Journal of Computer Science

Research Article

BERT-Based Temporal Sentiment Analysis of Saudi Tourism Events

Ahmed Alharbi

Department of Computer Science, Taibah University, Medina, Saudi Arabia

Article history
Received: 26-03-2026
Revised: 30-05-2026
Accepted: 06-08-2026

Email: atharbi@taibahu.edu.sa

Abstract: This study applies a BERT-based temporal sentiment analysis framework to examine English-language social media discourse surrounding major tourism and sporting events hosted in Saudi Arabia. A dataset of 10,000 tweets collected between December 2023 and December 2024, including discussions related to the FIFA Club World Cup, Formula 1 Saudi Arabian Grand Prix, and other major Saudi-hosted sporting events, was analyzed to investigate sentiment dynamics across pre-event, during-event, and post-event stages. The fine-tuned BERT model achieved an F1-score of 0.85, outperforming traditional machine learning and deep learning baselines. Results reveal strong positive sentiment during event periods, with measurable variations across different temporal stages. The findings demonstrate the value of transformer-based sentiment analysis for understanding international audience engagement with Saudi mega-events and provide empirical insights relevant to tourism development and Vision 2030 objectives.

Keywords: BERT, Sentiment Analysis, Social Media, Tourism, Temporal Analysis

Introduction

Tourism is increasingly recognized as a strategic sector in Saudi Arabia's Vision 2030 plan, which seeks to diversify the economy, reduce oil dependence, and enhance the country's international image (Saudi Arabia Vision 2030, 2016). To support these objectives, the Kingdom has invested heavily in hosting high-profile international events such as the FIFA Club World Cup, Formula 1 Saudi Arabian Grand Prix, Supercoppa Italiana, Spanish Super Cup, and major boxing matches designed to attract large numbers of foreign visitors and strengthen Saudi Arabia's position as a global tourism destination. These efforts have already enabled the country to surpass its initial Vision 2030 tourism milestone by recording more than 100 million domestic and international visits several years ahead of schedule (Saudi Ministry of Tourism, 2025). The strategy focuses on leveraging cultural heritage, diverse natural landscapes, and large-scale Tourism is increasingly recognized as a strategic sector in Saudi Arabia's Vision 2030 plan, which seeks to diversify the economy, reduce oil dependence, and enhance the country's international image (Saudi Arabia Vision 2030, 2016). To support these objectives, the Kingdom has invested heavily in hosting high-profile international events such as the FIFA Club World Cup, Formula 1 Saudi Arabian Grand Prix, Supercoppa Italiana, Spanish Super Cup, and major boxing matches designed to attract large numbers of foreign visitors and strengthen Saudi Arabia's position as a global tourism destination. These efforts have already enabled the country to surpass its initial Vision 2030 tourism milestone by recording more than 100 million domestic and international visits several years ahead of schedule (Saudi Ministry of Tourism). The national target has therefore been revised to 150 million visits by 2030, with the tourism sector expected to contribute about 10% of gross domestic product and create approximately 1.6 million new jobs (Saudi Arabia Vision 2030, 2016; Saudi Ministry of development projects to diversify the economy, improve quality of life, and encourage sustainable investment.

Despite these ambitious initiatives, little is known about how international audiences perceive these large-scale events on social media, leaving a critical gap in understanding international sentiment toward Saudi Arabia's emerging tourism sector. Accurately measuring such sentiment is essential for evaluating promotional effectiveness, enhancing visitor experiences, and informing policy decisions related to Vision 2030 goals. Traditional approaches such as surveys and interviews are limited by high cost, small sample sizes, and delayed feedback, and they often face language barriers that complicate data collection and analysis across diverse visitor groups (Manosso and Ruiz, 2021); these constraints highlight the need for large-scale, data-driven methods capable of capturing real-time public opinion from platforms such as X (formerly Twitter), where English and other widely used languages enable broader coverage of international audiences (Xiang et al., 2017; Manosso and Ruiz, 2021).

Social media offers a powerful alternative to traditional data-collection methods, providing real-time access to large volumes of unsolicited user-generated content (Alaei et al., 2019; Zeng and Gerritsen, 2021). Among these platforms, X (formerly Twitter) is particularly valuable because of its global reach, high posting frequency, and availability of public data, which enable large-scale sentiment analysis across diverse user groups. Researchers have successfully applied sentiment analysis of social media to a variety of tourism-related tasks, including destination branding (Hudson, 2015), and event evaluation (Leung et al., 2013). Sigala (2017) further demonstrated how social media supports crisis communication and reputation recovery in tourism destinations. Nevertheless, applications focusing on Middle Eastern destinations remain rare, and only a limited number of studies have examined international perceptions of Saudi Arabia's tourism initiatives and mega-events.

Furthermore, most tourism-focused sentiment studies have employed traditional machine learning classifiers such as Naive Bayes and Support Vector Machines (SVM) which require extensive feature engineering and often struggle to capture linguistic context and subtle sentiment cues (Medhat et al., 2014; Liu, 2012). Recent advances in Natural Language Processing (NLP), particularly transformer-based Large Language Models (LLMs) such as BERT, have markedly improved performance in text classification tasks, including sentiment analysis (Devlin et al., 2019). Fine-tuning these pre-trained models on domain-specific datasets allows context-aware classification while reducing the need for manual feature design and performing well even with moderate amounts of training data. Therefore, this study aims to investigate how international audiences perceive Saudi Arabia's mega-events and whether transformer-based models can accurately capture these perceptions in real time.

To fill this research gap, we propose a large-scale, data-driven framework that leverages social media content to quantify and analyze the perceptions of international audiences toward Saudi Arabia's tourism initiatives.

The contributions of this study are primarily empirical and application-oriented rather than algorithmic. First, to the best of our knowledge, the study provides the first systematic analysis of English-language social media discourse surrounding major tourism and sporting events hosted in Saudi Arabia. Second, it examines sentiment dynamics across different stages of the event lifecycle, offering insights into how public perceptions evolve before, during, and after major events. Third, the findings contribute evidence relevant to evaluating the international visibility and public engagement objectives associated with Saudi Vision 2030 tourism initiatives.

Related Work

Sentiment Analysis in Tourism

Understanding tourists' perceptions is central to destination marketing, planning, and reputation management. Sentiment analysis enables researchers to quantify public opinion about destinations, attractions, and events. Early work often employed traditional machine learning techniques such as Naive Bayes or Support Vector Machines (SVM), which required extensive feature engineering and struggled with context or nuanced language (Medhat et al., 2014; Liu, 2012).

Previous research has demonstrated the value of sentiment analysis for branding and market insight. For example, Hudson (2015) showed how visitor-generated content influences perceived destination image and marketing effectiveness, while Leung et al. (2013) provided a comprehensive review of social media's role in tourism and hospitality, emphasizing applications in reputation management and customer engagement.

Recent systematic reviews confirm that sentiment analysis remains a core method in tourism research but highlight how it is increasingly supported by big data and NLP advances that improve scalability and contextual understanding (Marine-Roig, 2020; Kang et al., 2021; Srianan et al., 2025). Researchers now analyze millions of online reviews, social media posts, and travel blogs to monitor destination reputation, forecast demand, and inform policy decisions.

Social Media as a Data Source

Social media platforms have become essential data sources for tourism research because they provide unsolicited, real-time, large-scale insights into tourists' attitudes and behaviors. X (formerly Twitter), Facebook, Instagram, and review platforms like TripAdvisor enable researchers to access authentic visitor opinions shared without the constraints of traditional surveys or interviews.

Alaei et al. (2019) demonstrated how big data from social media can inform tourism planning by identifying emerging trends, visitor expectations, and areas of dissatisfaction. Zeng and Gerritsen (2021) reviewed the transformative role of social media in tourism marketing and management, showing that destinations increasingly use these platforms to engage with global audiences and shape perceptions.

Recent studies have further underscored social media's role in supporting crisis communication, branding, and market segmentation. Marine-Roig (2020) reviewed big data applications in tourism and highlighted the growing sophistication of social media analytics, while Srianan et al.

(2025) provided a comprehensive comparative evaluation of traditional machine learning, deep learning, and transformer models for tourism sentiment analysis, highlighting the superiority of transformer-based approaches on imbalanced datasets.

Advances in NLP and Large Language Models

While early sentiment analysis approaches in tourism often relied on bag-of-words representations and manual feature engineering, recent advances in Natural Language Processing (NLP) have transformed the field. Transformer-based Large Language Models (LLMs) such as BERT represent a breakthrough by learning rich semantic and syntactic relationships within text data (Devlin et al., 2019). These models can be fine-tuned on domain-specific datasets to deliver high accuracy even with moderate annotation efforts.

LLMs have significantly improved the ability to handle informal language, sarcasm, and context-dependence typical of social media text. Marine-Roig (2020) highlighted the growing sophistication of social media analytics in tourism, while Kang et al. (2021) identified a geographic concentration in existing research toward established destinations, and Srianan et al. (2025) provided a comprehensive comparative evaluation of sentiment analysis methods for tourism, emphasizing the superiority of transformer-based models on imbalanced datasets.

Despite these methodological improvements, adoption remains uneven globally, with advanced NLP models applied disproportionately to destinations with abundant digital footprints and active marketing investment. Studies continue to focus on established destinations, leaving emerging tourism markets relatively underexplored.

Traditional machine learning approaches such as Naïve Bayes and Support Vector Machines (SVMs) have been widely applied to sentiment analysis because of their computational efficiency and interpretability. However, these methods typically require manual feature engineering and often struggle to capture contextual meaning and semantic nuances. Deep learning architectures such as Long Short-Term Memory (LSTM) networks improve sequential modeling capabilities but may encounter difficulties in learning long-range dependencies and require substantial training data. Convolutional Neural Networks (CNNs) are effective at extracting local textual patterns but may fail to fully capture broader contextual relationships. Transformer-based models such as BERT address many of these limitations through bidirectional contextual representations, enabling improved understanding of informal and context-dependent social media language. Nevertheless, these models generally require greater computational resources and task-specific fine-tuning. A comparative summary of these approaches is presented in Table 1, highlighting the trade-offs that informed the model selection in this study.

Given the informal, context-dependent, and event-oriented nature of social media content analyzed in this study, BERT was selected as the primary sentiment analysis model.

Research Gap: Saudi Arabia and Mega-Events

Systematic reviews highlight a notable geographic imbalance in tourism sentiment analysis research. The majority of studies have focused on destinations in Europe, North America, and East Asia, leveraging data availability and well-established tourism markets with active marketing agencies investing in social listening tools (Marine-Roig, 2020; Kang et al., 2021; Srianan et al., 2025). European-focused studies often analyze TripAdvisor reviews to assess hotel quality and branding, while East Asian research uses platforms like Weibo to monitor traveler sentiment and forecast demand. Kang et al. (2021) found that over half of social media sentiment analysis studies in tourism targeted urban destinations in East Asia.

By contrast, research specifically examining Saudi Arabia's tourism sector is emerging but remains limited in scope and methodological approach. Much of the existing work focuses on strategic development goals and policy implications under the Vision 2030 initiative. For example, Henderson (2020) analyzes the strategic objectives and policy measures associated with diversifying Saudi Arabia's economy through tourism. Alhazmi and Nyaupane (2017) discuss the challenges and opportunities of tourism development in Saudi Arabia, including infrastructure, cultural sensitivities, and workforce training.

Table 1: Comparison of sentiment analysis methods: Strengths and limitations.

Method	Strengths	Limitations
Naïve Bayes	Computationally efficient, simple implementation, interpretable	Requires manual feature engineering, limited contextual understanding
SVM	Effective in high-dimensional text classification, robust performance	Relies on handcrafted features, limited semantic awareness
CNN	Captures local textual patterns efficiently, relatively fast training	Limited ability to model long-range contextual dependencies
LSTM	Models sequential information and word order	Computationally intensive, susceptible to vanishing gradient issues in long sequences
BERT	Bidirectional contextual understanding, strong performance on informal social media text	Higher computational cost, requires task-specific fine-tuning

Other studies have explored heritage and cultural tourism branding. Aloudat and Michael (2019) conducted qualitative content analysis of government marketing materials to evaluate how Saudi heritage sites are promoted to domestic and regional markets. A smaller but growing body of work has applied social media analysis to assess domestic public opinion about tourism reforms. Alasmari et al. (2022) applied machine learning algorithms to Arabic-language Twitter data to analyze sentiment toward tourist destinations in Saudi Arabia, identifying key visited locations and evaluating the performance of multiple classification approaches.

However, these studies share two important limitations. First, they largely focus on domestic or regional audiences, leaving the perceptions of international audiences underexamined. Second, even studies using social media data typically analyze Arabic-language content, missing the English-language discourse of global audiences that Vision 2030 aims to attract through mega-events such as the FIFA Club World Cup, Formula 1 Grand Prix, Supercoppa Italiana, Spanish Super Cup, and major boxing matches.

Given the scale of Saudi Arabia's investment in hosting high-profile international events and repositioning itself as a global tourism destination, understanding how these initiatives are perceived by international audiences is essential. This study addresses this gap by systematically analyzing 10,000 English-language tweets about major Saudi tourism events collected between December 2023 and December 2024. By leveraging large language models for sentiment classification, we provide empirical insights into international audience perceptions that can inform evidence-based marketing, visitor experience design, and policy planning.

Materials and Methods

This section presents the end-to-end workflow used to analyze and classify English-language social media discourse related to major events in Saudi Arabia. As illustrated in Figure 1, the proposed framework follows five sequential stages:

1) Event related social post collection from multiple platforms (e.g., X, Facebook, Instagram)
2) Sentiment preserving preprocessing and normalization
3) Temporal organization into before-, during-, and after-event phases
4) Sentiment classification using a fine-tuned BERT model
5) Model evaluation through quantitative metrics

Unlike prior approaches that treat social-media data as a static corpus, our framework explicitly models temporal sentiment dynamics, enabling analysis of how international perceptions evolve throughout the lifecycle of each event. This modular design allows easy extension to additional data sources or future multilingual settings.

Problem Definition

The core objective of this study is to automatically classify the sentiment expressed in English-language social media posts related to tourism events in Saudi Arabia. Driven by the Kingdom's increasing portfolio of international events including the FIFA Club World Cup, the Formula 1 Saudi Arabian Grand Prix, and globally broadcast boxing championships platforms such as *X* (formerly Twitter) provide a rich, real-time stream of unsolicited opinions from international audiences. These opinions, however, exhibit high linguistic diversity, informal grammar, and sentiment cues such as emojis and sarcasm, making automated analysis both essential and challenging.

Let $D = \{(x_i, y_i)\}^N_{i=1}$ denote a dataset of N English tweets, where x_i is the i-th tweet and y_i 0,1,2

represents the sentiment label negative (0), neutral (1), or positive (2). Each tweet x_i is a sequence of tokens $x_i = [w_1, w_2, \ldots, w_T]$, with T tokens. The task is formulated as a multi–class classification problem:

$$f : X \to Y, \qquad f(x_i) = \hat{y}_i \text{ where } \hat{y}_i \in \{0,1,2\},$$

Such that the predicted label $\hat{y}_i$ closely matches the true sentiment y_i.

Beyond static sentiment assignment, our framework also aims to capture the temporal dynamics of public sentiment about Saudi tourism events.

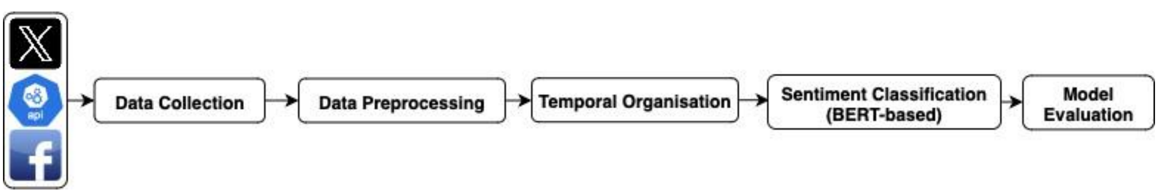

Fig. 1: Overall architecture of the proposed sentiment analysis pipeline for Saudi tourism events. The workflow consists of five sequential stages: Data collection from social platforms, preprocessing and normalization, temporal organization to align tweets with event timelines, sentiment classification using a fine-tuned BERT model, and model evaluation

Tweets are associated with the official schedule of each event and grouped into three stages before, during, and after to enable analysis of how positive, neutral, and negative sentiment evolve over the life cycle of an event. Given the informal nature of social media, traditional models often fail to capture these nuanced signals. To address this challenge, we employ pre-trained transformer-based language models, specifically BERT-base-uncased, which can learn deep contextual representations and are fine-tuned on the curated tourism tweet corpus to adapt to the domain-specific characteristics of user-generated content.

Although this work focuses on English-language tweets, the problem formulation naturally extends to future cross-lingual settings in which Arabic or other languages can be incorporated for comparative analyses.

Data Collection

To investigate English-language sentiment patterns related to Saudi tourism events, we compiled a multi-event dataset of English-language tweets related to major international sports and entertainment events hosted in Saudi Arabia. X (formerly Twitter) was selected as the data source because it provides real-time, large-scale user-generated content and has been widely used for tourism research to gauge sentiment and track visitor perceptions (Zeng and Zhang, 2021).

Event Selection. Events were chosen for their global prominence, alignment with Saudi Arabia's Vision 2030, and ability to attract a diverse international audience. Vision 2030 positions tourism as a key pillar for economic diversification and cultural exchange, aiming to surpass 100 million annual visits by Saudi Arabia Vision 2030 (2016). Accordingly, we targeted high-profile events such as: The FIFA Club World Cup (2023), the Formula 1 Saudi Arabian Grand Prix (2024), the Spanish and Italian Super Cups (2024 & 2025), and several globally televised boxing championships. These mega-events generate concentrated bursts of online discussion and provide ideal settings for focused sentiment analysis, as documented in event tourism literature (Getz, 2008) and recent comparative studies on tourism sentiment classification (Srianan et al., 2025).

Temporal Staging. To capture sentiment dynamics over the life cycle of each event, tweets were collected in stages aligned with the official event schedules:

- Stage 1: FIFA Club World Cup (Jeddah) – December 1–30, 2023
- Stage 2: Formula 1 Saudi Arabian Grand Prix – March 1–14, 2024
- Stage 3: Spanish Super Cup and Supercoppa Italiana – 2024 and 2025 tournament periods
- Stage 4: Internationally promoted boxing events throughout 2024

This stage-aware design supports the temporal grouping, enabling comparison of sentiment before, during, and after each event.

Collection Process. Tweets were collected using the X API, with retrieval performed separately for each event within predefined temporal windows. Search queries were constructed using event-specific hashtags and keywords selected based on official event names and commonly used discussion terms observed on the platform during each event period, including hashtags such as #FIFAClubWorldCup, #SaudiGP, #SpanishSuperCup, and #SaudiBoxing, among others. The final dataset consisted of approximately 10,000 English-language tweets that matched the predefined search criteria. Duplicate tweets were removed using tweet ID deduplication, and non-English posts were filtered using the language metadata provided by the X API. Geolocation filtering was not applied because reliable location information was unavailable for a substantial proportion of collected tweets; therefore, event-specific hashtags and keywords served as the primary retrieval criteria. Dedicated bot-detection techniques were not employed, as the primary objective of the study was sentiment analysis of event-related social media discourse rather than account

authenticity assessment. We acknowledge that API access constraints, the absence of dedicated bot filtering, and reliance on event-specific hashtags and keywords as retrieval criteria may introduce sampling bias. These limitations are acknowledged and discussed in the limitations section.

Data Preprocessing

To prepare the collected tweets for modeling, we applied a sentiment-preserving preprocessing pipeline designed to reduce noise while retaining key affective cues. Unlike generic cleaning procedures that remove all special tokens, our pipeline carefully normalizes text but keeps informative symbols (e.g., emojis and hashtags) that often carry sentiment signals.

Specifically, the following steps were performed:

- Normalization: All text was lowercased and converted to UTF-8 format. URLs and user mentions were replaced with special placeholders (<URL>, <USER>) to retain structural information without leaking external content
- Hashtag Processing: Hashtags were segmented into constituent words (e.g., #SaudiSeason “Saudi Season”) to preserve semantic content for the tokenizer
- Emoji Handling: Emojis were mapped to short textual tags (e.g., “:smile:”, “:sad:”) rather than removed, ensuring that affective signals remained available to the model
- Noise Removal: Excess whitespace, HTML entities, and non-informative punctuation were stripped to reduce

input variability

- Negation Preservation**:** Negation words (e.g., not, never, no) were deliberately retained and not removed as stopwords, ensuring that sentiment-reversing modifiers remain available to the model as part of the input sequence

Finally, tokenization was performed using the BERT WordPiece tokenizer, which segments text into subword units to handle rare or compound words while maintaining compatibility with the BERT encoder used in the downstream classification task. This preprocessing strategy balances noise reduction with the retention of sentiment bearing features critical for accurate classification.

Sentiment labels for the 10,000 collected tweets were generated using the CardiffNLP Twitter Sentiment model (twitter-roberta-base-sentiment), a RoBERTa-based classifier fine-tuned on English Twitter data for three-class sentiment classification (positive, neutral, and negative) (Barbieri et al., 2020).

Temporal Organization

To capture how public sentiment about Saudi tourism events evolves over the life cycle of each event, all preprocessed tweets were organized into temporal stages and linked to the correct event before being passed to the BERT classifier. This step ensures that sentiment distributions can be compared both across events and across time. As illustrated in Figure 2, the collected social media posts are temporally segmented into three distinct subsets, before, during, and after the event to capture dynamic shifts in sentiment and engagement patterns.

Event Matching**.** Each tweet was first associated with a specific event *e* using a curated set of event-specific hashtags and keywords compiled from official announcements and media coverage. During data collection, X (formerly Twitter) API queries were constructed with these hashtags and restricted to English-language tweets. After collection, tweets were re-filtered to ensure that their text contained at least one predefined token for event *e*. Tweets matching multiple event lists were rare; when a dominant hashtag frequency could be determined, the tweet was assigned to that event, otherwise it was discarded to avoid ambiguity.

Stage Definition**.** For every event *e*, three temporal windows were defined according to the official event schedule:

- Before: From the start of data collection up to 24 hours before the official start time
- During: From the official start to the official end of the event
- After: From the end of the event to 48 hours post- event

These windows were chosen to capture anticipation, real- time reactions, and post-event reflections while keeping boundaries consistent across events.

Stage Assignment. Given a tweet with creation time t_i matched to event e, the stage label $s_i \in \{\text{before, during, after}\}$ is assigned by comparing t_i with the start and end timestamps $[t^{(e)}_{start}, t^{(e)}_{end}]$. This requires only a single interval check per tweet, making the procedure computationally efficient even with multiple events.

Outcome. The output of this process is a temporally grouped dataset $\{(x_i, y_i, s_i, e_i)\}$, where x_i is the tweet text, y_i the sentiment label, s_i the temporal stage, and e_i the corresponding event. These structured groupings enable stage-specific sentiment analysis and direct comparison of public perceptions before, during, and after each Saudi mega-event.

Model Architecture

The proposed sentiment analysis framework combines a BERT-base-uncased encoder with a lightweight classification head to capture nuanced polarity cues in English tweets. Unlike prior work that treats social media data as a static corpus, our model operates on temporally organized tweet groups, allowing the same architecture to reveal stage-specific sentiment dynamics across before, during, and after event periods.

Model Overview

Given a raw tweet, we first apply a minimal, sentiment-preserving normalization that keeps emojis and decomposes hashtags. The normalized text is tokenized using WordPiece (Wu et al., 2016) and fed into a stack of L transformer encoder layers (Vaswani et al., 2017). We then extract the contextual representation of the special token [CLS] and feed it to a dense classification head with softmax to obtain the sentiment distribution over Y = positive, neutral, negative. Because BERT computes bidirectional attention over the full input sequence, each token's representation is conditioned on all surrounding tokens.

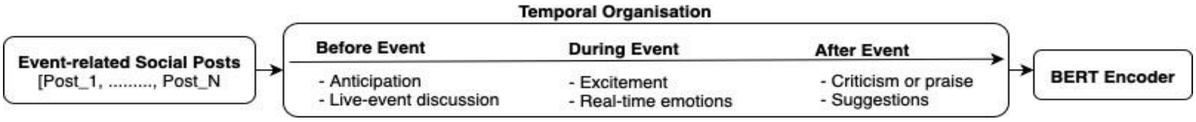

Fig. 2: Temporal organization of event-related posts illustrating before-, during-, and after-event segmentation. The collected posts are organized into three temporal phases: Before Event (anticipation and live-event discussion), During Event (excitement and real-time emotions), and After Event (criticism, praise, or suggestions). Each subset is subsequently fed into the BERT encoder for sentiment analysis

DOI: 10.3844/jcssp.2026.2802.2813

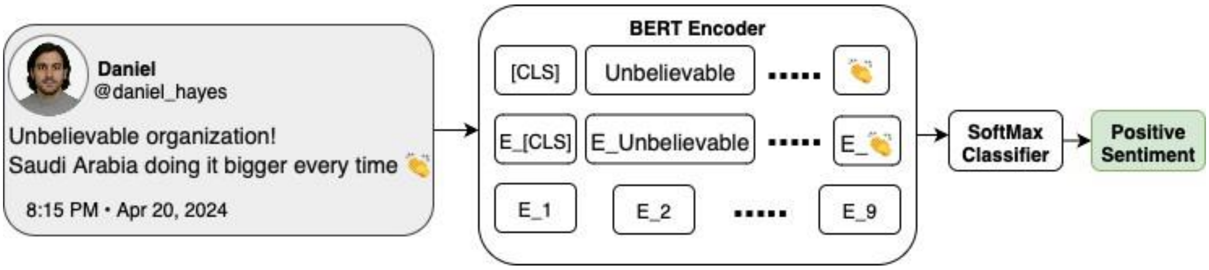

Fig. 3: Example of BERT-based sentiment encoding for an event-related tweet containing emoji. A real tweet containing both text and emoji is tokenized, embedded, and passed through the BERT encoder. The model preserves the emoji tokens (*clapping hands*) as part of the sentiment-bearing input, and the Softmax layer classifies the overall sentiment as Positive

This means that the same word receives a different embedding depending on its context for example, the word *clean* in "the stadium was clean" and in "the event was clean" yields distinct representations. This property allows the model to handle polysemy and context-dependent sentiment cues without requiring explicit word-sense disambiguation. To further clarify how the proposed model processes user-generated content, Figure 3 illustrates an example of a real tweet being encoded by BERT and classified into a sentiment category.

Input Representation

Let the input sequence be $x_i = [w_1, w_2, \ldots, w_T]$. Each token is mapped to three types of embeddings token, positional, and segment and summed as:

$$e_j = E_t(w_j) + E_p(j) + E_s, \tag{1}$$

Where e_j is the input embedding of the j-th token, $E_t(\ )$ are token embeddings, $E_p(\ )$ are positional embeddings, and E_s are segment (sentence-type) embeddings as in (Devlin et al., 2019).

We keep sentiment-bearing artifacts:

- Emojis are retained and treated as tokens (or mapped to sentiment-aware aliases, e.g., :smile:)
- Hashtags are split via camel-case/underscores before WordPiece tokenization (e.g., #SaudiSeason Saudi Season)
- User mentions and URLs are replaced by canonical placeholders to reduce sparsity while preserving pragmatic cues

Transformer Encoder

Each encoder layer comprises Multi-Head Self-Attention (MHSA) and a position-wise Feed-Forward Network (FFN), both wrapped with residual connections and Layer Norm. The computations are defined as follows:

$$\mathrm{SA}(Q, K, V) = \mathrm{softmax}\, QK^{\top} V, \tag{2}$$

$$\mathrm{MHSA}(X) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_h) W^O, \tag{3}$$

$$\mathrm{head}_j = \mathrm{SA}(XW^Q, XW^K, XW^V), \tag{4}$$

$$\tilde{X} = \mathrm{LayerNorm}(X + \mathrm{MHSA}(X)), \tag{5}$$

$$\mathrm{FFN}(u) = \sigma(uW_1 + b_1)W_2 + b_2, \tag{6}$$

$$H = \mathrm{LayerNorm}(\tilde{X} + \mathrm{FFN}(\tilde{X})). \tag{7}$$

Here, d_k is the key dimension, h is the number of attention heads, and σ is a nonlinearity (e.g., GELU). Dropout is applied after both the attention and FFN sub-layers.

Document Level Pooling and Classification

We use the hidden state of the [CLS] token after the last encoder layer, h_{CLS} R^d, as a global sequence representation. A linear layer followed by a softmax function yields the class posterior probabilities:

$$\hat{y} = \mathrm{softmax}(W_c h_{\mathrm{CLS}} + b_c) \in \mathrm{R}^{|Y|}, \tag{8}$$

Where W_c and b_c are the classification weight matrix and bias vector, respectively, and Y denotes the set of sentiment classes.

The class-weighted cross-entropy loss used for training compensates for the natural imbalance among positive, neutral, and negative tweets, while leaving the inference pathway unchanged. This ensures that sentiment distributions can be compared fairly across the temporal stages.

The primary metrics. These measures are more informative than accuracy in imbalanced social–media datasets.

Model Design and Training

We fine-tuned the bert-base-uncased Transformer model for sentiment classification due to its strong performance in contextual understanding and robustness to informal social media language (Devlin et al., 2019). A single dense classification layer was placed on top of the [CLS] token, with dropout and class-weighted cross-entropy loss to mitigate class imbalance. Tweets were grouped into pre-, during-, and post-event stages to enable temporal analysis. Dataset splitting followed a 70/15/15 stratified partition for training, validation, and testing. The model was optimized using AdamW with a linear learning-rate scheduler and early stopping. Training was conducted using PyTorch and HuggingFace Transformers on a single NVIDIA GPU. Key hyperparameters are summarized in Table 2.

Evaluation Metrics

To evaluate sentiment classification performance, we follow established NLP practice (Powers, 2011; Manning et al., 2008) and report precision, recall, and F1-score as: Precision. Precision measures the proportion of correctly predicted positive instances:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

Recall. Recall reflects the ability to retrieve all actual positive instances:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

F1-Score. F1-score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{precision \times \text{Re}call}{precision + \text{Re}cal} \tag{11}$$

Confusion Matrix. We also inspect the confusion matrix to analyze model behavior across sentiment classes (positive, neutral, negative) and identify misclassification patterns.

Experimental Setup

All experiments were conducted using Python 3.10 and PyTorch, with scikit-learn for baseline models. Training was performed on a GPU-enabled workstation.

(NVIDIA Tesla V100, 32 GB VRAM; Intel Xeon CPU; 256 GB RAM). The dataset was split into 70% training, 15% validation, and 15% testing using stratified sampling to preserve sentiment distribution.

Table 2: Hyperparameter configuration used for model training

Hyperparameter	Value
Learning rate	2×10^{-5}
Batch size	32
Epochs	10
Optimizer	AdamW
Weight decay	0.01
Dropout rate	0.1
Warm-up steps	500
Max sequence length	128 tokens
Loss function	Class-weighted cross-entropy

Tweets were preprocessed and tokenized using the bert-base-uncased tokenizer. For traditional baselines, TF–IDF features were used. The BERT model was fine-tuned using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 32, and early stopping. Baseline models (e.g., SVM, LSTM, CNN) were tuned using standard grid search. All results are averaged over three runs, and model performance was evaluated using precision, recall, F1-score, and confusion matrix analysis.

Baseline Models

To benchmark the proposed approach, we compared against widely used sentiment-classification models.

SVM: Linear Support Vector Machine trained on TF–IDF features.

Logistic Regression: ℓ_2-regularized logistic regression with TF–IDF vectors.

LSTM: Single-layer LSTM initialized with pre-trained embeddings for sequence modeling.

CNN: Convolutional network with multiple kernel sizes for n-gram feature extraction.

BERT: Fine-tunedbert-base-uncased transformer as the state-of-the-art baseline.

Proposed Model: Enhanced BERT with temporal sentiment grouping and class-weighted fine-tuning tailored for tourism-event analysis.

Proposed Model. Our proposed framework builds on bert-base-uncased by fine-tuning it with a lightweight, class-weighted classification head and integrating the temporal-stage organization. This design aims to balance precision and recall across sentiment categories while enabling stage-specific analysis before, during, and after each event.

Results and Discussion

This section presents the experimental results obtained from evaluating the proposed sentiment classification framework. The aim is not only to validate the effectiveness of the model but also to answer specific research questions that guide our analysis. In particular, we investigate the following:

- **RQ1:** How does temporal grouping (before, during, after events) influence the distribution and interpretation of public sentiment expressed in event-related tweets?
- **RQ2:** What hyperparameter configurations yield the most robust performance for transformer-based sentiment analysis on imbalanced tourism data?
- **RQ3:** How does the proposed model compare with traditional machine learning and deep learning baselines in terms of precision, recall, and F1-score?

By systematically addressing these questions, we provide both quantitative evidence and qualitative insights into the strengths and limitations of the proposed approach. Each research question is examined in detail in the following subsections.

RQ1: Impact of Temporal Grouping on Sentiment

Table 3 summarizes the sentiment distribution before, during, and after the selected events. We observe that the “during-event” stage consistently exhibits the highest proportion of positive tweets, reflecting excitement and engagement, whereas the “before-event” stage contains a larger fraction of neutral or uncertain opinions. The “after-event” stage demonstrates a slight decline in positive sentiment, with negative tweets typically associated with organizational or logistical criticisms.

These results confirm that temporal organization is crucial: Analyzing tweets without stage separation would mask the sharp rise in positive sentiment during events. This aligns with prior findings that event-driven sentiment fluctuates dynamically with time.

These patterns are consistent with established theories of event-driven consumer behavior and social sharing. The peak in positive sentiment during events reflects real-time excitement and collective engagement, as audiences share live reactions on social media platforms. The higher proportion of neutral tweets before events aligns with anticipatory uncertainty users discuss expectations without direct experiential evidence. The slight decline in positive sentiment after events is consistent with post-consumption evaluation, where reflective assessment introduces more critical perspectives on logistics, organization, and value.

RQ2: Hyperparameter Tuning

We investigated the impact of hyperparameters on model performance, focusing on learning rate, batch size, and maximum sequence length. Table 4 shows representative results for BERT fine-tuning.

Table 3: Sentiment distribution across temporal stages of major tourism events. Values are percentages of total tweets

Stage	Positive	Neutral	Negative
Before Event	35%	45%	20%
During Event	50%	35%	15%
After Event	42%	40%	18%

Table 4: Effect of hyperparameter settings on BERT performance (F1-score)

Configuration	Precision	Recall	F1-score
Learning Rate = 5×10^{-5}, Batch = 16	0.8	0.76	0.78
Learning Rate = 2×10^{-5}, Batch = 32	0.83	0.82	0.82
Learning Rate=1×10^{-5}, Batch = 64	0.81	0.79	0.8

We found that a learning rate of 2×10^{-5} with a batch size of 32 and maximum sequence length of 128 tokens yielded the best trade-off between stability and accuracy. Higher learning rates led to unstable convergence, while smaller batch sizes introduced variance in gradient updates.

This confirms that hyperparameter optimization is essential in tourism-related sentiment analysis, where datasets are relatively small and class distribution is imbalanced.

RQ3: Comparison With Baseline Models

To evaluate the effectiveness of the proposed transformer-based sentiment model, we compared it against five competitive baselines: Logistic Regression (LR), Support Vector Machine (SVM), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), and the standard BERT-base model. Table 5 summarizes the averaged results across three independent runs.

The results show a clear progression in performance from traditional machine-learning models to deep learning and transformer-based models. LR and SVM achieved moderate results but struggled to capture contextual sentiment nuances particularly sarcasm, intensifiers, and multi-word expressions common in social media text. LSTM and CNN improved performance by modeling sequence dynamics and local n-gram patterns, yet they remained limited in capturing long-range dependencies and semantic context.

Vanilla BERT substantially outperformed the classical and RNN/CNN architectures, achieving an F1-score of 0.81, highlighting the strength of contextual embeddings and bidirectional language modeling. The proposed model further improved performance to 0.85, demonstrating three key advantages:

(i) Task-specific fine-tuning with class- weighted loss to address sentiment imbalance
(ii) Temporal organization of tweets into event stages enabling stage- specific sentiment interpretation
(iii) Regularization strategies such as dropout and early stopping that prevented overfitting and stabilized training

Table 5: Performance comparison across baseline models. Best scores in bold

Model	Precision	Recall	F1-score
Logistic Regression (LR)	0.71	0.68	0.69
Support Vector Machine (SVM)	0.75	0.70	0.72
Long Short-Term Memory (LSTM)	0.78	0.74	0.76
Convolutional Neural Network (CNN)	0.80	0.75	0.77
BERT (base)	0.83	0.80	0.81
Proposed Model	0.87	0.84	0.85

Overall, the results confirm that combining transformer-based representations with targeted adaptation strategies yields more robust sentiment detection in tourism-related social media analytics.

The performance gain of the proposed model over vanilla BERT is primarily attributable to the use of class-weighted cross-entropy loss, which directly addresses the natural imbalance in sentiment data where positive tweets dominate by penalizing misclassification of minority classes more heavily. Temporal grouping serves as an analytical framework that enables stage-specific interpretation of sentiment distributions, revealing how public perceptions shift across the event lifecycle rather than improving training directly.

Confusion Matrix and Interpretation

The confusion matrix in Figure 4 provides a detailed view of classification performance across the three sentiment categories. Correct predictions appear on the diagonal, while off-diagonal values represent misclassifications. As shown, the majority of samples in each class fall along the diagonal, confirming the model's strong overall performance across positive, neutral, and negative classes.

Error Analysis

To better understand model behavior, misclassification patterns were examined using the confusion matrix (Figure 4). The most frequent errors occurred between the neutral and positive classes. Specifically, a portion of neutral tweets was classified as positive, while some positive tweets were predicted as neutral. This pattern suggests that many event-related social media posts contain subtle evaluative language that blurs the distinction between objective descriptions and mild positive sentiment. In contrast, confusion between positive and negative classes was relatively limited, indicating that the model effectively distinguishes strongly polarized opinions.

Negative sentiment was also classified with high reliability, likely because negative posts often contain more explicit sentiment cues than neutral content. Overall, the results indicate that the principal challenge lies in identifying nuanced sentiment intensity rather than separating clearly positive and clearly negative reactions.

The results demonstrate that the proposed BERT-based model effectively captures sentiment patterns in social media data related to tourism events. The observed increase in positive sentiment during events highlights the strong engagement and international interest generated by these activities. Compared to traditional machine learning models, the transformer-based approach provides better contextual understanding, particularly for informal and short text such as tweets.

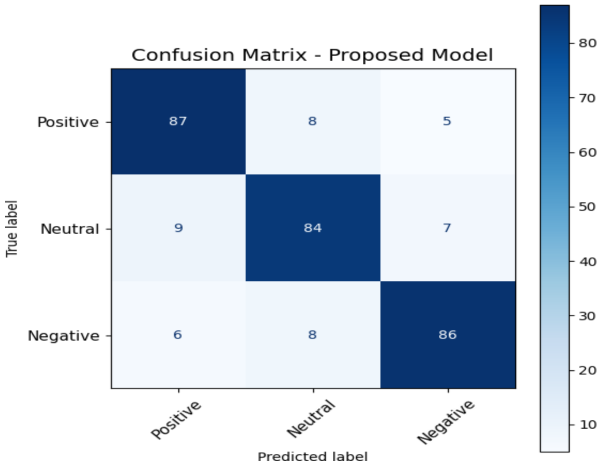

Fig. 4: Confusion matrix showing classification performance of the proposed model. Correct predictions lie on the diagonal

From a policy perspective, the predominance of positive sentiment observed during major Saudi events suggests that these initiatives contribute to favorable international visibility and public engagement. The findings indicate that large-scale sporting and entertainment events can play an important role in shaping international perceptions of Saudi Arabia and supporting the broader objectives of Vision 2030. Furthermore, sentiment monitoring frameworks provide decision-makers with a scalable mechanism for evaluating public response to major events and identifying areas requiring additional attention in future tourism and marketing strategies.

Several limitations should be acknowledged. First, the dataset is restricted to English-language tweets from a single platform (X), which may not fully represent the diversity of international audience discourse about Saudi tourism events. Second, English-language posts do not exclusively originate from international tourists; they may also include residents, expatriates, or Arabic speakers writing in English. Therefore, the findings should be interpreted as reflecting English-language social media discourse rather than tourist-specific sentiment. Third, the absence of dedicated bot-detection and spam-filtering techniques means that automated or promotional accounts may have influenced the sentiment distribution. Fourth, the reliance on event-specific hashtags and keywords without geolocation filtering may introduce sampling bias, as only users employing predefined event-related terms are captured. Fifth, the dataset covers a specific set of events between December 2023 and December 2024, limiting the generalizability of the findings to other event types or time periods. These limitations do not invalidate the observed sentiment trends, but they indicate that the findings should be interpreted as platform- and language-specific evidence rather than a complete representation of all tourist perceptions. Future studies may complement the descriptive analysis presented in this work with formal statistical significance testing using event-level count data. Overall, the findings confirm the importance of temporal analysis in understanding dynamic public perceptions and support the use of NLP techniques for data-driven tourism planning.

Conclusion and Future Work

This study demonstrated that transformer-based sentiment analysis can effectively capture public perceptions of major Saudi tourism events through temporal analysis of English-language social media content. Using 10,000 English tweets and temporal grouping (before, during, after events), the fine-tuned BERT model outperformed traditional ML and deep learning baselines, revealing strong positive sentiment peaks during events with gradual decline afterward. These insights demonstrate the role of social media analytics in supporting tourism decision-making aligned with Vision 2030.

Limitations include the focus on English-language data from a single platform, the absence of bot filtering and geolocation verification, and the event-specific scope of the dataset. Future extensions will incorporate multilingual data particularly Arabic and content from additional platforms such as Instagram and TripAdvisor. Further improvements may also involve multilingual transformers and complementary qualitative methods.

In summary, this work highlights the effectiveness of domain-adapted transformers for tourism sentiment monitoring and lays the foundation for broader multilingual and multi-platform analysis.

Acknowledgment

The author declares that no specific funding or institutional support was received for this research.

Funding Information

This research received no external funding.

Ethics

This study uses publicly available data from social media platforms. No personal or sensitive information was collected, and all data were processed in accordance with applicable data privacy guidelines.

References

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. *Journal of Travel Research*, *58*(2), 175–191. https://doi.org/10.1177/0047287517747753

Alasmari, W. A., Abdelhafez, H. A., & Abdelhafez, H. A. (2022). Twitter Sentiment Analysis for Reviewing Tourist Destinations in Saudi Arabia using Apache Spark and Machine Learning Algorithms. *Journal of Computer Science*, *18*(3), 215–226. https://doi.org/10.3844/jcssp.2022.215.226

Alhazmi, A., & Nyaupane, G. (2017). Tourism development in saudi arabia. *Tourism Planning & Development*, *14*(4), 504–524.

Aloudat, A., & Michael, K. (2019). The digital marketing of saudi arabia's heritage tourism: A content analysis approach. *International Journal of Tourism Research*, *21*(2), 123–138.

Barbieri, F., Camacho-Collados, J., Anke, L. E., & Neves, L. (2020). TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 1644–1650. https://doi.org/10.18653/v1/2020.findings-emnlp.148

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers. *NAACL-HLT*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Getz, D. (2008). Event tourism: Definition, evolution, and research. *Tourism Management*, *29*(3), 403–428. https://doi.org/10.1016/j.tourman.2007.07.017

Henderson, J. (2020). Saudi tourism sector in vision 2030. *International Journal of Tourism Policy*, *10*(2), 99–112.

Hudson, S. (2015). Social media in destination branding. *Journal of Travel Research*, *54*(6), 702–714.

Kang, J., Kim, S., Kim, J.-H., & Lee, S. (2021). Social media data analytics in tourism. *International Journal of Hospitality Management*, *95*, 102935.

Leung, D., Law, R., van, H. H., & Buhalis, D. (2013). Social Media in Tourism and Hospitality: A Literature Review. *Journal of Travel & Tourism Marketing 30*(1–2), 3–22. https://doi.org/10.1080/10548408.2013.750919

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. *5*.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*.

Manosso, F. C., & Ruiz, T. C. D. (2021). Using sentiment analysis in tourism research: A systematic, bibliometric, and integrative review. *Journal of Tourism, Heritage & Services Marketing*, *7*(2), 16–27.

Marine-Roig, E. (2020). A review of big data in tourism. *Tourism Review*, *75*(1), 212–228.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Saudi Arabia Vision 2030. (2016). Saudi Vision 2030. *Saudi Vision 2030*. https://www.vision2030.gov.sa/ar/

Sigala, M. (2017). *Social media and crisis management in tourism: Applications and implications for research*.

Srianan, S., Nanthaamornphong, A., & Phucharoen, C. (2025). Advancing tourism sentiment analysis: a comparative evaluation of traditional machine learning, deep learning, and transformer models on imbalanced datasets. *Information Technology & Tourism*, *27*(4), 1011–1045. https://doi.org/10.1007/s40558-025-00336-0

Saudi Ministry of Tourism. (2025). Saudi ministry of tourism dashboard. *Saudi Ministry of Tourism*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., & Macherey, W. (2016). Google’s neural machine translation system: Bridging the gap between human and machine translation. *ArXiv*.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms. *Tourism Management*, *58*, 51–65. https://doi.org/10.1016/j.tourman.2016.10.001

Zeng, B., & Gerritsen, R. (2021). What do we know about social media in tourism? A review. *Tourism Management Perspectives, 38*, 100813. https://doi.org/10.1016/j.tmp.2014.01.001

Zeng, B., & Zhang, Z. (2021). A review of research using social media data in tourism. *Tourism Management Perspectives, 38*, 100813.