

A Hybrid NLP Framework for Topic Relevance and Advertisement Detection in Trending Hashtags on X (Twitter)

Aadil Alshammari¹, Khulud Alashjaee^{2*}, Balsam Alzamam^{3*}

Aryam Alshammari⁴, Milaf Alshammari⁵, Mohamed Ayari⁶

Department of Information Systems-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia^{1,2,3,4,5}

Department of Information Technology-Faculty of Computing and Information Technology,
Northern Border University, Rafha 91911, Saudi Arabia⁶

Abstract—Trending hashtags on X (formerly Twitter) are widely used for real-time news tracking and public discourse, yet they are increasingly polluted by advertisements and off-topic posts that exploit hashtag visibility. This study proposes a hybrid natural language processing (NLP) framework for improving the reliability of hashtag streams through two complementary modules: 1) semantic topic relevance detection and 2) supervised advertisement classification. Using 9,671 Arabic tweets retained in the six final research datasets collected via the X API from multiple trending hashtags spanning diverse domains, we apply a dedicated Arabic preprocessing pipeline including noise removal and normalization. Topic relevance is assessed using AraBERT contextual embeddings by computing cosine similarity between each tweet and a concise hashtag topic description, followed by threshold-based labeling of on-topic versus off-topic content. In parallel, advertisement detection is formulated as a binary classification task using TF-IDF features and standard machine learning models (Logistic Regression, SVM, Random Forest, Naive Bayes, and SGD). Experimental results show that advertisement detection performance varies by hashtag and topic type: technical hashtags yield stronger and more stable F1-scores, while broad event hashtags are more challenging due to heterogeneous language and content. Across multiple hashtags, Random Forest and linear classifiers provide the most consistent trade-off between precision and recall. The proposed framework enables fine-grained quantification of promotional intrusion and semantic drift within hashtag discussions, providing a practical foundation for trustworthy hashtag analytics, improved trend monitoring, and downstream social media analysis in Arabic settings.

Keywords—Twitter; hashtags; advertisement detection; topic relevance detection; NLP; machine learning; Arabic text processing; social media analytics

I. INTRODUCTION

Online social networks have become a major channel for real-time communication, public discourse, and large-scale information diffusion, while social-media mining provides established foundations for extracting useful patterns from noisy, dynamic user-generated data [1], [2]. Among these platforms, Twitter (now X) plays a central role by allowing users to post short messages (tweets) that are organized and discovered

through *hashtags*. Hashtags act as user-generated labels that group tweets under a shared topic or event, and they are widely used to follow breaking news, public campaigns, and trending discussions. Consequently, analyzing hashtag-based streams provides valuable insights into collective attention, emerging issues, and user interests at scale.

However, the open and dynamic nature of Twitter also makes popular hashtags attractive targets for spam, promotion, and coordinated publicity campaigns. Previous studies have shown that many trending hashtags are polluted by spam accounts, automated bots, or promotional activities that exploit hashtag visibility to disseminate unsolicited content, advertisements, or misleading links [3]–[5]. Such hashtag-oriented promotion not only degrades user experience, but also distorts the perceived meaning and credibility of hashtags, which in turn affects downstream tasks such as trend monitoring, public opinion analysis, and event understanding. Automatically identifying promotional and irrelevant content within hashtag streams therefore constitutes an important and challenging research problem.

From a content analysis perspective, tweets associated with a given hashtag are not always uniformly related to the hashtag's intended meaning. While many tweets genuinely contribute to the discussion, others may be weakly related or entirely irrelevant, including advertisements that exploit popular hashtags for increased visibility. Rather than discovering latent topics through clustering or probabilistic topic modeling, an alternative approach is to assess the semantic relevance of individual tweets with respect to the intended topic of each hashtag. This perspective enables the separation of on-topic and off-topic content while preserving the original hashtag structure.

In parallel, natural language processing (NLP) and machine learning techniques have been extensively applied to Twitter data for tasks such as sentiment analysis, stance detection, and content classification. Traditional approaches typically rely on bag-of-words or TF-IDF representations combined with linear classifiers [6]. More recent advances leverage contextual embeddings derived from transformer-based architectures, building on BERT-style language representations [7]. Sentence-level embedding methods further support efficient semantic

*Corresponding author.

similarity comparisons using contextual representations and cosine similarity [8], which is directly relevant to short and noisy texts such as tweets.

The analysis of Arabic Twitter content introduces additional challenges due to the language's rich morphology, dialectal variation, and informal writing styles. Transformer-based models pretrained on Arabic-specific corpora, such as AraBERT, have demonstrated strong performance across a wide range of Arabic social media tasks, including sentiment analysis, spam detection, and content classification [9]–[11]. Despite these advances, most existing Arabic-focused studies concentrate on sentiment or toxicity detection, with limited attention given to hashtag-centered analysis and the identification of promotional content within individual hashtag streams.

Although spam and campaign detection on Twitter has received growing research interest, relatively few studies address the combined problem of assessing topical relevance and detecting promotional tweets at the level of individual hashtags, particularly in Arabic contexts. Much of the existing literature focuses on account-level or campaign-level detection [3], [5], without providing fine-grained tweet-level analysis that distinguishes organic contributions from advertisement-driven content within hashtag discussions.

In this work, we address these gaps by proposing a hashtag-centered NLP framework that integrates topic relevance detection with supervised promotional-content classification. Starting from a large collection of tweets grouped by trending Arabic hashtags, we apply preprocessing steps tailored to noisy and dialectal Arabic text. Contextual embeddings generated by an Arabic-specific transformer model (AraBERT) are used to measure the semantic alignment between tweets and predefined hashtag topic descriptions. An embedding-based similarity approach with a predefined threshold is then applied to determine whether a tweet is on-topic or off-topic. In parallel, supervised machine learning models trained on TF-IDF features are employed and comparatively evaluated to detect promotional and advertising-related tweets.

The main contributions of this study are threefold. First, we construct and describe a dataset of hashtag-based Twitter streams focusing on Arabic hashtags, preserving full tweet text and relevant metadata for fine-grained analysis. Second, we propose and evaluate a hybrid NLP pipeline that combines embedding-based topic relevance detection with supervised advertisement classification at the tweet level. Third, we present an empirical analysis that quantifies the presence of promotional content within trending hashtags and demonstrates the effectiveness of the proposed framework using standard classification metrics. The remainder of this study is organized as follows: Section II reviews related work; Section III describes the dataset and preprocessing steps; Section IV presents the proposed methodology; Section V reports and discusses the experimental results; and Section VI concludes the study and outlines directions for future research.

II. RELATED WORK

Research on analyzing Twitter content has grown substantially in recent years, driven by the platform's role in real-time communication, public opinion monitoring, and user-generated

event reporting. Existing studies generally fall into three major themes relevant to this work: 1) hashtag-oriented spam and promotional content detection, 2) topical analysis and relevance assessment in microblogs, and 3) natural language processing (NLP) techniques for Arabic Twitter analysis.

A. Hashtag Spam, Promotion, and Campaign Detection

Hashtags provide visibility and virality, making them attractive targets for spammers and promotional campaigns. Early foundational work by Sedhai and Sun introduced large-scale datasets such as HSpam14 [4] and conducted in-depth analyses of hashtag-oriented spam behavior across millions of tweets [3]. Their findings demonstrated that hashtags are frequently misused for unsolicited advertising, misleading links, and low-quality promotional content. Subsequent research examined the impact of spam on hashtag recommendation systems, showing that unsolicited promotional tweets significantly degrade recommendation quality and overall user experience [12].

Another line of research focuses on detecting large-scale promotional campaigns at the platform level. Zhang et al. proposed a behavioral detection framework that classifies accounts into normal, spam, and promotional campaigns by analyzing posting patterns and URL-sharing behaviors [5]. While these studies confirm the prevalence and sophistication of hashtag-oriented promotional activities, they primarily operate at the account or network level. However, such approaches do not explicitly address the identification of promotional content at the individual tweet level within specific hashtag streams. In contrast, the present work focuses on fine-grained tweet-level analysis, aiming to identify promotional tweets and assess their impact on the quality of hashtag-based discussions.

B. Topical Analysis and Relevance in Twitter

Analyzing topical content in short texts such as tweets is inherently challenging due to sparsity, noise, and limited context. Early studies highlighted the limitations of traditional topic modeling techniques, such as Latent Dirichlet Allocation (LDA), when applied to Twitter data and proposed adaptations for microblog environments [13]. Hong and Davison examined topic modeling in Twitter and demonstrated the importance of adapting topic-model training to short microblog messages [13].

Hashtag structures have been used to enhance topical interpretation, including graph-based hashtag analysis [14]. More recent work has established transformer-based tweet classification as a strong approach across heterogeneous social-media tasks [6]. While these methods are effective in discovering latent thematic structures, they primarily focus on topic extraction rather than evaluating whether individual tweets semantically align with the intended meaning of a specific hashtag. In contrast, our work adopts a relevance-based perspective, where predefined hashtag topic descriptions and embedding-based similarity measures are used to distinguish on-topic and off-topic tweets within hashtag streams.

C. NLP and Machine Learning for Arabic Twitter Analysis

Arabic Twitter content presents additional challenges due to the language's rich morphology, dialectal variation, and

informal writing styles. Traditional lexical representations such as TF-IDF remain useful baselines for social-media classification, while standardized tweet-classification benchmarks have increasingly emphasized transformer-based representations [6].

The introduction of transformer-based language models pretrained on Arabic corpora, such as ARBERT, MARBERT, and AraBERT, has significantly improved performance across a variety of Arabic Twitter analysis tasks [9], [10]. These models have demonstrated strong effectiveness in areas such as sentiment analysis, spam detection, sarcasm recognition, and offensive language classification [6], [11]. Despite these advances, relatively limited work has explored hashtag-centered analytical frameworks that jointly assess topical relevance and promotional content at the tweet level within Arabic hashtag streams.

D. Summary

Recent work on the evolution of cloud-to-edge digital applications further illustrates the broader movement toward intelligent, data-driven services and machine-learning-enabled analytics [15]. Overall, prior research has made substantial contributions to spam detection, topical analysis, and Arabic NLP. However, most existing studies either focus on discovering latent topics or detecting spam and promotional campaigns at the account or network level. Few works address the combined problem of assessing tweet-level topical relevance and detecting promotional content within individual hashtags, particularly in Arabic contexts. This study addresses this gap by proposing a hybrid, hashtag-centered NLP framework that integrates embedding-based topic relevance detection with supervised promotional-content classification to improve the reliability and interpretability of hashtag-based Twitter analysis.

III. MATERIALS AND METHODS

This section describes the dataset construction process, data collection strategy, preprocessing pipeline for Arabic tweets, and the feature representations used in this study. The objective is to enable tweet-level analysis of hashtag streams for 1) topic relevance assessment and 2) advertisement detection.

A. Dataset Collection

Arabic tweets were collected through the official X (formerly Twitter) API while the selected Arabic hashtags were active. The final archived research files supplied for this study contain 9,671 Arabic-language records spanning 9 October–15 December 2025 (UTC). The six streams cover technology, social issues, religious discussion, sports, and public-event topics. Their archived sizes are 501 tweets for #War_on_Drugs, 309 for #Data_Analysis, 2,162 for #Artificial_Intelligence, 1,908 for #World_Cup, 1,772 for #Friday, and 3,019 for #Study_Suspension. All records in these six files are marked with the Arabic language code extttar.

The archived files preserve tweet identifiers, timestamps, conversation identifiers, account identifiers, public profile metadata, and engagement fields. Exact duplicate checks on the supplied files identified 637 repeated TweetID occurrences and 817 repeated text occurrences across the combined archive; these fields make duplicate-aware filtering and user-aware

evaluation possible. Each record includes core fields provided by the API:

- Tweet and conversation identifiers,
- Tweet text content and timestamp,
- Public account/profile metadata and engagement fields when available.

Only publicly available posts were used. The archived research files retain the fields required for reproducibility, including TweetID, ConversationID, text, date, language, engagement counts, username, public profile location, verification status, biography, and follower/following statistics. Personally identifying account fields are not required by the proposed text-classification models and should not be redistributed in a public derived dataset.

The archive contains 4,433 distinct usernames; TweetID, username, and text therefore provide explicit keys for duplicate-aware and user-disjoint validation.

Fig. 1 presents English translations of the analyzed hashtags, and Fig. 2 shows the distribution of tweets across hashtags.

Hashtag in English	Hashtag in Arabic
#War_on_Drugs	#الحرب_على_المخدرات
#Data_Analysis	#تحليل_البيانات
#Artificial_Intelligence	#الذكاء_الاصطناعي
#World_Cup	#كأس_العالم
#Class_Suspension	#تعطيل_الدراسة
#Friday	#يوم_الجمعة

Fig. 1. English translation of the Arabic hashtags analyzed in this study.

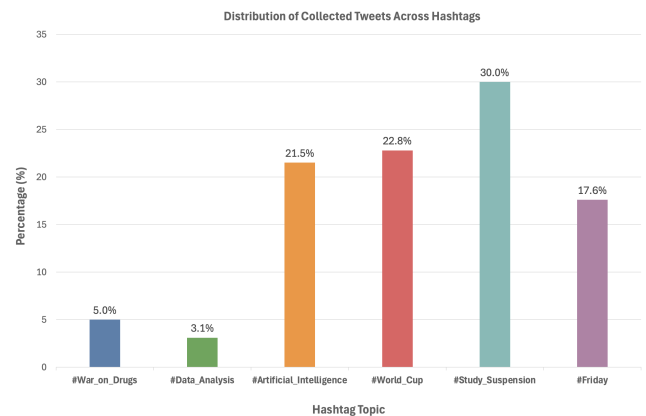


Fig. 2. Percentage distribution of tweets across the selected hashtags.

B. Data Preprocessing

Tweets are noisy and informal by nature (e.g., URLs, mentions, emojis, repeated characters). Arabic content further increases complexity due to rich morphology, dialectal variation, and non-standard spelling in social media. A multi-step preprocessing pipeline was applied to clean and normalize tweet text:

- Text cleaning: Removal of URLs, user mentions (@), emojis, hashtag symbols, and irrelevant special characters while preserving core semantic content.
- Arabic normalization: Unification of letter variants, removal of diacritics and tatweel, reduction of elongated/repeated characters, correction of common orthographic inconsistencies, and normalization of punctuation.
- Tokenization: Segmentation into tokens using Arabic-appropriate rule-based tokenization.
- Stop-word removal: Removal of frequent Arabic stop words using standard lists, supplemented with social-media filler terms when needed.

To quantify the effect of preprocessing, Fig. 3 compares average tweet length before vs. after preprocessing, showing a clear reduction due to removal of non-informative noise.

Hashtags	Avg. words before preprocessing	Avg. words after preprocessing	Noise reduction in text
#War_on_Drugs	18.58	11.12	%40.12
#Data_Analysis	30.93	16.33	%47.20
#Artificial_Intelligence	28.65	18.41	%35.75
#World_Cup	22.02	12.44	%43.52
#Class_Suspension	24.73	16.72	%32.40
#Friday	27.08	19.77	%27.02

Fig. 3. Average tweet length before vs. after preprocessing.

C. Exploratory Metadata Analysis

Exploratory analysis was performed on selected metadata fields to contextualize hashtag dynamics. User-provided location information was examined to capture broad participation patterns across regions, noting that many users omit location and are therefore treated as unknown. Additionally, user activity was analyzed by computing tweet frequency per user within each hashtag to assess whether content was concentrated among a small number of accounts.

During Arabic text visualization (e.g., word-frequency plots), standard plotting libraries may render Arabic characters incorrectly (disconnected or reversed) due to left-to-right layout. When such visualizations were used, Arabic reshaping was applied before plotting to ensure correct right-to-left rendering and improved readability.

D. Feature Representation

Two complementary feature representations were employed, aligned with the framework's two tasks:

- TF-IDF vectors: Extracted from the preprocessed tweet text (Clean Text) and used for supervised advertisement classification. This representation captures lexical cues and recurring promotional patterns.
- Transformer-based contextual embeddings: Contextual representations generated using AraBERT (aubmindlab/bert-base-arabertv02) and used to compute cosine similarity between tweet representations and predefined hashtag topic descriptions for relevance assessment. The archived CSV files do not contain cached embedding tensors;

unverifiable implementation details are therefore not asserted.

E. Topic Relevance Detection Module

Topic relevance detection evaluates whether a tweet aligns semantically with the intended meaning of its hashtag. For each hashtag, a concise topic description was manually defined to represent its semantic scope. AraBERT embeddings were computed for tweets and topic descriptions, and cosine similarity was used to quantify semantic relatedness. A fixed decision threshold, established during the original experimental analysis, classified tweets as *on-topic* or *off-topic*. The archived CSV files do not store similarity scores or threshold values; no unverified threshold is therefore reported.

The semantic relevance module was evaluated independently from advertisement classification: its on-topic/off-topic decision was based only on the cosine-similarity rule. Advertisement predictions were incorporated only in the final fused stream label, where promotional tweets were excluded from the organic on-topic subset. This separation preserves independent assessment of semantic relevance while retaining the practical filtering objective of the hybrid framework.

F. Advertisement Classification

Advertisement detection was formulated as supervised binary classification, independent of topic relevance detection. TF-IDF features were extracted from Clean Text, and a binary label Has_Ad was used as the target variable (1 for promotional content; 0 otherwise).

Prior to training, the dataset was split into training and testing subsets using stratified sampling to preserve label proportions under class imbalance. Five classifiers were evaluated:

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Naive Bayes
- Stochastic Gradient Descent (SGD) Classifier

This module outputs Has_Ad predictions that are later integrated into the final relevance labeling (Section IV).

G. Evaluation Metrics

Advertisement classification was evaluated using Accuracy, Precision, Recall, and F1-score:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$
$$\text{Precision} = \frac{TP}{TP + FP}, \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN}.$$

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

where, TP , TN , FP , and FN denote true positives, true negatives, false positives, and false negatives, respectively.

Topic relevance detection was assessed through the observed off-topic proportions and comparison with manual off-topic estimates reported in Section V. These manual estimates are used as a descriptive comparison rather than as a held-out accuracy benchmark; consequently, no unsupported accuracy statistic is reported for the relevance module.

IV. METHODOLOGY

This section presents the proposed hybrid NLP framework for analyzing hashtag-based streams on X. The framework combines semantic topic relevance detection using transformer embeddings with supervised advertisement classification using TF-IDF features. Fig. 4 illustrates the overall pipeline.

A. Overview of the Proposed Framework

The processing workflow consists of four stages:

- Hashtag-based collection and preprocessing: Tweets are collected per hashtag and standardized using the preprocessing pipeline described in Section III.
- Parallel feature generation: AraBERT embeddings are generated for topic relevance detection, and TF-IDF vectors are generated for advertisement classification.
- Topic relevance detection: Tweet embeddings are compared with the corresponding hashtag topic description embedding using cosine similarity and a threshold rule to classify tweets as on-topic or off-topic.
- Advertisement classification: A supervised model predicts whether a tweet is promotional using TF-IDF features and the Has_Ad label.

B. System Architecture

The pipeline branches into two parallel analysis paths after feature representation: 1) topic relevance detection based on semantic similarity and 2) advertisement classification based on supervised learning. The two branches are evaluated separately. Their outputs are combined only at the final filtering stage, where predicted advertisements are excluded from the organic on-topic subset.

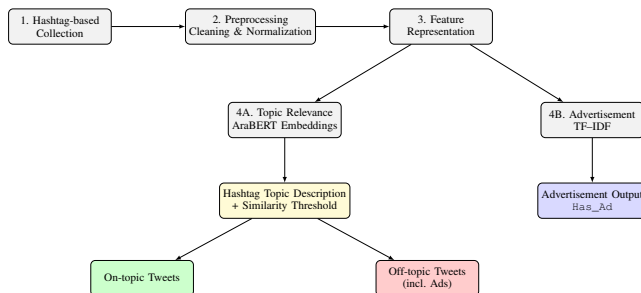


Fig. 4. Architecture of the proposed hybrid NLP framework for embedding-based topic relevance detection and supervised advertisement classification within hashtag-based streams on X.

C. Topic Relevance Detection

For each hashtag, a topic description is manually defined to represent the intended semantic scope. AraBERT embeddings are computed for tweets and topic descriptions, and cosine similarity is used to measure semantic alignment. Tweets are labeled as on-topic when their similarity exceeds the decision threshold and off-topic otherwise. This semantic decision is retained independently for evaluation. In the final fused output only, predicted advertisements are excluded from the organic on-topic subset so that advertisement-classifier errors do not alter the reported semantic-module decision.

D. Advertisement Classification

Advertisement detection is implemented as a binary decision module that flags whether a tweet contains promotional intent. Given a cleaned tweet, the classifier outputs a label $\text{Has_Ad} \in \{0, 1\}$, where $\text{Has_Ad} = 1$ denotes advertisement content. This output provides a direct estimate of promotional intrusion within each hashtag stream and is combined with the semantic label only in the final filtering stage. The advertisement prediction therefore does not overwrite the independently reported semantic relevance decision.

V. RESULTS AND DISCUSSION

This section reports the experimental results of the proposed hybrid framework, covering:

- Supervised advertisement detection using TF-IDF features and classical machine learning classifiers and
- Embedding-based topic relevance detection using AraBERT similarity scores.

In addition to per-hashtag model tables, we provide cross-hashtag summary visualizations and aggregated comparisons to highlight robustness, variability across topics, and practical implications for hashtag stream credibility.

A. Advertisement Classification Results

Advertisement detection was formulated as a binary classification task ($\text{Has_Ad} \in \{0, 1\}$), using TF-IDF features extracted from the cleaned tweet text. Performance was evaluated using Accuracy, Precision, Recall, and F1-score. Because promotional tweets are typically the minority class in several hashtags, Precision and Recall provide a more faithful assessment than Accuracy alone.

1) *Per-hashtag performance*: Tables I–VI summarize model performance for each hashtag. Overall, we observe that model behavior varies substantially across hashtags, reflecting differences in topic specificity, lexical patterns of promotions, and class imbalance.

TABLE I. ADVERTISEMENT CLASSIFICATION FOR #WAR_ON_DRUGS

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.940	0.142	1.00	0.25
SVM	0.950	0.166	1.00	0.28
Random Forest	0.950	0.166	1.00	0.28
SGDClassifier	0.950	0.166	1.00	0.28

Interpretation. The very high Recall indicates that most advertisements are detected, but low Precision implies more false positives. This is consistent with strong imbalance or overlap between promotional keywords and organic discussion terms in this stream.

TABLE II. ADVERTISEMENT CLASSIFICATION FOR #FRIDAY

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.960	0.444	0.666	0.533
SVM	0.932	0.285	0.666	0.400
Random Forest	0.923	0.222	0.500	0.307
Naive Bayes	0.977	1.000	0.333	0.500

Interpretation. Naive Bayes achieves extremely high Precision but low Recall, meaning it flags ads only when evidence is very strong, missing many subtle advertisements. Logistic Regression provides a better Precision–Recall tradeoff for this hashtag.

TABLE III. ADVERTISEMENT CLASSIFICATION FOR #DATA_ANALYSIS

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.887	0.692	0.750	0.720
SVM	0.903	0.750	0.750	0.750
Random Forest	0.935	1.000	0.667	0.800
Naive Bayes	0.903	1.000	0.500	0.667

Interpretation. Random Forest achieves the best F1-score, suggesting it captures non-linear combinations of TF–IDF signals that are characteristic of promotional tweets in a technical hashtag.

TABLE IV. ADVERTISEMENT CLASSIFICATION FOR #ARTIFICIAL_INTELLIGENCE

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.983	1.000	0.788	0.881
SVM	0.991	1.000	0.879	0.935
Random Forest	0.983	1.000	0.788	0.881
Naive Bayes	0.942	1.000	0.273	0.418

Interpretation. This hashtag yields consistently high Precision across models, indicating that advertisements contain strong lexical cues. SVM produces the highest F1-score by maintaining high Recall while preserving perfect Precision.

TABLE V. ADVERTISEMENT CLASSIFICATION FOR #WORLD_CUP

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.976	0.500	0.455	0.476
SVM	0.976	0.500	0.455	0.476
Random Forest	0.985	1.000	0.364	0.533
Naive Bayes	0.976	0.500	0.364	0.476

Interpretation. General-event hashtags often contain diverse language and high lexical overlap between organic posts and promotions, reducing separability. Random Forest achieves the best F1 primarily via higher Precision.

TABLE VI. ADVERTISEMENT CLASSIFICATION FOR #CLASS_SUSPENSION

Model	Accuracy	Precision	Recall	F1
Logistic Regression	0.900	0.770	0.750	0.760
SVM	0.890	0.750	0.750	0.750
Random Forest	0.920	0.920	0.710	0.800
Naive Bayes	0.920	0.940	0.660	0.770

Interpretation. Random Forest yields the strongest balance between Precision and Recall, suggesting that promotional tweets in this stream are detectable but require modeling interactions between multiple lexical indicators.

2) *Cross-hashtag summary (best model + robustness):* To summarize performance trends, Table VII reports the best-performing classifier per hashtag based on F1-score. Fig. 5 provides a compact visualization of the F1 variation across models and hashtags, and Fig. 6 highlights the Precision–Recall tradeoff, which is especially important under class imbalance.

TABLE VII. BEST ADVERTISEMENT CLASSIFIER PER HASHTAG BASED ON F1-SCORE

Hashtag	Best Model (F1)	F1-score
#War_on_Drugs	SVM / RF / SGD (tie)	0.28
#Friday	Logistic Regression	0.533
#Data_Analysis	Random Forest	0.800
#Artificial_Intelligence	SVM	0.935
#World_Cup	Random Forest	0.533
#Class_Suspension	Random Forest	0.800

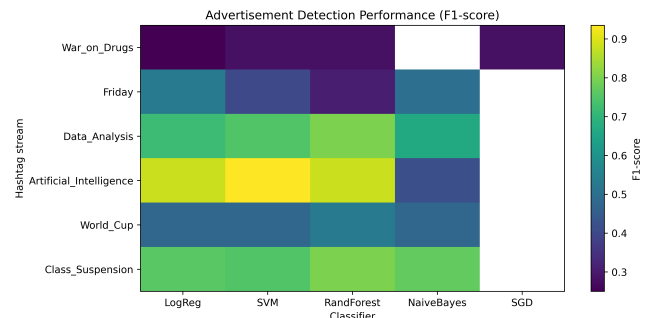


Fig. 5. F1-score heatmap for advertisement detection across hashtags and classifiers. The figure highlights that model effectiveness is hashtag-dependent, motivating per-hashtag model selection rather than a single universal classifier.

B. Topic Relevance Detection Results

Fig. 7 summarizes the best advertisement-detection F1-score for each hashtag and highlights the strong dependence of classifier performance on hashtag context.

Topic relevance detection was performed using AraBERT embeddings and cosine similarity between each tweet and a predefined hashtag topic description. The percentages reported here are produced by the semantic relevance branch before advertisement-based fusion. Table VIII compares manual off-topic estimation versus embedding-based detection.

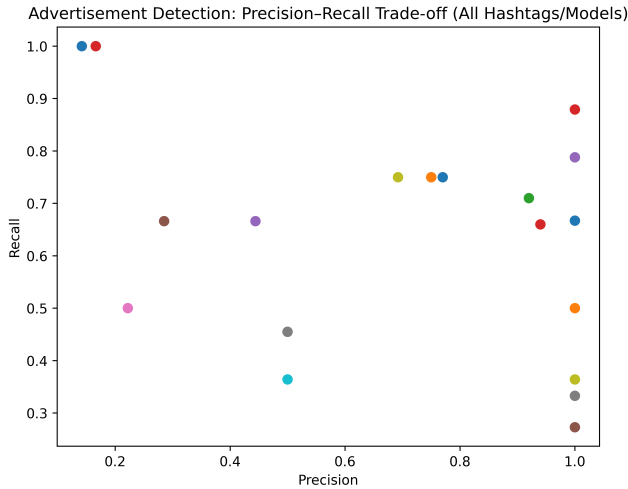


Fig. 6. Precision–Recall tradeoff across hashtags and models for advertisement detection. Models with high Recall but low Precision tend to over-flag promotions, while high Precision with low Recall indicates missing subtle advertisements.

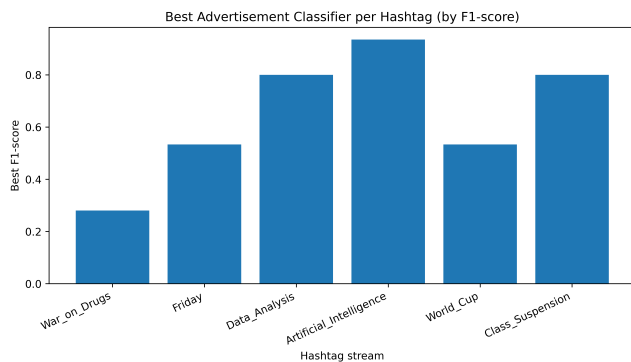


Fig. 7. Best F1-score achieved per hashtag for advertisement detection. The variability confirms that topic specificity and class balance strongly influence separability.

TABLE VIII. OFF-TOPIC DETECTION RESULTS (MANUAL VS. EMBEDDING-BASED MODEL)

Hashtag	Total Tweets	Off-topic (Manual %)	Off-topic (Model %)
#War_on_Drugs	501	2.60	3.40
#Data_Analysis	309	46.28	55.00
#World_Cup	2292	3.18	60.00
#Study_Suspension	3019	59.88	91.40
#Artificial_Intelligence	2162	9.34	89.00
#Friday	1772	10.80	67.00

Interpretation. The embedding-based model often reports higher off-topic proportions than manual inspection, especially for highly active hashtags. This suggests that semantic similarity filtering captures subtle drift and unrelated conversational branches that may not be counted as spam/ads during manual checks. At the same time, the gap indicates that similarity thresholds can be conservative, which motivates future calibration using larger annotated relevance datasets.

C. Comparative Discussion and Key Findings

Fig. 8 and 9 complement Table VIII by visualizing the manual–model off-topic percentages and their differences across hashtags.

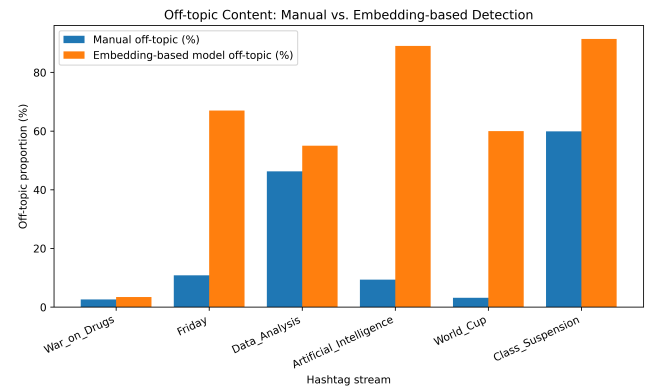


Fig. 8. Manual vs. embedding-based off-topic percentage across hashtags. The embedding-based method generally applies a stricter semantic criterion, yielding higher off-topic rates in high-activity streams.

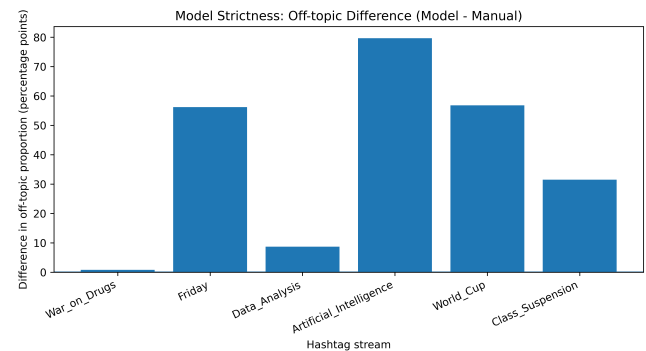


Fig. 9. Difference between embedding-based and manual off-topic estimates (Model% – Manual%). Large positive gaps indicate streams where semantic drift is strong or where manual inspection underestimates off-topic diversity.

Across tasks, the results support three consistent observations:

1) *Hashtag dependency is substantial:* Advertisement detection performance varies widely between technical hashtags (more consistent promotional language) and broad event hashtags (higher linguistic diversity and overlap). This justifies evaluating models per hashtag rather than reporting a single global score.

2) *Precision–recall tradeoffs reveal practical behavior under imbalance:* Several models achieve high Accuracy while yielding weak Precision, meaning they may over-flag ads when the positive class is rare. Therefore, Precision, Recall, and F1-score provide more informative signals for model selection.

3) *Embedding-based relevance detection is sensitive but can be conservative:* The topic relevance module captures semantic divergence beyond explicit advertisements/spam, improving stream credibility filtering. However, the model may label a larger fraction of tweets as off-topic in high-activity

streams, indicating that threshold calibration and richer ground truth labels would further improve reliability.

In summary, combining TF-IDF-based supervised detection for promotional intent with AraBERT-based semantic relevance filtering provides a practical and interpretable framework for improving the trustworthiness of trending hashtag streams, especially in Arabic contexts.

VI. CONCLUSION

This study presented a hybrid NLP framework for improving the reliability of trending hashtag streams on Twitter/X, with a focus on Arabic-language content. The proposed approach integrates two complementary components: 1) embedding-based topic relevance detection, which measures semantic alignment between tweets and the intended hashtag meaning using AraBERT representations, and 2) supervised advertisement detection using TF-IDF features with classical machine learning classifiers. By separating *semantic relevance* from *promotional intent*, the framework provides clearer interpretability of hashtag dynamics and supports fine-grained filtering of noisy content while preserving meaningful organic discussion.

Experimental results across multiple trending hashtags indicate that advertisement detection is generally stronger in domain-specific hashtags (e.g., technology-oriented streams), where promotional language is more consistent and lexically separable, while broader event-oriented hashtags exhibit greater linguistic diversity and lexical overlap between organic and promotional tweets, leading to more variable performance. In parallel, the topic relevance module often reports higher off-topic proportions than manual inspection in highly active hashtags, suggesting that contextual embeddings can capture subtle semantic drift beyond surface keyword matching. Manual off-topic estimates were retained as a descriptive comparison with the embedding-based outputs rather than being presented as an independent accuracy benchmark.

Overall, the findings confirm that combining contextual semantic representations with lightweight supervised classifiers is an effective and scalable strategy for analyzing Arabic hashtag streams and mitigating promotional intrusion. Future improvements will focus on expanding and refining manual annotation to enable stronger statistical evaluation, replacing fixed similarity thresholds with learning-based relevance models, and exploring transformer-based or hybrid classifiers for advertisement detection to better capture implicit promotional intent. Additional extensions include imbalance-aware calibration to reduce false positives when ads are rare, and incorporating temporal and metadata signals (e.g., activity bursts, repeated posting patterns) to better distinguish coordinated campaigns from organic engagement, as well as validating the framework in multilingual settings and across other social media platforms.

ACKNOWLEDGMENT

The authors extend their appreciation to the Deanship of Scientific Research at Northern Border University, Arar, KSA

for funding this research work through the project number "NBU-FFRGS-2026".

REFERENCES

- [1] P. Gundecha and H. Liu, "Mining social media: A brief introduction," in *2012 TutORials in Operations Research: New Directions in Informatics, Optimization, Logistics, and Production*, INFORMS, 2012, pp. 1–17, doi: 10.1287/educ.1120.0105.
- [2] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*. Cambridge, U.K.: Cambridge University Press, 2014, doi: 10.1017/CBO9781139088510.
- [3] S. Sedhai and A. Sun, "An analysis of 14 million tweets on hashtag-oriented spamming," *Journal of the Association for Information Science and Technology*, vol. 68, no. 7, pp. 1638–1651, 2017, doi: 10.1002/asi.23836.
- [4] S. Sedhai and A. Sun, "HSpam14: A collection of 14 million tweets for hashtag-oriented spam research," in *Proc. 38th Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 2015, pp. 223–232, doi: 10.1145/2766462.2767701.
- [5] X. Zhang, Z. Li, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in Twitter," *ACM Transactions on the Web*, vol. 10, no. 1, Art. 4, pp. 1–28, 2016, doi: 10.1145/2846102.
- [6] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1644–1650, doi: 10.18653/v1/2020.findings-emnlp.148.
- [7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. 2019 Conf. North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.
- [8] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. 2019 Conf. Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. Natural Language Processing*, 2019, pp. 3982–3992, doi: 10.18653/v1/D19-1410.
- [9] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based model for Arabic language understanding," in *Proc. 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, 2020, pp. 9–15, doi: 10.48550/arXiv.2003.00104.
- [10] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep bidirectional transformers for Arabic," in *Proc. 59th Annual Meeting of the Association for Computational Linguistics and 11th Int. Joint Conf. Natural Language Processing*, 2021, pp. 7088–7105, doi: 10.18653/v1/2021.acl-long.551.
- [11] A. Wadhawan, "AraBERT and Farasa segmentation based approach for sarcasm and sentiment detection in Arabic tweets," in *Proc. Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 395–400, doi: 10.48550/arXiv.2103.01679.
- [12] S. Sedhai and A. Sun, "Effect of spam on hashtag recommendation for tweets," in *Proc. 25th Int. Conf. World Wide Web Companion*, 2016, doi: 10.1145/2872518.2889404.
- [13] L. Hong and B. D. Davison, "Empirical study of topic modeling in Twitter," in *Proc. First Workshop on Social Media Analytics*, 2010, pp. 80–88, doi: 10.1145/1964858.1964870.
- [14] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in Twitter: A graph-based hashtag sentiment classification approach," in *Proc. 20th ACM Int. Conf. Information and Knowledge Management*, 2011, pp. 1031–1040, doi: 10.1145/2063576.2063726.
- [15] S. M. Altowaijri and M. Ayari, "The synergistic impact of 5G on cloud-to-edge computing and the evolution of digital applications," *Mathematics*, vol. 13, no. 16, Art. 2634, 2025, doi: 10.3390/math13162634.