Sentiment Analysis of Tweets During the Riyadh Season Using Deep Learning

Sara Alsahafi
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
salsahafi0070.stu@uj.edu.sa (corresponding author)

Areej Alshutayri
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
aoalshutayri@uj.edu.sa

Shahd Alahdal
Department of Computer Science and Artificial Intelligence, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
saalahdal@uj.edu.sa

Received: 29 October 2025 | Revised: 12 December 2025 and 24 December 2025 | Accepted: 29 December 2025

 DOI: https://doi.org/10.48084/etasr.15862

ABSTRACT

Social media platforms have transformed interpersonal communication, reshaping the dynamics of human interaction and information dissemination. Platform X serves as a medium for global real-time communication, enabling individuals to share thoughts and ideas. Its open accessibility makes it a valuable resource for researchers seeking domain-specific data. Saudi Arabia prioritizes the development of its tourism sector, recognizing its significance for economic diversification and national advancement, in alignment with the goals of the Vision 2030 initiative. This study evaluates attendees' sentiments about the Riyadh Season in Saudi Arabia, an entertainment event. Tweets are categorized into positive, negative, and neutral sentiments using sentiment analysis techniques, allowing a comprehensive understanding of their perceptions and emotions. We collected the dataset using the Application Programming Interface (API) of X during the Riyadh Seasons of 2020 and 2021. The dataset contains 33,300 tweets. Through rigorous comparative analysis, the study assesses the performance of deep learning models, including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, hybrid CNN–Bidirectional LSTM (CNN-BiLSTM) architectures, and BiLSTM models with attention mechanisms. Notably, the BiLSTM model with an attention layer and Arabic Bidirectional Encoder Representations from Transformers (AraBERT) word embeddings emerges as the top-performing model, achieving an accuracy rate of 83.37%. This underscores the efficacy of BiLSTM models in capturing nuanced sentiment patterns, highlighting the importance of attention mechanisms and advanced word embeddings in enhancing sentiment analysis performance. In conclusion, this study contributes to sentiment analysis by providing insights into the perceptions of the Riyadh Season in Saudi Arabia and demonstrating the effectiveness of deep learning models in analyzing sentiment-rich X data.

Keywords-deep learning; sentiment analysis; CNN; BiLSTM;attention layer; CNN-BiLSTM

I. INTRODUCTION

Social platforms generate massive amounts of data, making them a perfect domain for big data analysis. With the growing importance of social media data, sentiment analysis has become increasingly valuable as a critical tool for individuals and businesses to understand emotions and sentiments [1]. This analytical approach enables the precise exploration of areas of interest within the expansive realm of social media content. Consequently, sentiment analysis tools play an integral role in extracting valuable insights from the abundant data streams generated by these platforms.

Sentiment analysis is a field that focuses on analyzing ideas, sentiments, emotions, and behavioral patterns to classify documents, sentences, or aspects as either positive, negative, or neutral. Emotion analysis, a subset of sentiment analysis, delves deeper into users' emotional states and psychological responses, providing enhanced insights. Both sentiment and emotion analysis represent applications of Natural Language

Processing (NLP), which studies human languages from a computational standpoint.

In the field of Arabic sentiment analysis, researchers face several challenges. These were identified by authors in [2] and summarized here as follows: Arabic is a complex morphological language, making it challenging to identify the root of words due to multiple surface forms and various forms of affixes, clitics, and diacritics. Arabic speakers often use compound idioms and phrases to express emotions, making it difficult for sentiment systems to detect implicit sentiments if the lexicon or experimental dataset does not include such terms. For instance, the idiom 'يا الشيخ' (used to express disagreement or disbelief regarding someone's statement) and the phrase 'ما □ ء الله' (used to express admiration or praise) illustrate how Arabic expressions convey implicit sentiment. Since the article addresses a global readership, providing English explanations for such expressions ensures clarity and accessibility. Scam Detection (SCD) pertains to situations where a person expresses a positive opinion but means the opposite, or vice versa [2]. Addressing these challenges requires ongoing research and innovation in the field of sentiment analysis, with a focus on developing robust algorithms and methodologies that account for the intricacies of the Arabic language and its diverse linguistic expressions.

This research paper employs deep learning algorithms implemented in Python and Google Colab to analyze Arabic tweet data to detect emotions expressed during the Riyadh Season. We experimented with multiple models, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks. Then, we evaluated a CNN–Bidirectional Long Short-Term Memory (CNN-BiLSTM) hybrid. We also applied a BiLSTM model with an attention layer. This study aims to examine public sentiment toward the Riyadh Season through deep learning-based analysis of Arabic tweets, offering valuable insights into understanding audience perceptions.

II. LITERATURE REVIEW

This section synthesizes previous studies on Arabic sentiment and emotion analysis under three analytical themes:

1. Challenges in Arabic NLP and dialectal diversity,
2. Comparative performance of deep learning architectures, and
3. Transformer-based advancements in Arabic sentiment analysis.

The review integrates the main findings with corresponding research gaps to clarify methodological limitations and position the current study accordingly.

A. Arabic Natural Language Processing and Dialect Challenges

Although sentiment analysis and emotion extraction have been extensively explored in English and other European languages, research on Arabic remains comparatively limited. The complexity of Arabic morphology, rich inflection, shallow discretization, and wide dialectal variation introduce modeling challenges that mainstream NLP methods often struggle to handle. Several studies [3-6] have highlighted these obstacles, noting the need for task-specific preprocessing, stemming choices, and normalization strategies to mitigate linguistic noise.

Earlier works relied on classical deep learning architectures—LSTM, CNN, and hybrid CNN-LSTM models—to process Arabic tweets and multi-class emotion labels including anger, fear, sadness, joy, and surprise. However, these models typically required extensive preprocessing to compensate for their limited ability to model long-range context. For instance, the study in [4] emphasized that Arabic stop-word removal significantly affected CNN performance, and the Sentiment and Emotion Detection in Arabic Text (SEDAT) [5] improved results only after intensive preprocessing of raw Arabic tweets and their English translations.

Research on Arabic dialects further illustrated the difficulty of modeling informal, user-generated content [7]. The work in Arabic Online Commentary (AOC)–based dialect classification demonstrated that even with LSTM and BiLSTM architectures, accuracies plateaued around ~71%, largely due to dialect variation and insufficient normalization. These findings collectively show that traditional models lack a robust mechanism to encode the linguistic richness and dialectal diversity of Arabic, indicating a need for architectures that inherently incorporate contextual and sub-word information. While most works describe these linguistic challenges, few propose systematic frameworks to handle sarcasm, idioms, Arabizi, or dialectal shifts. Moreover, domain-specific dialects—such as the Saudi dialect used in entertainment contexts—remain underexplored within sentiment analysis research. Although this gap has been widely recognized, addressing the full complexity of dialectal variation and idiomatic expression lies beyond the scope of the present study. Instead, this research focuses on evaluating model performance within a Saudi dialect context, thereby providing partial insights into this broader challenge and laying groundwork for future exploration.

B. Comparison of Deep Learning Models for Arabic Sentiment and Emotion Analysis

Deep learning architectures have played a pivotal role in advancing Arabic sentiment analysis, yet their performance varies widely depending on datasets, preprocessing strategies, and embedding choices.

Recurrent architectures (Recurrent Neural Networks (RNNs), LSTM, and BiLSTM) have demonstrated strong sequential modeling capacity. Recent studies applied LSTM for multi-label emotion classification and achieved 61.7% accuracy [3, 8], whereas LSTM variants used during the COVID-19 pandemic reached up to 83% [9]. Convolutional architectures such as CNNs have also yielded promising results—other research achieved 99.82% accuracy on the SemEval-2018 dataset, underscoring the critical influence of preprocessing steps such as stop-word removal [4].

Hybrid architectures integrating CNN and LSTM components—such as the SEDAT model [5]—have shown

accuracy above 80%, combining local feature extraction with temporal sequence modeling. Similarly, Bidirectional Gated Recurrent Unit (BiGRU)-based systems reached 92.96% accuracy on the ATDFS dataset [10], whereas differential evolution was employed to optimize CNN hyperparameters, obtaining an accuracy of 86.64% [11].

Despite these promising results, several consistent limitations emerge. Most studies rely on single datasets, hindering generalizability. Deep learning models also demand extensive preprocessing to compensate for their limited ability to capture long range dependencies and context. Their reliance on word- or character-level embeddings restricts semantic depth—particularly problematic for morphologically rich and dialectal Arabic.

Consequently, while CNN, LSTM, and BiLSTM architectures provide strong baselines, they may not fully exploit the linguistic depth of Arabic. These constraints reinforce the need for contextualized models, such as Arabic Bidirectional Encoder Representations from Transformers (AraBERT) and multilingual Bidirectional Encoder Representations from Transformers (BERT), which integrate self-attention and sub-word modeling to better handle dialectal nuances.

C. Transformer-Based Models in Arabic Sentiment Analysis

Transformer-based architectures have revolutionized Arabic NLP by offering greater contextual understanding and scalability [12]. A hyper-tuned BERT model achieved a 99% accuracy, surpassing BiLSTM (98%), LSTM, GRU, and CNN models [13]. Similar findings in Facebook- and X-based analyses confirm that Transformer architectures consistently outperform recurrent and convolutional baselines.

The theoretical foundations for this superiority include:

- Self-attention mechanisms enable the model to capture dependencies between all tokens simultaneously, which is essential for understanding Arabic syntax and long-range relationships.
- Large-scale pretraining models such as AraBERT to leverage massive Arabic corpora, embedding rich syntactic and semantic knowledge absent in task-specific models.

These properties collectively empower Transformers to generalize across dialects and noisy social media text with minimal preprocessing. Yet, notable research gaps persist: few studies systematically compare Transformer variants (e.g., AraBERT vs. MARBERT) or examine their robustness across multiple dialectal datasets. Additionally, many works assume Transformer superiority without empirically analyzing why these models perform better, particularly in handling idioms, sarcasm, and contextual ambiguity within Saudi dialect tweets.

The current study addresses these limitations by conducting a controlled comparative evaluation of multiple deep learning architectures (CNN, CNN-LSTM, LSTM, and BiLSTM) alongside Transformer- and embedding-based models (AraBERT and FastText) using Saudi dialect tweets. The analysis examines both quantitative performance metrics and qualitative linguistic factors to interpret variations in model behavior.

D. Literature Gap and Study Contribution

Synthesizing across these analytical dimensions, several research gaps are evident:

- Dialectal and linguistic complexities are well-documented but remain methodologically unresolved.
- Deep learning studies lack analytical comparisons across multiple datasets and model types, limiting reproducibility and generalizability.
- Transformer superiority is reported but rarely theorized or empirically explained, especially in dialectal contexts.
- Saudi dialect sentiment—particularly in entertainment-related tweets—remains underexplored, despite its social and commercial significance.

To bridge these gaps, the present study provides a comparative, dialect-specific evaluation of multiple deep learning architectures (CNN, CNN-LSTM, LSTM, and BiLSTM) and Transformers architectures on Saudi social media data, offering both performance benchmarks and interpretive insights into their linguistic and contextual strengths.

III. METHODOLOGY

This section briefly describes the methodology used in this research. X data were collected using the snscrape library in Python to analyze the overall study. Deep learning methods were used to train the models. After the models were trained, a comprehensive evaluation was conducted, and the results were compared. The overall methodology for this study is shown in Figure 1.

A. Data Collection

This section provides a detailed description of our data collection process and data annotation. The data collection process aims to create a dataset for the entertainment event-making industry, which researchers can use for Arabic sentiment analysis and comparison. Collecting data can be challenging due to the limitations of the Application Programming Interface (API) of X, which only allows users to collect data for up to 7 days and has a limited number of tweets that can be gathered.

We collected data between October 2021 and January 2023, targeting the most common hashtags related to the Riyadh Season and official accounts to ensure comprehensive coverage of the season's events and their associated hashtags. The hashtags used to collect data are listed in Table I.

1) Dataset Availability

The final dataset, comprising 81,175 manually annotated tweets collected from the Riyadh Season events held in 2020 and 2021, is available upon request from the corresponding author. To protect user privacy and adhere to ethical research standards, all user identifiers, including usernames and profile information, have been removed.

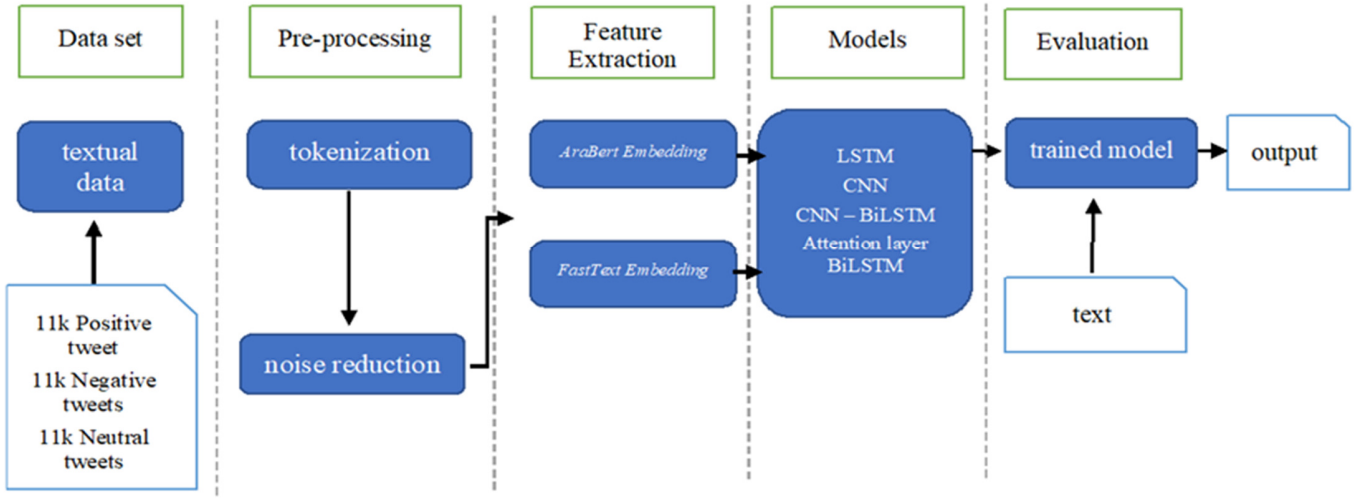

Fig. 1. The proposed methodology.

TABLE I. TWEETS DATA FOR RIYADH SEASON HASHTAGS

No.	Hashtags	Number of tweets
1	#RiyadhSeason	10,002
2	#BlvdRuhCity	5,009
3	#boulevardRiyadhCity	684
4	#FormulaE	95,845
5	#JoyAwards	766
6	#MDLBEAST	30,447
7	#Thegroves	13,803
8	#WinterWonderland	18,214
9	#تخيل_أكثر	100,002
10	#وليفارد_الرياض_سيتي	20,521
11	#فورميلا_اي_الدرعية	12,447
12	#ليلة_السندباد	16
13	#مدل_بيست	204
14	#مسابقات_موسم_الرياض	63,603
15	#وينتر_ويندر_لاند	529
Total tweets		372,092

B. Data Preprocessing and Cleaning

After collecting the tweets, which totaled 372,092 entries, we manually cleaned the dataset by removing advertisements and spam. Following this cleaning process, the number of remaining tweets was reduced to 81,175. An additional 66,166 tweets from the Riyadh Season 2020 dataset [14] were incorporated. Finally, the dataset was filtered to include only Arabic tweets (both Modern Standard Arabic and dialectal Arabic). This ensured that the corpus reflected the linguistic characteristics of real-world user-generated content relevant to the Riyadh Season.

Following the collection, tweets were labeled into positive, negative, and neutral categories. Class imbalance was observed, with positive tweets being underrepresented. To address this, random undersampling was applied to negative and neutral tweets to match the count of positive ones, resulting in a balanced dataset of 33,300 tweets (11,100 per class).

Data cleaning involved several steps, including handling Arabic stop words and removing tags, mentions, URLs, and usernames. Although we initially used an existing Python library for Arabic stop words, we found that it was overly formal for our dataset. Therefore, we added additional stop words commonly used in the Saudi dialect, as many of these terms frequently appeared in the tweets. To ensure comprehensive coverage, we supplemented the list with dialect-specific stop words as well as additional terms sourced from a publicly available GitHub repository [15]. Table II presents examples of added stop words.

TABLE II. EXAMPLES OF ADDED STOP WORDS

Stop words
امس،السابق ,التي ,التي ,اكثر , أيضا, ثلاثة, الذاتي ,الأخيرة, الثاني ,الثانية, الذى ,الذي ,الان, امام

Overall, our data cleaning process involved the following steps:

1. Removing tags, mentions, URL links, and usernames.
2. Removing symbols and special characters.
3. Removing duplicate letters in words, such as "كبيبير" to become "كبير."
4. Removing emojis, as studies such as [16, 17] have demonstrated that emojis can introduce noise and ambiguity due to inconsistent or ironic usage, whereas [9] specifically highlighted that emojis in Arabic texts can conflict with textual sentiment, leading to misclassification and reduced model performance.
5. Removing Arabic stop words.

Following these steps, as shown in Figure 2, the data were cleaned effectively and prepared for further analysis. Table III shows an example of the tweet cleaning process.

Fig. 2. Data cleaning process.

TABLE III. EXAMPLE OF THE TWEET CLEANING PROCESS

Original tweet	"اليوم جربت بعض الألعاب في الرياض ونتر لاند ... الألعاب الموجودة□ي خورااافي ، ٣ ايّام ويالله تكفيك 😍 ❤️ 🔥 ... https://t.co/DKcypepueB"
Remove hashtags and URL	"اليوم جربت بعض الألعاب في الرياض ونتر لاند ... الألعاب الموجودة شي خورااافي ، ٣ ايّام ويالله تكفيك 😍 ❤️ 🔥 ... "
Remove symbols and special characters	"اليوم جربت بعض الألعاب في الرياض ونتر لاند الألعاب الموجودة شي خورااافي ٣ ايّام ويالله تكفيك 😍 ❤️ 🔥 "
Remove duplicate letters	"اليوم جربت بعض الألعاب في الرياض ونتر لاند الألعاب الموجودة شي خورافي ٣ ايّام ويالله تكفيك 😍 ❤️ 🔥 "
Remove emoji	"اليوم جربت بعض الألعاب في الرياض ونتر لاند الألعاب الموجودة شي خورافي ٣ ايّام ويالله تكفيك"
Remove stop words	"جربت الألعاب الرياض ونترلاند الموجودة خورافي تكفيك"
Cleaned text	"جربت الألعاب الرياض ونترلاند الموجودة خورافي تكفيك"

C. Data Labeling

The data labeling phase is essential and affects the results; it differs based on people's perspectives. Therefore, three independent Arabic-speaking annotators were employed to label each of the 81,175 tweets into one of three categories: positive, negative, or neutral. The labeling process was designed to treat all annotators equally, with the final label determined by majority voting. Specifically, if at least two annotators agreed on a label, the tweet was assigned to that label. In cases where all three annotators disagreed (i.e., each assigned a different label), the tweet was labeled as neutral. This approach ensures a balanced and fair annotation process while minimizing ambiguity.

To assess the consistency and reliability of the annotations, the inter-annotator agreement was analyzed using Fleiss' kappa, a statistical measure suitable for evaluating agreement among multiple annotators. This approach ensures the robustness and credibility of the labeled dataset. The kappa coefficient, denoted as κ, is calculated as:

$$\kappa = \frac{\bar{P} - \overline{P_e}}{1 - \bar{P}_e} \quad (1)$$

The details of the annotation process are summarized in Table IV, which shows the distribution of labels based on majority agreement and complete disagreement among the annotators. Out of the total tweets, 76,875 (94.7%) had majority agreement, whereas 4,300 (5.3%) resulted in complete disagreement and were labeled as neutral. This reflects the inherent subjectivity in sentiment analysis.

The Fleiss' Kappa value was calculated as shown in Table V, resulting in $\kappa = 0.72$, which indicates substantial agreement among the annotators.

TABLE IV. SUMMARY OF AGREEMENT AND DISAGREEMENT

Agreement type	**Positive**	**Negative**	**Neutral**	**Total**
Majority agreement	49,832	23,332	3,711	76,875
Complete disagreement	-	-	4,300	4,300
Total	49,832	23,332	8,011	81,175

TABLE V. FLEISS KAPPA CALCULATION

Parameter	**Value**
Total tweets (N)	81,175
Number of annotators (n)	3
Number of categories (k)	3
Sum of squared agreements	340,935
Observed Agreement (P_o)	$P_o = \frac{340{,}935}{81{,}175 \times 3 \times 2} \approx 0.70$
Expected Agreement (P_e)	$P_e = (0.614)^2 + (0.287)^2 + (0.099)^2 \approx 0.469$
Fleiss' Kappa (κ)	$\kappa = \frac{0.70 - 0.469}{1 - 0.469} \approx 0.72$
Interpretation	Substantial agreement (κ = 0.72)

D. Feature Extraction

This section discusses two advanced word embedding techniques for feature extraction in Arabic sentiment analysis: AraBERT and FastText. Both approaches convert textual data into continuous vector representations, enabling deep learning models to interpret linguistic information effectively.

1) AraBert Embeddings

AraBert is a state-of-the-art Transformer-based language model that provides high-quality word embeddings. We used pretrained AraBert embeddings to convert textual data into continuous vector representations. These embeddings capture rich semantic information, making them suitable for sentiment analysis [18].

2) FastText Embeddings

FastText is an efficient word embedding technique that can handle out-of-vocabulary words. We employed FastText embeddings to compare their performance with AraBert embeddings. FastText embeddings are trained on sub-word information, which can benefit morphologically rich languages like Arabic [6].

E. Models

This section describes the architecture of different neural network models used for sentiment analysis, including LSTM, CNN, CNN-BiLSTM, and attention-based BiLSTM models:

1. LSTM model: The LSTM model employs LSTM, a type of RNN. Its architecture comprises an embedding layer, a bidirectional LSTM layer, a fully connected (dense) layer, and a final dense layer for sentiment classification. This design mainly captures long-term dependencies within sequential data, such as text [6].
2. CNN model: The CNN model implements a CNN for sentiment prediction. Its architecture encompasses an embedding layer, Conv1D layer(s), a MaxPooling1D layer, a flatten layer, a dense layer with 32 hidden nodes, and a final dense layer for sentiment classification. CNNs are recognized for their proficiency in extracting pertinent features crucial for practical classification tasks [19].
3. CNN-BiLSTM model: The CNN-BiLSTM model combines the advantages of CNNs and BiLSTM. Its architecture includes Conv1D, MaxPooling1D, BiLSTM, and a final dense layer. This integration enables CNN-based feature extraction whereas BiLSTM captures contextual information in both forward and backward directions, yielding a robust sentiment analysis model [20].
4. Attention BiLSTM model: The attention BiLSTM model integrates attention mechanisms to concentrate on significant segments of the input sequence. Its architecture encompasses an embedding layer, a BiLSTM layer, an attention mechanism layer, a dense layer for feature extraction, and a final dense layer for sentiment classification. This design enhances the model's ability to discern essential patterns in sequential data, making it well-suited for sentiment analysis tasks [21].

In sentiment analysis, LSTM and BiLSTM models are commonly used for analyzing text sentiment based on labeled data. The attention BiLSTM model enhances this by incorporating attention mechanisms to better focus on the input sequence's relevant parts.

F. Experimental Setup

1) Training and Validation

The dataset was randomly partitioned into a training set (80%), a validation set (10%), and a testing set (10%) for model training, validation, and evaluation.

2) Hyperparameter Configuration

The models were trained with the following hyperparameters:

- Learning rate: 0.001 for LSTM, attention LSTM, and CNN models.
- Batch size: 128 for all experiments.
- Epochs: 100 for all models.
- Optimizer: Adam was used to optimize the loss function.

3) Experimental Procedure

To investigate the effect of stop-word removal on sentiment analysis accuracy, experiments were conducted both with and without the removal of common Arabic stop words. These experiments were performed using AraBERT and FastText embeddings.

4) Evaluation Metrics

The performance of the proposed model was assessed using four standard evaluation metrics: accuracy, precision, recall, and F1-score. These metrics are widely used in classification tasks to measure the model's ability to detect and differentiate between classes correctly. The formulas for each metric are as follows:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \quad (2)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (3)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \quad (4)$$

$$\text{F1} - \text{score} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 \quad (5)$$

where TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative) represent the classification outcomes of the model. Accuracy reflects the overall correctness of the model, precision measures the proportion of correctly predicted positive samples, recall evaluates the model's ability to identify all actual positives, and the F1-score provides a balanced harmonic mean of precision and recall.

These metrics were subsequently used to evaluate and compare the performance of the proposed models, as presented in the results section.

IV. RESULTS

Sentiment analysis is a widely used method for analyzing the emotional tone of texts, such as social media posts or product reviews. In this study, we analyzed the sentiment on a dataset of 33,300 Arabic tweets from Riyadh Season 2020, 2021, and 2022 using four deep learning models: CNN, LSTM, CNN-BiLSTM, and attention BiLSTM. These models were selected based on their reported performance in previous studies [3, 5, 22].

Tables VI and VII present the performance of all models without removing stop words, employing two types of embedding, AraBERT and FastText. With AraBERT embeddings, the CNN achieved 82.23% accuracy, LSTM 82.84%, CNN-BiLSTM 79.59%, and attention BiLSTM 83.37%. Using FastText embeddings, the CNN achieved 80.10%, LSTM 80.80%, CNN-BiLSTM 79.72%, and attention BiLSTM 80.91%.

Table VIII and IX show model performance after removing stop words. Using AraBERT embeddings, the CNN achieved 81.75% accuracy, LSTM 82.29%, CNN-BiLSTM 78.79%, and attention BiLSTM 82.38%. With FastText embeddings, the CNN achieved 79.60%, LSTM 80.19%, CNN-BiLSTM 80.16%, and attention BiLSTM 80.28%.

TABLE VI. ARABERT WORD EMBEDDINGS WITHOUT STOP WORD REMOVAL

Model	Accuracy	Precision	Recall	F1-score
CNN	82.23	81.95	82.23	82.04
BiLSTM	82.84	82.76	82.84	82.65
CNN-BiLSTM	79.59	79.49	79.58	79.50
BiLSTM + attention	83.37	83.23	83.37	83.27

TABLE VII. FASTTEXT WORD EMBEDDINGS WITHOUT STOP WORD REMOVAL

Model	Accuracy	Precision	Recall	F1-score
CNN	80.10	80.19	80.09	79.90
BiLSTM	80.80	80.56	80.80	80.66
CNN-BiLSTM	79.72	79.71	79.72	79.57
BiLSTM + attention	80.91	80.68	80.90	80.74

TABLE VIII. ARABERT WORD EMBEDDINGS WITH STOP WORD REMOVAL

Model	Accuracy	Precision	Recall	F1-score
CNN	81.75	81.62	81.71	81.58
BiLSTM	82.29	82.09	82.28	82.11
CNN-BiLSTM	78.79	78.61	78.79	78.69
BiLSTM + attention	82.38	82.19	82.37	82.20

TABLE IX. FASTTEXT WORD EMBEDDINGS WITH STOP WORD REMOVAL

Model	Accuracy	Precision	Recall	F1-score
CNN	79.60	80.06	79.60	79.34
BiLSTM	80.19	80.03	80.18	79.97
CNN-BiLSTM	80.16	80.17	80.15	80.02
BiLSTM + attention	80.28	80.35	80.27	80.00

Across all models, AraBERT embeddings generally yielded slightly higher accuracy rates compared with FastText embeddings, and the attention BiLSTM model achieved the highest accuracy.

A. Influences on Sentiment Analysis Accuracy: Feature Extraction, Stop Word Removal, and Model Architectures

The results indicate that feature extraction significantly impacts model performance. AraBert embeddings consistently outperformed FastText embeddings, likely due to their ability to capture deeper semantic meaning [23].

Interestingly, removing stop words slightly decreased the accuracy across all models, which is not uncommon according to other research [3]. In contrast, this practice resulted in a notable enhancement in accuracy when applied to the SemEval-2018 dataset. However, in our specific case, it is essential to recognize that stop words may contain vital contextual information, and their removal could potentially lead to a loss of accuracy. This discrepancy in outcomes prompts a closer examination. It is worth noting that the efficacy of stop word removal in the other study might be attributed to their extensive preprocessing methodology, which comprised approximately 22 steps, extending beyond mere word removal. This suggests that the decision to retain or remove stop words should be contingent upon the requirements of the task and the nuances of the text data under analysis.

Moreover, we observed that the attention mechanism improved LSTM performance for both embeddings. BiLSTM produced more meaningful output because it combines LSTM layers in both forward and backward directions, allowing each word in the input sequence to incorporate contextual information from both directions.

Figures 3 and 4 show the accuracy of all models with and without stop-word removal. Figures 5 and 6 show that BiLSTM with FastText embeddings predicts positive tweets accurately without removing stop words. However, it tends to confuse negative and neutral tweets.

B. Challenges Related to the Dataset: Insights from the Riyadh Season Tweets

The analyzed Arabic tweets from our Riyadh Season dataset present several linguistic and contextual challenges that likely contributed to the observed decrease in classification accuracy across various models. Two primary factors underscore these challenges.

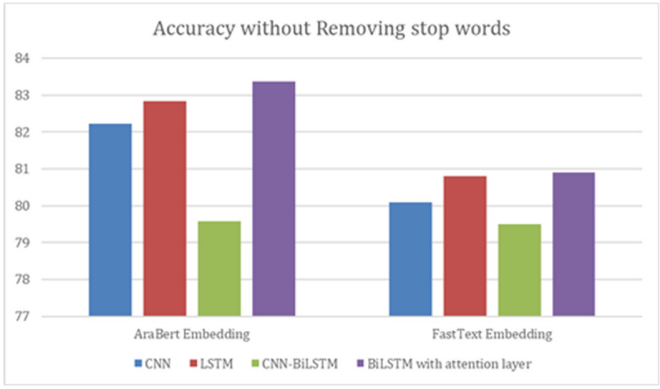

Fig. 3. Accuracy of all models using AraBERT and FastText embeddings without stop-word removal.

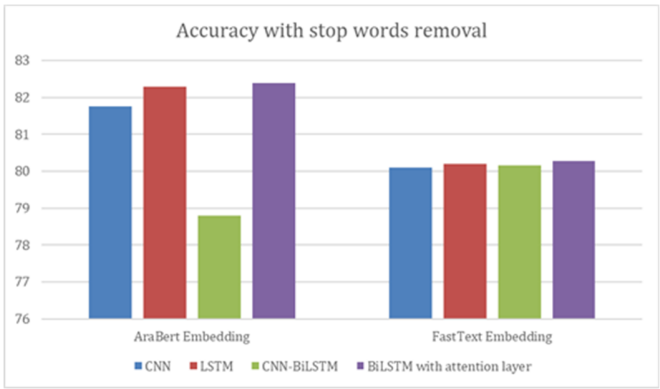

Fig. 4. Accuracy of all models using AraBERT and FastText embeddings with stop-word removal.

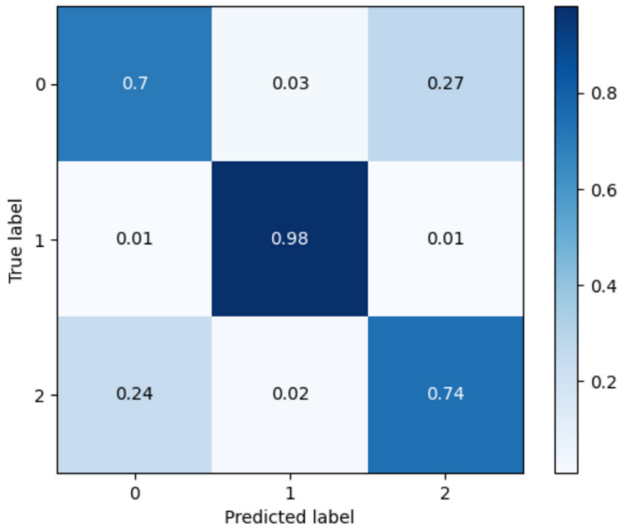

Fig. 5. Confusion matrix of the BiLSTM model using FastText embeddings with stop-word removal.

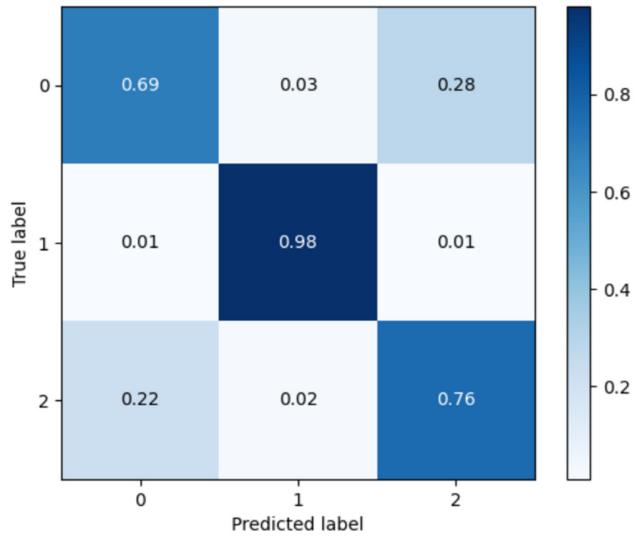

Fig. 6. Confusion matrix of the BiLSTM model using FastText embeddings without stop-word removal.

Firstly, using local Arabic dialect in tweets, mainly Saudi, introduces complexities related to sarcasm and exaggerated descriptions. Sarcasm entails employing positive language to convey negative meanings, a linguistic phenomenon common across diverse cultures, including Saudi culture. Furthermore, using exaggerated descriptions involves expressing positive sentiments using negative words, essentially the opposite of sarcasm. The contradictory nature of these social phenomena makes it challenging for models to accurately capture and classify sentiment. Examples illustrating these challenges are provided in Table X.

Secondly, the accuracy in tweet identification is also influenced by the nature of the event. Users frequently quote song lyrics, including somber ones performed live on stage. Although this denotes engagement and an overall positive reaction, the choice of words may mislead models into categorizing such tweets as negative.

TABLE X. DATASET CHALLENGES

Tweet	Label	Context
يلا حلها يا مدل بيست يا رقص يا فله	1	Sarcasm: Positive words are used to express anger
ليلة_سهم # حزبيينه ليش اللستة صارت حلوه فجاءه كذا؟!!	-1	Exaggerated descriptions: Negative words are used to express positive emotion
كل شخص حاظر أحسد #ليلة_سهم	-1	
" خايف اتمنى لقاك 💔 💔 💔 ؟ #رايح صقر ليلة سهم "	-1	Quotations: Users write song lyrics to evoke excitement and emotion

V. CONCLUSION

This study presented a novel dataset of locally generated Arabic content collected from Riyadh Season events in 2020 and 2021, thereby contributing to the limited resources available for sentiment analysis in Saudi dialects. The dataset demonstrates several linguistic challenges, including sarcasm, irony, exaggerated expressions, and the frequent use of song lyrics, all of which complicate accurate sentiment interpretation. These characteristics highlight the need for improved processing of entertainment-related Arabic text, particularly as Saudi Arabia continues to expand its entertainment sector under Vision 2030.

Comprehensive experiments were conducted using four deep-learning architectures—Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, hybrid CNN–Bidirectional LSTM (CNN-BiLSTM), and attention BiLSTM—along with two feature-extraction techniques, FastText and Arabic Bidirectional Encoder Representations from Transformers (AraBERT). The attention BiLSTM model combined with AraBERT embeddings achieved the highest accuracy of 83.37%. These results are consistent with recent literature reporting the superior performance of Transformer-based representations in Arabic sentiment analysis. The proposed system also demonstrated strong capability in detecting emotions within Saudi dialect tweets, with notable effectiveness in identifying positive sentiment.

Several limitations were identified. The presence of sarcasm, irony, and region-specific expressions introduces ambiguity that may reduce classification accuracy. Dialectal diversity across Saudi regions and potential annotation bias also present challenges that may affect the generalizability of the results. Additionally, the domain-specific nature of entertainment-related content may limit its broader applicability.

The findings offer practical implications for event management, tourism, and social media analytics. More accurate sentiment and emotion detection can assist stakeholders in assessing audience engagement, understanding public reactions, and improving decision-making processes for large-scale events. Future work will focus on expanding the dataset to include additional dialects and on evaluating

advanced Transformer-based architectures to further enhance predictive performance and robustness. Such efforts are expected to provide deeper insight into sentiment dynamics within Saudi Arabia's evolving digital and entertainment landscape.

REFERENCES

[1] Q. A. Xu, V. Chang, and C. Jayne, "A systematic review of social media-based sentiment analysis: Emerging trends and challenges," *Decision Analytics Journal*, vol. 3, June 2022, Art. no. 100073, https://doi.org/10.1016/j.dajour.2022.100073.

[2] M. Zayed, H. Mousa, and M. Elmenshawy, "Sentiment Analysis for Arabic Social Media," *IJCI. International Journal of Computers and Information*, vol. 7, no. 1, pp. 14–31, Oct. 2020, https://doi.org/10.21608/ijci.2020.16170.1004.

[3] F. Abdullah, M. Al-Ayyoub, I. Hmeidi, and N. Alhindaw, "A Deep Learning Approach to Classify and Quantify the Multiple Emotions of Arabic Tweets," in *2021 12th International Conference on Information and Communication Systems*, Valencia, Spain, 2021, pp. 399–404, https://doi.org/10.1109/ICICS52457.2021.9464548.

[4] M. Baali and N. Ghneim, "Emotion analysis of Arabic tweets using deep learning approach," *Journal of Big Data*, vol. 6, no. 1, Oct. 2019, Art. no. 89, https://doi.org/10.1186/s40537-019-0252-x.

[5] M. Abdullah, M. Hadzikadicy, and S. Shaikhz, "SEDAT: Sentiment and Emotion Detection in Arabic Text Using CNN-LSTM Deep Learning," in *2018 17th IEEE International Conference on Machine Learning and Applications*, Orlando, FL, USA, 2018, pp. 835–840, https://doi.org/10.1109/ICMLA.2018.00134.

[6] M. Heikal, M. Torki, and N. El-Makky, "Sentiment Analysis of Arabic Tweets using Deep Learning," *Procedia Computer Science*, vol. 142, pp. 114–122, Jan. 2018, https://doi.org/10.1016/j.procs.2018.10.466.

[7] L. Lulu and A. Elnagar, "Automatic Arabic Dialect Classification Using Deep Learning Models," *Procedia Computer Science*, vol. 142, pp. 262–269, Jan. 2018, https://doi.org/10.1016/j.procs.2018.10.489.

[8] S. Alzu'bi, O. Badarneh, B. Hawashin, M. Al-Ayyoub, N. Alhindawi, and Y. Jararweh, "Multi-Label Emotion Classification for Arabic Tweets," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security*, Granada, Spain, 2019, pp. 499–504, https://doi.org/10.1109/SNAMS.2019.8931715.

[9] A. Al-Laith and M. Alenezi, "Monitoring People's Emotions and Symptoms from Arabic Tweets during the COVID-19 Pandemic," *Information*, vol. 12, no. 2, Feb. 2021, Art. no. 86, https://doi.org/10.3390/info12020086.

[10] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, and E. A. Retta, "Improving Arabic Sentiment Analysis Using CNN-Based Architectures and Text Preprocessing," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, Sept. 2021, Art. no. 5538791, https://doi.org/10.1155/2021/5538791.

[11] A. Dahou, M. A. Elaziz, J. Zhou, and S. Xiong, "Arabic Sentiment Classification Using Convolutional Neural Network and Differential Evolution Algorithm," *Computational Intelligence and Neuroscience*, vol. 2019, no. 1, Feb. 2019, Art. no. 2537689, https://doi.org/10.1155/2019/2537689.

[12] A. Ouza, A. Ouacha, A. Rachidi, M. El Ghmary, and A. Choukri, "Enhancing Arabic Sentiment Analysis Using AraBERT and Deep Learning Models," in *Modern Artificial Intelligence and Data Science 2024: Tools, Techniques and Systems*, A. Idrissi, Ed. Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 189–200, https://doi.org/10.1007/978-3-031-65038-3_15.

[13] N. Sureja, N. Chaudhari, P. Patel, J. Bhatt, T. Desai, and V. Parikh, "Hyper-tuned Swarm Intelligence Machine Learning-based Sentiment Analysis of Social Media," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15415–15421, Aug. 2024, https://doi.org/10.48084/etasr.7818.

[14] W. Bajaber, "Sentiment Analysis for Entertainment Events," M.S. thesis, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia, 2018, https://doi.org/10.13140/RG.2.2.16825.31848.

[15] H. M. Hosny. "ArabicDialectClassification/utils.py at main hazemhosny/ArabicDialectClassification." GitHub. https://github.com/hazemhosny/ArabicDialectClassification/blob/main/utils.py.

[16] C. Liu *et al.*, "Improving sentiment analysis accuracy with emoji embedding," *Journal of Safety Science and Resilience*, vol. 2, no. 4, pp. 246–252, Dec. 2021, https://doi.org/10.1016/j.jnlssr.2021.10.003.

[17] P. K. Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of Emojis," *Plos One*, vol. 10, no. 12, Dec. 2015, Art. no. e0144296, https://doi.org/10.1371/journal.pone.0144296.

[18] A. Abo-Elghit, T. Hamza, and A. Al-Zoghby, "Embedding Extraction for Arabic Text Using the AraBERT Model," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1967–1994, Feb. 2022, https://doi.org/10.32604/cmc.2022.025353.

[19] P. Huang, L. Zheng, Y. Wang, and H. J. Zhu, "Sentiment Analysis of Chinese Text Based on CNN-BiLSTM Serial Hybrid Model," in *Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition*, Shanghai, China, 2022, pp. 309–313, https://doi.org/10.1145/3497623.3497673.

[20] W. Wei, Y. Zhang, R. Duan, and W. Zhang, "Microblog Sentiment Classification Method Based on Dual Attention Mechanism and Bidirectional LSTM," in *Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers*, Beijing, China, 2019, pp. 309–320, https://doi.org/10.1007/978-3-030-38189-9_33.

[21] S. A. A. Hakami, R. Hendley, and P. Smith, "Emoji as Sentiment Indicators: An Investigative Case Study in Arabic Text," in *HUSO 2020 : The Sixth International Conference on Human and Social Analytics*, Porto, Portugal, 2020, pp. 26–32.

[22] T. A. Al-Qablan, M. H. Mohd Noor, M. A. Al-Betar, and A. T. Khader, "A survey on sentiment analysis and its applications," *Neural Computing and Applications*, vol. 35, no. 29, pp. 21567–21601, Oct. 2023, https://doi.org/10.1007/s00521-023-08941-y.

[23] A. Masmoudi, J. Hamdi, and L. Hadrich Belguith, "Deep Learning for Sentiment Analysis of Tunisian Dialect," *Computación y Sistemas*, vol. 25, no. 1, pp. 129–148, Feb. 2021, https://doi.org/10.13053/cys-25-1-3472.