrid approach, combining AraELECTRA with TF-IDF features, provided a substantial improvement of 20 percentage points in F1-score over the baseline. The accuracy achieved by our system was also recorded, though the primary ranking was based on the macro F1-score. Detailed per-class precision, recall, and F1-scores, if provided by the organizers or obtainable from our experiment logs, would offer further insights but are summarized here by the macro F1-score.

Table 1 summarizes the key results of our system in comparison to the baseline.

5 Discussion

The performance of our AraNLP system, achieving a macro F1-score of 76% and ranking 5th in the Ahasis shared task, is encouraging. The substantial improvement over the baseline (56% F1) highlights the efficacy of our hybrid approach. The fusion of contextual embeddings from AraELECTRA with traditional TF-IDF features appears to provide a synergistic effect, capturing both deep semantic understanding and salient lexical cues. AraELECTRA, pre-trained on a vast Arabic corpus, offers robust representations of Arabic text, including dialectal variations to some extent. The TF-IDF features, on the other hand, can effectively highlight words that are strongly indicative of a particular sentiment, which might be particularly useful for domain-specific jargon or highly polar expressions not fully captured by the general pre-training of the transformer.

The challenges inherent in Arabic dialect sentiment analysis, such as the lack of standardized orthography, code-switching, and the nuanced expression of sentiment, are significant (Birjali et al., 2021). Our model's ability to perform well despite these challenges suggests that the combination of pre-trained transformers and carefully selected classical features is a promising direction. The 300-dimensional TF-IDF vector, passed through a linear

System	Macro F1-Score (%)	Rank
AraNLP (Our System)	76	5
BaseLine	56	13

Table 1: Performance comparison of AraNLP-SENT with the baseline system on the Ahasis shared task test set.

transformation and dropout, likely helped in regularizing the model and projecting these sparse features into a denser space that could be effectively combined with the transformer embeddings.

However, the error analysis reveals several limitations of our current model. We analyzed a subset of development set instances where AraNLP's prediction was incorrect, to understand the failure modes. The following examples highlight four such misclassifications, along with possible reasons:

Example 1: الفندق جميل و لكن الخدمة سيئة جدا – True label: Negative; Predicted: Positive. This is a code-mixed sentiment within one sentence: "The hotel is beautiful but the service is very bad." Our model likely picked up on the word "**جميل**" ("beautiful") as a strong positive indicator from both AraELECTRA and TF-IDF perspectives. The presence of "**سيئة جدا**" ("very bad") should denote negativity, but it appears the model either gave more weight to the positive part or failed to properly model the contrast introduced by "**لكن**" ("but"). This suggests difficulty in handling sentences with mixed sentiment. A better handling of contrastive conjunctions or a more fine-grained sentiment analysis (aspect-based) might be needed to get these correct.

Example 2: ما عجبنيش الثمن، الغرفة صغيرة بزاف – True label: Negative; Predicted: Neutral. This Moroccan Darija review translates to: "I did not like the price; the room is very small." It is clearly negative, complaining about cost and room size. The model predicted Neutral, possibly because the sentence structure is slightly complex (with negation "**ما عجبنيش**" meaning "did not please me") and multiple issues listed. AraELECTRA might have struggled with the dialect negation construct ("**ش...**" suffix) if it wasn't common in pretraining data. Also, "**بزاف**" ("very") amplifies negativity but without a direct negative word next to it, the model might not strongly connect it to negative sentiment.

Example 3: الغرفة مو بطّالة لكن التكييف مزعج طوال الليل – True label: Positive; Predicted: Negative. This Saudi dialect sentence means: "The room is not bad, but the air conditioning is noisy all night." The model predicted Negative, likely focusing on the complaint. The phrase "**مو بطّالة**" ("not bad") is faint praise. The model may have been confused by "**مزعج**" ("noisy"), without understanding that a minor complaint does not negate overall satisfaction.

Example 4: كان المتوقع أفضل بكثير – True label: Negative; Predicted: Neutral. This short review means: "The expected (experience) was much better." It implies disappointment. There is no frankly written negative word; the sentiment is implicit. This suggests a limitation in understanding nuanced or implied sentiment.

These examples illustrate that while AraNLP handles straightforward language well, it can stumble on contrast, negation, and mixed sentiments. Improvements could include contrast modeling, sentiment lexicons, or more training data, especially in dialects. Lastly, sentiment is inherently subjective (See Example 3). Some reviews are borderline. A multi-label or continuous sentiment score model might better reflect these cases in future work. A better accurate solution and better reflecting the value of of customers review is using aspect-based sentiment analysis (Pontiki et al., 2016).

While our system performed well, there is still room for improvement. The gap between our F1-score and those of the top-ranked systems suggests that further refinements could be beneficial. One area for future exploration could be more sophisticated feature fusion techniques. Instead of simple concatenation, attention mechanisms could be employed to allow the model to dynamically weigh the importance of transformer embeddings versus TF-IDF features for different inputs. Additionally, incorporating other linguistic features, such as those derived from morphological analysis of dialectal reviews, might provide further gains, as suggested by prior work like (Al-Smadi et al., 2019).

Another aspect to consider is the handling of neutral reviews. Often, neutral sentiment is harder to classify as it can encompass a wider range of expressions, including factual statements, mixed

opinions, or irrelevant content. Analyzing the per-class performance, if available, could shed light on whether our model struggled more with the neutral class compared to positive and negative classes. Tailoring specific strategies for neutral class detection or employing a hierarchical classification approach might be beneficial.

6 Conclusion

In this paper, we presented AraNLP, a hybrid deep learning system for sentiment analysis of Arabic hotel reviews, developed for the "Ahasis" shared task. Our model combines the strengths of the pre-trained AraELECTRA transformer model with classical TF-IDF features to classify sentiment in diverse Arabic dialects, specifically Saudi and Darija. The AraNLP-SENT system achieved a macro F1-score of 76%, securing the 5th rank and significantly outperforming the baseline system. This result underscores the effectiveness of integrating deep contextual embeddings with traditional lexical features for tackling the complexities of Arabic dialect sentiment analysis.

Future work will focus on exploring more advanced feature fusion techniques, incorporating dialect-specific linguistic resources, and investigating methods to better handle the nuances of neutral sentiment expressions. Further research into larger and more diverse dialectal datasets will also be crucial for advancing the field of Arabic sentiment analysis.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations. *arXiv preprint arXiv:2102.10684*.

Muhammad Abdul-Mageed, Abdelrahim Elmadany, et al. 2021. Arbert & marbert: Deep bidirectional transformers for arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105.

Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yaser Jararweh, and Omar Qawasmeh. 2019. Enhancing aspect-based sentiment analysis of arabic hotels' reviews using morphological, syntactic and semantic features. *Information Processing & Management*, 56(2):308–319.

Mohammad AL-Smadi, Mahmoud M. Hammad, Sa'ad A. Al-Zboon, Saja AL-Tawalbeh, and Erik Cambria. 2023. Gated recurrent unit with multilingual universal sentence encoder for arabic aspect-based sentiment analysis. *Knowledge-Based Systems*, 261:107540.

Mohammad Al-Smadi, Omar Qawasmeh, Mahmoud Al-Ayyoub, Yaser Jararweh, and Brij Gupta. 2018. Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. *Journal of computational science*, 27:386–393.

Maram Alharbi, Salmane Chafik, Saad Ezzini, Ruslan Mitkov, Tharindu Ranasinghe, and Hansi Hettiarachchi. 2025a. Ahasis: Shared task on sentiment analysis for arabic dialects. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

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

Salha Alyami, Areej Alhothali, and Amani Jamal. 2022. Systematic literature review of arabic aspect-based sentiment analysis. *Journal of King Saud University-Computer and Information Sciences*, 34(9):6524–6551.

Mohammed Alzahrani, Ashraf Elnagar, and James O'Shea. 2024. Arabert-lstm: Improving arabic sentiment analysis based on transformer model and long short-term memory. *Frontiers in Artificial Intelligence*, 7:1408845.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195.

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134.

Ghizlane Bourahouat, Manar Abourezq, and Najima Daoudi. 2024. Improvement of moroccan dialect sentiment analysis using arabic bert-based models. *J. Comput. Sci*, 20(2):157–167.

Hanane Elfaik and El Habib Nfaoui. 2020. Deep bidirectional lstm network learning-based sentiment analysis for arabic text. *Journal of Intelligent Systems*, 30(1):395–412.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Omar Mansour, Eman Aboelela, Remon Talaat, and Mahmoud Bustami. 2025. Transformer-based ensemble model for dialectal arabic sentiment classification. *PeerJ Computer Science*, 11:e2644.

Abubakr H Ombabi, Wael Ouarda, and Adel M Alimi. 2020. Deep learning cnn–lstm framework for arabic sentiment analysis using textual information shared in social networks. *Social Network Analysis and Mining*, 10:1–13.

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *International workshop on semantic evaluation*, pages 19–30.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained arabic dialect identification. In *Proceedings of the 27th international conference on computational linguistics*, pages 1332–1344.

Hager Saleh, Abdulaziz AlMohimeed, Rasha Hassan, Mandour M Ibrahim, Saeed Hamood Alsamhi, Moatamad Refaat Hassan, and Sherif Mostafa. 2025. Advancing arabic dialect detection with hybrid stacked transformer models. *Frontiers in Human Neuroscience*.

Bashar Talafha, Mohammad Ali, Muhy Eddin Za'ter, Haitham Seelawi, Ibraheem Tuffaha, Mostafa Samir, Wael Farhan, and Hussein Al-Natsheh. 2020. Multi-dialect arabic bert for country-level dialect identification. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 111–118.