AUIQ Technical Engineering Science

Volume 2 | Issue 3 Article 8

2025

Predicting the Next Trending Hussaini Poems: A Machine Learning Approach Based on Textual and Engagement Features

Elaf Adel Abbas
College of Computer Science and Information Technology, University of Karbala, elaf.adil@uokerbala.edu.iq

Recommended Citation

Abbas, Elaf Adel (2025) "Predicting the Next Trending Hussaini Poems: A Machine Learning Approach Based on Textual and Engagement Features," *AUIQ Technical Engineering Science*: Vol. 2: Iss. 3, Article 8.
DOI: https://doi.org/10.70645/3078-3437.1049

AUIQ TECHNICAL ENGINEERING SCIENCE 2025;2:117–128

ORIGINAL STUDY

Predicting the Next Trending Hussaini Poems: A Machine Learning Approach Based on Textual and Engagement Features

Elaf Adel Abbas

College of Computer Science and Information Technology, University of Karbala

ABSTRACT

Hussaini's poem videos have garnered substantial viewership on digital platforms; nevertheless, the determinants of why just a select number achieve "trending" status remain inadequately comprehended. This study presents a machine learning approach aimed at forecasting the trending potential of Hussaini poetry videos through the integration of text- and interaction-based variables. A dataset including 270 YouTube videos was generated, encompassing each video's title and audience engagement parameters, including views, likes, comments, duration, and upload time. Textual attributes were converted using TF-IDF representations, whereas numerical attributes were standardized and amalgamated via a unified ColumnTransformer pipeline. Five machine learning classifiers—Naive Bayes, Logistic Regression, Linear SVC, Random Forest, and XGBoost—were trained and assessed by cross-validation. The findings indicated that ensemble models yielded superior performance, with Random Forest attaining the highest accuracy (0.86) and F1-score ($\sim$0.85) in identifying trending poems, followed by XGBoost (F1 $\sim$ 0.83). Feature-importance analysis indicated that engagement metrics, specifically the quantity of views, likes, and shares, were the most significant predictors of virality, although title keywords and video time contributed somewhat. The results indicate that integrating linguistic cues with engagement behavior facilitates accurate predictions of digital virality. The suggested framework exemplifies one of the initial computational methodologies for modeling the online transmission of Hussaini poetry, providing both methodological and cultural contributions through the integration of machine learning and digital studies of Islamic and Arabic heritage.

Keywords: Trend forecasting, Social media analytics, Digital humanities, Cultural content prediction, Hussaini poetry

1. Introduction

Cultural and religious expressions now have more chances than ever to reach large audiences because of the proliferation of digital media [1]. The proliferation of digital content in religious and cultural domains, particularly Hussaini poetry, has created a rich repository of expressive media that combines spiritual themes with artistic performance. Specifically, Hussaini's poetry, which is fundamental to remembering the sorrow of Karbala, is making its way across the internet via sites like Facebook and YouTube [2]. In the same way that popular music and videos battle for attention and views, these poems help to maintain cultural identity in the age of social media. Cultural scholars, media analysts, and religious content creators are vested in knowing which poetry will be "trending" soon [3]. These poetic recitations, typically disseminated through video formats, contain multifaceted characteristics encompassing vocal delivery, linguistic content, visual elements, and emotional intensity. The phenomenon of trending content has been extensively examined in the realms of popular music, internet news, and viral marketing [4]. While traditional analysis of Hussaini media has focused predominantly on textual and theological aspects, the rapid expansion of digital platforms has generated vast collections of video-recorded

Received 16 October 2025; revised 29 October 2025; accepted 30 October 2025.
Available online 08 November 2025

E-mail address: elaf.adil@uokerbala.edu.iq (E. A. Abbas).

https://doi.org/10.70645/3078-3437.1049

poetic performances that remain underexplored through computational methods. Contemporary research frequently relies on machine learning methodologies that utilize textual, visual, and engagement attributes to forecast audience behavior. Nevertheless, there has been no research conducted on Hussaini's poetry, particularly within the framework of Arabic cultural media [5]. This distinction highlights the necessity of creating future models that align with this domain, integrating both textual components (e.g., poem titles) and performance measures (e.g., visuals, selections, comments, length).

The development of multimedia analysis techniques offers novel prospects for comprehending the intricate nature of Hussaini's poetry videos. These digital artifacts include the poem's semantic content and its performative aspects, such as recitation style, visual symbolism, and audience participation patterns [6]. Prior studies in multimedia analysis have predominantly focused on conventional video genres, with less exploration of religious and cultural material, especially within Arabic and Islamic frameworks. This distinction is significant due to the distinctive features of Hussaini's poetry performances, which incorporate particular rhetorical frameworks, musical components, and visual metaphors that set them apart from other video genres.

Recent methodologies in video analysis generally emphasize either content-based attributes (such as color, texture, and motion) or textual/audio components in isolation [7], failing to establish a cohesive framework that accommodates the distinctively multimodal characteristics of religious poetry texts. The lack of distinct methodologies for analyzing Hussaini media is a significant research void, particularly considering the cultural significance and increasing digital prominence of this material category.

This investigation resolves these constraints by creating a comprehensive computational framework that integrates machine learning classification and feature extraction to analyze Hussaini poetry videos. Our research utilizes sophisticated video processing techniques to extricate the visual characteristics (visual structure, symbolic elements) and audio-textual features (vocal patterns, poetic structure, emotional tone) that characterize this distinctive genre. By employing a multi-modal analysis approach, we can identify the fundamental characteristics that differentiate Hussaini's poetry performances from other forms of religious and cultural material.

This study offers three primary contributions:

- A bi-modal analytical methodology is proposed that amalgamates textual features (poem titles via TF-IDF) and engagement-based metrics (views, likes, comments, duration).
- Offers a practical framework for evaluating the digital virality of religious and cultural poetry videos.
- Feature importance analysis identifies the key aspects influencing a poem's popularity.
- Provides the initial computer-based framework for mimicking Hussaini poetry recitals in online media.
- Integrates machine learning with digital humanities to aid Islamic cultural legacy analysis and preservation.
- Contributes to the improvement of Arabic NLP and cultural media analysis studies.

2. Research objectives

The main objective of the present study is to develop a predictive framework to identify trending Hussaini poetry videos on digital platforms using computational modeling and cultural media analysis.

The specific objectives are as follows:

➢ Develop and validate a machine learning model that predicts whether Hussaini poetry videos will trend based on textual and engagement-related features. (main objective)
➢ Compare the performance of multiple classifiers (Naive Bayes, Logistic Regression, Linear SVC, Random Forest, and XGBoost) to determine the most effective trend detection approach.
➢ Identify and analyze the most significant textual and engagement-related factors that facilitate virality in religious and cultural contexts.
➢ Apply computational tools to Arabic cultural media to advance digital humanities and Arabic NLP research.

3. Research questions

RQ1: Which textual and interaction-based characteristics most effectively distinguish trending from non-trending Hussaini poems?

RQ2: Which machine learning models exhibit the highest predictive accuracy in identifying prevalent religious content?

RQ3: In what ways can computational modeling facilitate the comprehension and preservation of digital media that are cultural or religious in nature?

4. Related work

This section reviews prior studies relevant to the prediction of online content virality and its cultural dimensions. To provide a clear conceptual flow, the literature is categorized into three thematic directions:

Trend and Popularity Prediction in Online Content – summarizing previous works on computational models that forecast digital engagement and virality in general media contexts.

Machine Learning for Cultural and Religious Content – highlighting research efforts that apply artificial intelligence and natural language processing techniques to cultural, spiritual, or religious datasets, particularly within Arabic and Islamic domains.

Novelty of the Present Study – outlining how the current work bridges these two research domains by introducing a Hussaini-poetry–specific dataset and a hybrid linguistic–engagement modeling approach. The following subsections discuss these themes in detail.

4.1. Trend and popularity prediction in online content

Machine learning techniques have been widely applied to predict content popularity and trending items on social media platforms. For example, researchers have used algorithms such as support vector machines (SVM), logistic regression, and decision trees to build classifiers to determine whether a YouTube video will be labeled as "trending" [8]. Such models leverage both content features and interaction metrics to predict virality. In one study, a survey of web video popularity prediction methods identified the integration of textual features (such as video titles or descriptions) with metadata as a promising approach [9]. Their case study combined engineered features (views, upload times, etc.) with text embeddings, and a Random Forest classifier achieved approximately 87% accuracy in predicting which videos get the top audience attention. Other work similarly reports that ensemble models can perform well: for example, combining audio and visual features of songs on YouTube with Random Forest improved hit-song classification accuracy by 82% [10]. Previous studies have also experimented with gradient-boosted trees (XGBoost) on content features – an early hit-song predictor using XGBoost achieved about 71% accuracy using acoustic features. In general, incorporating diverse feature types (textual content, numerical engagement data, and even visual cues) and using hybrid models (e.g., combining TF-IDF text vectors with classification algorithms such as Random Forest or XGBoost) has proven effective for popularity prediction tasks.

On platforms such as Twitter and TikTok, the virality of content has been studied through both content-based and network-based features. Mahdikhani (2022) [11] analyzed over one million tweets, extracting topic models and TF-IDF features from tweet texts to estimate retweet popularity. Their findings showed that emotionally driven tweets were more likely to go viral than purely informational tweets, highlighting the importance of emotion in predicting popularity. In the realm of hashtags, Suthanthiradevi and Geetha (2020) in [12] developed a system to classify the future popularity of Twitter hashtags using two types of features: content features (textual properties of the hashtag's tweets) and contextual features (user networks and temporal trends). Their evaluation showed that relevant social-network features were far more predictive – with decision tree classifiers providing approximately 94% accuracy – while content-only models performed worse. Interestingly, combining content and context did not improve performance much in that study, suggesting that social diffusion characteristics dominated the prediction of trend longevity. Nonetheless, many studies emphasize a multi-model approach. A recent model for short video apps (such as YouTube Shorts or TikTok) has introduced an attention-based multi-modal network that incorporates full video frames and metadata to predict audience popularity. By capturing intra- and inter-modal patterns (visual, textual, and potentially audio), this model outperformed baseline machine learning algorithms in predicting which short clips would achieve high popularity [13]. Overall, prior work on trend forecasting spans a range of content domains (videos, social posts, songs), and typically leverages feature fusion (e.g. textual TF-IDF with numerical interaction achievement).

4.2. Machine learning for cultural and religious content

Concurrently with broad social media analytics, scholars have investigated computational techniques for Islamic and Arabic cultural content. Natural language processing has been utilized on sacred texts; for instance, efforts have been undertaken to categorize Quranic passages employing conventional classifiers. Elmitwalli and Alsayet (2020) developed a multi-class model for the automatic classification of Quranic verses by chapter (surah) utilizing Support Vector Machine (SVM) and Naive Bayes, attaining approximately 80% accuracy with SVM on a selected selection of chapters [14]. These endeavors illustrate the viability of machine learning in identifying linguistic patterns within religious literature. Recent surveys underscore the extensive range of machine learning applications in Quranic studies, encompassing text categorization and audio analysis. Iqbal and Hasan (2024) examine numerous studies in which algorithms like k-nearest neighbors, neural networks, and support vector machines have been employed for the classification of Quranic text and the automated examination of Tajweed (pronunciation rules) [15]. These encompass systems designed to aid users in identifying and comprehending genres, emphasizing that machine learning may derive significant aspects from Arabic religious material beyond mere text labeling. Likewise, computational analysis has been noted in the domain of hadith (prophetic traditions); for instance, scholars have explored machine learning techniques for hadith authentication and classification, and recent reviews in the ACM TALP documentation have recorded advancements in text classification methodologies to assess hadith reliability. This collection of studies on Islamic texts illustrates the increasing interrelation of AI and cultural heritage organizations [16].

Another relevant area is the computational analysis of Arabic poetry and spiritual literature. Arabic poetry is a cultural art form and, in religious contexts (such as Hussaini elegy), a medium of devotion. NLP techniques have advanced rapidly in line with Arabic poetry. Qarah (2024) introduced AraPoemBert [17], a language model trained on millions of verses of Arabic poetry, which achieved state-of-the-art results on several analysis tasks. Impressively, the model can classify the rhythmic meter of a poem with approximately 99% accuracy and can even perform emotional analysis of poetic stanzas (identifying emotional tone) with approximately 79% accuracy. Other tasks, such as identifying the gender of the poet or the rhyme scheme of a poem, were also solved with high accuracy. These results illustrate how specific models can capture the rich linguistic patterns in Arabic poetic and religious texts. There have also been attempts to predict trends in cultural media – for example, using social media indicators to forecast the popularity of nasheeds (Islamic devotional songs) or other religious content – although such studies are still limited in number. Overall, prior literature demonstrates the application of ML and NLP in a variety of Islamic content: from Quran and Hadith classification to sentiment analysis in Arabic text and even automatic recognition of poetic structures.

4.3. Novelty of the present study

The convergence of the two research threads discussed above – content popularity prediction and computational analysis of Islamic cultural texts – defines the unique contribution of our work. Previous studies have modeled viral trends in general media content, and isolated works have applied machine learning to analyze religious literature, but to our knowledge, no study has specifically predicted popularity for Hussaini poems (devotional poetry honoring Imam Hussein). Unlike earlier viral-content predictors, we focus on the culturally significant genre of Arabic poetry and leverage both textual characteristics and engagement metrics to predict which Hussaini poems will trend. This blend of popularity analysis with cultural content modeling is unique. By building on the techniques of social media trend forecasting and incorporating domain-specific text features (e.g., poetry lyrics, themes), our study fills a gap in the literature. This extends popularity prediction into the realm of religious and heritage content, providing new insights that were not addressed in earlier works on trending multimedia or computational Islamic studies. The approach and findings presented here thus break new ground at the intersection of machine learning, social trend prediction, and cultural/religious text analysis.

5. Methodology

5.1. Data acquisition

We assembled a collection of Hussaini poetry videos from YouTube. Every video in the collection was meticulously annotated with a binary target variable, trend, denoting whether the poem was trending (1) or not trending (0) at the time of observation. The trend label was determined through an expert-based manual review rather than automated metrics. A domain specialist with long-standing experience in Hussaini media evaluated each video and judged its cultural virality based on prior audience response, recitation circulation, and its recognition across major platforms. This human-guided process ensured that the trend annotation reflected the contextual and cultural aspects of virality, which cannot be captured by engagement counts or algorithmic trending lists alone. For each video, we gathered an extensive array of attributes that encapsulate content and popularity measures, comprising both textual metadata and quantitative engagement statistics:

Textual metadata: Title of the poem, name of the narrator, and name of the poet (as provided on the platform).

Engagement indicators include the quantity of video views, likes, comments, and shares. An external share proxy. A non-native sharing signal was acquired from public third-party websites that embed or link to each video; this optional function may be inaccessible for certain users. Temporal characteristics: video length (in minutes/seconds) and publication time (upload timestamp).

The dataset used in this research includes 270 Hussaini poetry videos collected from 2020 to 2023 through manual curation from official YouTube channels. Each video was reviewed and annotated for its virality status. Due to the manual labeling process and the need to ensure content authenticity, the dataset size was intentionally limited. Consequently, the present work should be considered a pilot study aimed at verifying the feasibility of the proposed pipeline before scaling up to larger datasets.

5.2. Data Preprocessing

Before modeling, we applied several preprocessing steps to clean and transform the raw data. Missing values were addressed by removing entries with missing target labels and imputing missing feature values where possible, to ensure that no information was lost due to zero entries. We also standardized text and categorical entries by fixing inconsistent formatting (for example, ensuring consistent spelling/casing of reader and poet names). Temporal characteristics were extracted from raw data fields to furnish additional model-compatible information. The video's duration, initially provided as a timestamp in hh: mm: ss or mm: ss format, was parsed and transformed into video_minutes, a numerical characteristic denoting the video's length in minutes. The publishing time of each video was segmented into components, including publication year, month, and day of the year, allowing the model to identify seasonal or annual trends. A video released in Muharram of a specific year may exhibit distinct trending behavior compared to others, rendering that pattern discernible. Ultimately, we transformed the target column trend into a binary numeric format (0/1) for application in machine learning techniques. All non-standard inputs for trend (e.g., "yes," "trending," etc.) were standardized to 1, while "no"/"not trending" were standardized to 0, so establishing a clear binary aim.

5.3. Feature engineering

Using feature engineering, we transformed numerical and textual data streams into a modelable format. The Term Frequency-Inverse Document Frequency (TF-IDF) representation was used to convert textual features, which are the names of poems, into numerical vectors. This method extracts the relative relevance of each word in a collection of poem titles by transforming the collection into a TF-IDF feature matrix. To improve the titles' text features before vectorization, we did some basic text preparation, such as eliminating punctuation, normalizing Arabic letters (by doing things like removing diacritics), and removing frequent stop words. In this way, we know that the TF-IDF model paid close attention to the significant terms utilized in the poems' titles. The classifier was given a quantitative representation of the title's content by the resultant high-dimensional TF-IDF feature space, which included each unique word as a feature. To highlight the influence of linguistic content, a TF-IDF analysis was performed on poem titles. The results revealed that emotionally charged and spiritually themed words are strongly correlated with virality. Table 1 lists representative high-ranking keywords, showing that emotionally evocative expressions such as "*زينب (Zainab)*", "*الحسين (Al-Hussain)*", "*كربلاء (Karbala)*", and "*الشهادة (Martyrdom)*" often appear in trending poems. These terms reflect deep cultural and religious resonance, eliciting affective engagement that enhances audience interaction and the probability of virality.

Table 1. High-weighted emotional keywords identified by TF-IDF.

Rank	Keyword	TF-IDF Weight	Emotional Category
1	زينب (Zainab)	0.087	Reverence/Compassion
2	الحسين (Al-Hussain)	0.081	Devotion/Loyalty
3	كربلاء (Karbala)	0.074	Tragedy/Memory
4	الشهادة (Martyrdom)	0.071	Sacrifice/Faith
5	العباس (Al-Abbas)	0.067	Courage/Brotherhood

For numerical features (views, likes, comments, shares, duration, and publication date parts), we performed scaling and modeling as needed. All numerical features were standardized (for example, through z-score normalization using standard scalars) so that they had zero mean and unit variance. This step prevents features with larger numerical ranges (such as scene numbers) from dominating features with smaller ranges. Categorical fields such as reader or poet names (if used in modeling) will be encoded in an appropriate numerical form (e.g., one-hot encoding), although the primary emphasis in our experiments was on title text and numerical engagement metrics.

We combined textual and numerical features into a single feature space using the ColumnTransformer pipeline. This approach allowed us to preprocess each feature type separately and then merge them efficiently. Poem titles were transformed into TF-IDF vectors to capture important words, while numerical features (such as views, likes, and duration) were plotted and scaled for consistency. ColumnTransformer then combined both outputs into a single feature matrix used for model training. This method preserved the strengths of each feature type, keeping the TF-IDF text sparse and numerical data normalized providing a balanced input to the classifier. Fig. 1 illustrates the overall workflow of the proposed machine learning framework, from data acquisition to model evaluation.

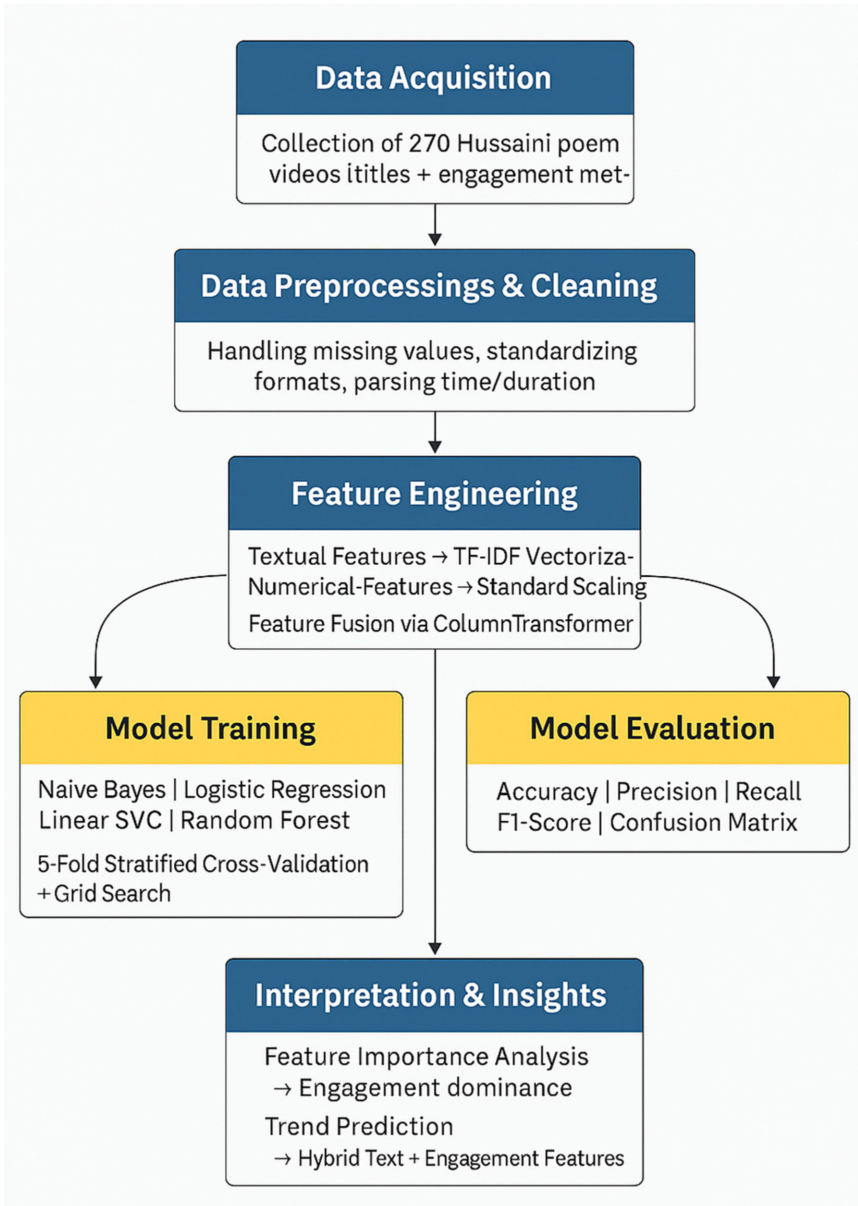

Fig. 1. Proposed machine learning framework for trend prediction of hussaini poems.

Modeling

With the feature set prepared, we evaluated a suite of machine learning algorithms for a binary classification task to predict whether a poetry video is trending or not. We considered five different classifiers, representing a mix of linear, probabilistic, and combinatorial methods commonly used in text and metadata classification:

Multinomial Naive Bayes (NB): A probabilistic classifier suitable for discrete features such as word count/TF-IDF, often used as a baseline for text classification [18].

Logistic Regression (LR): A linear model that outputs probabilities for two classes with L2 regularization to prevent overfitting [19].

Linear Support Vector Classifier (Linear SVC): A support vector machine with a linear kernel, which is effective for high-dimensional feature spaces such as text TF-IDF vectors [20].

Random Forest: A set of decision trees that bootstrap samples and features, providing robust predictions and the ability to handle non-linear relationships as well as feature importance estimates [21].

All models were implemented using the scikit-learn library (except XGBoost, which used the XGBoost

Python package). To ensure fair comparison, each classifier was trained and evaluated on the same stratified 5-fold splits with identical feature inputs. We performed hyperparameter tuning via GridSearchCV. All preprocessing (TF-IDF vectorization and numerical scaling) was performed *within each cross-validation fold* using a single scikit-learn pipeline to prevent data leakage. Stratification preserved the class distribution of trending vs. non-trending videos in every fold, which is important under potential class imbalance. During grid search, candidate configurations were evaluated on the training folds and the best hyperparameters were selected based on validation performance.

5.4. Model evaluation

To objectively evaluate model performance, four common classification metrics were employed: accuracy, precision, recall, and F1-score. Each provides a different perspective on the model's ability to correctly identify trending poems (positive class) versus non-trending (negative class).

1- Accuracy

Accuracy measures the overall proportion of correctly classified examples, both trending and non-trending.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP (True Positives) are trending poems predicted correctly, TN (True Negatives) are non-trending poems predicted correctly, FP (False Positives) are non-trending poems predicted as trending, and FN (False Negatives) are trending poems that are missed by the model. Accuracy gives a general overview of performance, but can be misleading if the dataset is unbalanced.

2- Precision

Precision determines how many of the items predicted as trending are actually trending.

$$Precision = \frac{TP}{TP + FP}$$

High precision means that the model makes few false-positive errors (i.e., it rarely predicts that a poem is trending when it is not).

3- Recall

Recall how many actual trending poems have been successfully modeled

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

High recall indicates that the model captures most of the trending poems but does not take into account false positives. This metric is especially important when missing a trending item is more costly than generating a few false alarms.

4- F1-Score

The F1-score is the harmonic mean of precision and recall, which balances both types of errors.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

A higher F1-score reflects a better trade-off between identifying all trending poems (remember) and ensuring that predictions are reliable (accurate).

Although the F1-score was used as the primary evaluation metric, other measures – accuracy, precision, and recall – were also reported to provide a more comprehensive assessment of model behavior. Accuracy offers an overall view of how many predictions were correct, but it can be biased if the dataset is unbalanced (as in our case, where non-trending rhymes are more frequent). Precision is important to evaluate the reliability of the model, ensuring that predictions labeled as "trending" actually represent trending poems. In contrast, recall reflects the model's ability to capture all actual trending poems, which prevents important cases from being missed.

6. Results and discussion

6.1. Model performance comparison

Classification results (Table 2) reveal a consistent trend across all performance metrics: ensemble-based classifiers significantly outperformed simple linear and probabilistic models in predicting the virality of Hussaini poems. Among all the models tested, Random Forest achieved the best overall balance in the four main metrics—Accuracy (0.8629), Precision (0.8461), Recall (0.8610), and F1-score (0.8503).

These values indicate that Random Forest not only made correct predictions overall (high precision), but also managed to correctly identify most of the trending poems (high recall) while maintaining low false positives (high precision). The close alignment

Table 2. The model's results.

Model	Accuracy	Precision	Recall	F1 (trend = 1)
Random Forest	0.8629	0.8461	0.8610	0.8503
XGBoost	0.8481	0.8334	0.8377	0.8325
Linear SVC	0.7296	0.7726	0.5847	0.6515
Logistic Regression	0.6815	0.6972	0.5350	0.5942
Naïve Bayes	0.7222	0.8739	0.4623	0.5910

between its precision and recall indicates that the model is well-calibrated, able to distinguish trending content without overfitting in any class. The XGBoost model followed closely with accuracy = 0.8481, precision = 0.8334, recall = 0.8377, and F1 = 0.8325. This performance supports the effectiveness of gradient boosting ensembles in handling heterogeneous data types – textual features (TF-IDF representation of titles) combined with numerical engagement indicators (views, likes, comments, etc.). Both ensemble models were trained using class-balanced weights, which reduced the bias towards the majority class (non-trending poems) and allowed the models to maintain strong recall on the minority class (trending poems). Their consistency across all four metrics shows that they have successfully captured the non-linear dependency between linguistic expression and audience engagement – factors that drive digital virality in real-world platforms. In contrast, linear classifiers such as logistic regression and LinearSVC (Support Vector Machine) showed weaker and less balanced behavior. LinearSVC achieved moderate precision (0.7726) but remarkably low recall (0.5847), meaning that although it was selective (few false positives), it actually failed to detect many trending poems. Its overall accuracy (0.7296) and F1-score (0.6515) reinforce this trade-off, as the model favors precision at the expense of recall. Similarly, logistic regression performed even worse (accuracy = 0.6815, precision = 0.6972, recall = 0.5350, F1 = 0.5942), confirming that linear decision boundaries are inadequate to represent the complex interplay between text semantics and engagement patterns in this dataset. These findings align with established results in social media analytics, where linear models often struggle to capture the nonlinear, feature-interactive dynamics present in viral prediction tasks.

The Naive Bayes classifier displayed an interesting but limited behavior. Despite achieving the highest precision (0.8739) among all models, it suffered from the lowest recall (0.4623) and a marginal precision of 0.7222, generating an F1-score of 0.5910. This discrepancy highlights the conservatism of the model: it was extremely cautious in labeling poems as trending, predicting only a few instances as trend = 1. When it did this, those predictions were often correct (hence high precision), but it failed to identify most of the cases that were actually going on. Such a pattern is typical for probabilistic models that assume feature independence, which is unrealistic for this dataset, where textual and interaction features interact heavily.

Comparing all metrics, several insights emerge:

While accuracy does a good job of capturing the overall precision of predictions, it masks any disparities between classes. Random Forest and XGBoost both proved to be quite resilient by attaining the top accuracy.

Precision evaluates the reliability of the prediction – here, Naive Bayes performed excellently, but at the cost of low coverage, while Random Forest maintained high precision with equally strong recall, achieving the best overall balance.

Recalls reveal sensitivity to the trending class, where aggregate models again dominated, effectively capturing rare but significant positive cases.

The F1-score, as the harmonic mean of precision and recall, summarizes the trade-off between the two and confirms that Random Forest provided the most stable and effective predictive performance.

Overall, the results provide clear evidence that ensemble classifiers, particularly Random Forest and XGBoost, are superior to this task. Their ability to integrate multiple feature interactions and resist overfitting provides high and balanced scores across all evaluation dimensions. The performance difference — where Random Forest's F1-score is about 26 percentage points higher than Naive Bayes is not only statistically significant, but also practically meaningful for trend forecasting applications. These findings reinforce that the relationship between linguistic cues, user engagement, and digital virality is highly non-linear and context-dependent, and that robust, ensemble-based learning methods are best suited for modeling such cultural and social phenomena.

6.2. Confusion matrix analysis

A row-normalized confusion matrix (with counts in parentheses) was created to demonstrate the classification efficacy of the optimal model (Random Forest) utilizing 5-fold cross-validation (Fig. 2). In the complete test sample of 270 videos (trending = 123; non-trending = 147), the model accurately recognized 127 non-trending videos (true negatives) and 106 trending videos (true positives), while committing 17 false negatives and 20 false positives. An excess of false negatives relative to false positives suggests a conservative inclination, specifically a propensity to underestimate the trend state rather than overestimate it. This behavior is consistent with the model's balanced learning objective, which prioritizes precision in identifying genuinely trending poems over recall. Such misclassifications are often observed in borderline cases where engagement metrics are moderate and linguistic cues exhibit mixed emotional tones. Overall, the confusion matrix confirms the robustness and interpretability of the Random Forest classifier within the 5-fold cross-validation framework.

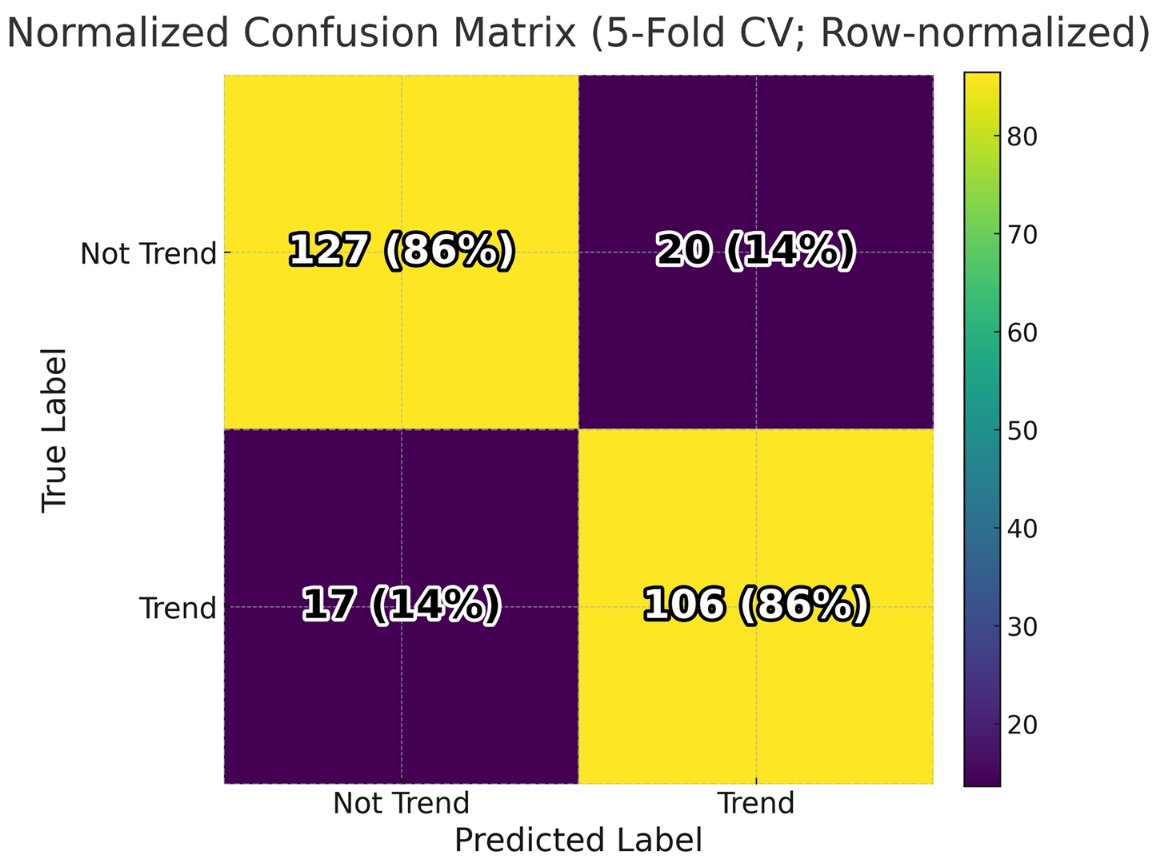

Fig. 2. Normalized confusion matrix of the Random Forest classifier for trend classification.

These findings validate the previously found equilibrium between precision and recall in quantitative measurements (Precision = 0.846, Recall = 0.861). The modest frequency of misclassifications indicates that the model proficiently distinguishes between the two groups without exhibiting bias towards the majority class. The marginally greater incidence of false negatives compared to false positives indicates that the model exhibits a degree of conservatism, occasionally prioritizing prudence by over-predicting specific trending poetry instead of neglecting to identify them.

6.3. Feature importance and key predictors of virality

To understand what drives these predictions, we examined feature importance from the best-performing random forest models. The analysis revealed that engagement metrics were key predictors of poetry's virality, reinforcing the intuition that popular content attracts more user interactions. In fact, as Fig. 3 shows, "number of views" was the single most important attribute by a wide margin. This is expected, as a video with more views is naturally more likely to be trending or popular. Similarly, "number of likes" and "number of posts" (shares) were among the top features – trending Hussaini poems accumulated significantly higher like counts and shares. This finding is consistent with prior studies on YouTube trends that use the number of views, likes, comments, and shares as key indicators of a video's popularity. "Number of comments" also had a significant impact, although slightly less than views and likes, suggesting that the amount of discussion contributes to virality (but is not sufficient for it alone).

Notably, video duration also emerged as an influential numerical feature. Random Forest's importance score for "video_minutes" (length of poem video) indicates that trending poems often have an optimal duration range. Extremely long or very short videos were less likely to take off, meaning audience engagement could decline outside a certain length window. This observation matches recent findings that video length can play an important role in whether a video goes viral. In our context, many trending Hussaini poetry videos were of medium length, which could maximize audience retention and sharing. Temporal posting features (publication year, month, day) had less importance relative to engagement metrics. There was a slight increase in posted content in some

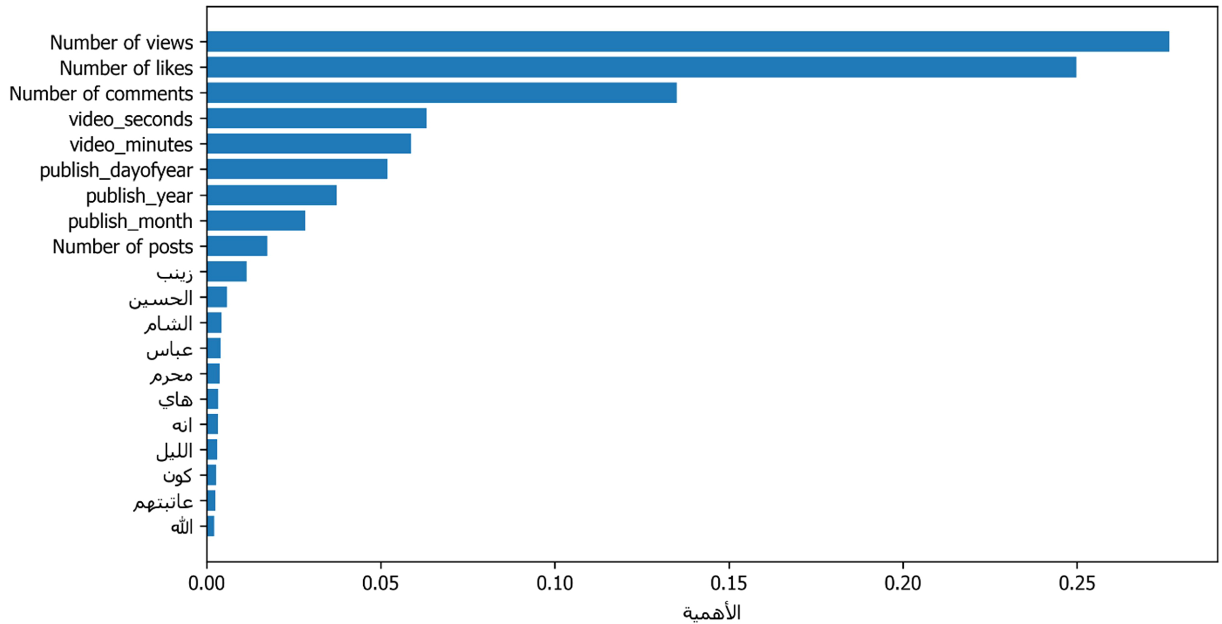

Fig. 3. Top 20 features of trend forecasting.

months, which may coincide with periods of high interest (e.g. Muharram, when there is demand for remembrance poetry), but these time-dependent effects were modest compared to direct engagement variables.

Importantly, the textual features of the poem titles also contributed to the model, albeit to a lesser extent than the numerical metrics. Random forest identified some TF-IDF features (words or capital letters in titles) as moderately predictive of virality. Upon inspection, many of the top-ranked headline words were those evoking strong emotional or religious themes. For example, words meaning " الحسين ", " زينب ", or the name "الشام " (a revered person in Hussaini contexts) appeared more frequently in trending headlines and received higher feature importance weights. This suggests that headlines with emotionally resonant phrases or references to prominent figures attract audiences and are slightly more likely to go viral. Such a result is intuitive and confirms observations in social media analytics that content signals such as titles and keywords can influence the popularity of a post. Indeed, other work on YouTube trending videos has noted that "good titles" – with attractive thumbnails and proper tagging – are important factors for attracting views. In our results, while no single word can match the predictive power of a strong numerical feature, the combination of several subtle textual cues helped groups differentiate borderline cases (for example, a poem with moderate statistics may still be trending if its title taps highly relevant emotions).

In summary, the model performance and feature analysis highlight that virality in Hussaini's poems is driven by a mix of high engagement and relevant content topics. The surest predictor was engagement metrics: if a poetry video received a large number of views, likes, and shares, it was very likely to be trending. This underlines that audience attention (as determined by these metrics) is a prerequisite for virality. However, content features were not irrelevant – the titles of trending poems often contained particular phrases that resonated with the community, giving their trending potential an additional boost. The interplay of these factors may explain why combination models performed best: they can simultaneously take advantage of continuous numerical features and sparse textual signals, capturing non-linear synergistic effects (for example, a compelling headline can only lead to virality if paired with sufficient initial engagement, and vice versa). These findings are consistent with a broader understanding of online content popularity, where both user engagement signals and the content itself (topic and presentation) jointly determine what is shared widely. By combining textual TF-IDF features with metadata, our approach was able to achieve high accuracy in predicting trending poems. This result demonstrates the value of multi-modal feature sets for trend prediction and provides insight for content creators: To maximize the chances of a poem going viral, one should not only aim for high engagement (through outreach and platform dynamics), but also create titles and content that deeply resonate with the target audience. The current model is already performing well, but further improvements could include incorporating additional textual analysis (such as sentiment or topic modeling on the poem content) and experimenting with engagement dynamics (e.g., growth rate of views) to capture viral momentum.

7. Conclusion

This study presents a machine learning approach that integrates textual and engagement data to forecast which Hussaini poetry videos are likely to trend. This combined methodology demonstrated efficacy: the top-performing model (Random Forest) attained an F1-score of roughly 0.85 in detecting trending videos, markedly surpassing basic linear classifiers. Our investigation indicated that audience interaction metrics—specifically view counts, likes, and shares—are the principal determinants of virality, but the incorporation of textual cues, even brief headlines, offers an additional predictive advantage by engaging emotionally resonant subjects.

From a practical perspective, our findings offer concrete guidance for digital creators of religious and cultural content. To maximize a video's trending potential, creators should aim for an optimal video length (avoiding extremely long or very short videos that might lose audience interest) and craft compelling, emotion-laden titles that resonate with the community. Equally important is actively encouraging early audience engagement: prompting viewers to like, comment, and share can rapidly boost the metrics that platforms use to identify trending content. By combining high-quality, meaningful content with strategic engagement efforts, content creators can significantly enhance the likelihood of their work achieving viral reach.

In the future, there are multiple avenues to advance this research. Future endeavors may encompass a comprehensive multimodal analysis by deriving insights from the acoustic and visual elements of the film. These components will enhance the textual and numerical features employed here by encapsulating performance aspects such as text style and emotional tone. Applying advanced text analysis techniques, such as sentiment or topic modeling, to the lyrics or descriptions of poems could uncover deeper linguistic elements that affect their appeal. Another significant focus is to analyze the temporal dynamics of interaction, such as monitoring the initial increase rate of views or assessing the timing of releases in relation to major religious events, to more accurately identify patterns of viral momentum. Assessing the framework on larger or more varied datasets, encompassing additional genres of cultural media, will further examine its generalizability and resilience.

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

References

1. T. Işık, "The effects of digital culture and new media on religious identity in the postmodern age: The case of Türkiye," *Medya ve Din AraştırmalarıDergisi*, no. Special Issue 1, pp. 253–280, Nov. 2024, doi: 10.47951/mediad.1524883.
2. M. Mirshahvalad, "Transnational karbala: From rebellion to reconciliation," *Religions (Basel)*, vol. 15, no. 12, Dec. 2024, doi: 10.3390/rel15121536.
3. C. H. Soelseth, *Instapoetry as a post-digital phenomenon The infrastructural effects of platformization on contemporary pop poetry*.
4. H. Son and Y. E. Park, "Predicting user engagement with textual, visual, and social media features for online travel agencies' Instagram post: evidence from machine learning," *Current Issues in Tourism*, vol. 27, no. 22, pp. 3608–3622, 2024.
5. M. Mustak, H. Hallikainen, T. Laukkanen, L. Plé, L. D. Hollebeek, and M. Aleem, "Using machine learning to develop customer insights from user-generated content," *Journal of Retailing and Consumer Services*, vol. 81, Nov. 2024, doi: 10.1016/j.jretconser.2024.104034.
6. S. Jamil Shahwan, "The impact of social media on literature," *Arab World English Journal*, no. 1, pp. 226–245, Jan. 2023, doi: 10.24093/awej/comm1.18.
7. K. L. O'Halloran, G. Pal, and M. Jin, "Multimodal approach to analysing big social and news media data," *Discourse, Context and Media*, vol. 40, Apr. 2021, doi: 10.1016/j.dcm.2021.100467.
8. M. S. Irshad, A. Anand, and M. Ram, "Trending or not? Predictive analysis for youtube videos," *International Journal of System Assurance Engineering and Management*, vol. 15, no. 4, pp. 1568–1579, Apr. 2024, doi: 10.1007/s13198-023-02034-8.
9. S. L. de Sá, A. A. de A. Rocha, and A. Paes, "Predicting popularity of video streaming services with representation learning: A survey and a real-world case study," *Sensors*, vol. 21, no. 21, p. 7328, 2021.
10. C.-Y. Lee and Y.-N. Tu, "Predicting hit songs using audio and visual features," *Engineering Proceedings*, vol. 89, no. 1, p. 43, 2025.
11. M. Mahdikhani, "Predicting the popularity of tweets by analyzing public opinion and emotions in different stages of Covid-19 pandemic," *International Journal of Information Management Data Insights*, vol. 2, no. 1, Apr. 2022, doi: 10.1016/j.jjimei.2021.100053.
12. P. S. Devi, R. Geetha, and S. Karthika, "Trendingtags—classification & prediction of hashtag popularity using twitter features in machine learning approach proceedings," in *Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018*, Springer, 2019, pp. 161–177.
13. M. Cho, D. Jeong, and E. Park, "AMPS: Predicting popularity of short-form videos using multi-modal attention mechanisms in social media marketing environments," *Journal of Retailing and Consumer Services*, vol. 78, p. 103778, 2024.
14. N. Sabri Elmitwally and A. Alsayat, "The Multi-class classification for the first six surats of the holy Quran," 2020. [Online]. Available: www.ijacsa.thesai.org.

15. A. Iqbal and S. H. Hassan, "Impact of machine learning integration in Qur'anic studies," *Mach Learn*, vol. 9, no. 2, pp. 54–63, 2024.
16. B. Sulistio, A. Ramadhan, E. Abdurachman, M. Zarlis, and A. Trisetyarso, "The utilization of machine learning on studying Hadith in Islam: A systematic literature review," *Educ Inf Technol (Dordr)*, vol. 29, no. 5, pp. 5381–5419, Apr. 2024, doi: 10.1007/s10639-023-12008-9.
17. F. Qarah, "AraPoemBERT: A pretrained language model for arabic poetry analysis," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.12392.
18. Nadira Alifia Ionendri, Feri Candra, and Afdi Rizal, "News classification using natural language processing with TF-IDF and multinomial Naïve bayes," *Journal of Applied Computer Science and Technology*, vol. 6, no. 1, pp. 37–45, Jun. 2025, doi: 10.52158/jacost.v6i1.1099.
19. M. Cherifi, M. N. El Korso, S. Fortunati, A. Mesloub, and L. Ferro-Famil, "Robust inference with incompleteness for logistic regression model,".
20. B. Gaye, D. Zhang, and A. Wulamu, "Improvement of support vector machine algorithm in big data background," *Math Probl Eng*, vol. 2021, 2021, doi: 10.1155/2021/5594899.
21. E. A. Abbas and N. A. Hussein, "Algorithm comparison for data mining classification: Assessing bank customer credit scoring default risk," *Jurnal Kejuruteraan*, vol. 36, no. 5, pp. 1935–1944, Sep. 2024, doi: 10.17576/jkukm-2024-36(5)-13.
22. M. U. N. Nisa, D. Mahmood, G. Ahmed, S. Khan, M. A. Mohammed, and R. Damaševičius, "Optimizing prediction of YouTube video popularity using XGBoost," *Electronics (Basel)*, vol. 10, no. 23, p. 2962, 2021.