A COMPARATIVE MACHINE LEARNING STUDY FOR THEME AND EMOTION EXTRACTION FROM ENGLISH AND BANGLA POETRY

Zinia Rahman[1], Wang Zheng[1], Refat Khan Pathan[2]

[1]School of Automation, Department of Control Science and Engineering, Southeast University Nanjing, China
[2]School of Computing and Artificial Intelligence, Faculty of Engineering and Technology Sunway University, Malaysia

ABSTRACT

Automatic interpretation of poetry presents significant challenges for natural language processing due to figurative language, cultural symbolism, and subtle emotional cues. This study proposes a comparative computational framework for extracting themes and emotions from English and Bangla poems using TF-IDF features and multiple supervised algorithms. Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees (DT), Random Forests (RF) - and a Multilayer Perceptron (MLP)were evaluated for both thematic and emotional categorization. For English poetry, ensemble and margin-based models achieved the highest performance, with SVM and Random Forest attaining up to 88.7% accuracy for emotion and 85.5% for theme classification. In Bangla poetry, theme classification remained highly discriminative, with Random Forest achieving 94% accuracy.The study demonstrates the effectiveness of traditional machine learning approaches for bilingual poetic analysis in low-resource literary domains.

KEYWORDS

Poetry Analysis, Emotion and theme classification, Deep Learning, MLP, ML

1. INTRODUCTION

Poetry represents one of the most linguistically rich and semantically complex forms of human expression. Unlike factual or narrative prose, poetic language frequently relies on metaphor, symbolism, ambiguity, and culturally embedded emotion, making automated interpretation a challenging task for natural language processing (NLP) systems. The extraction of themes and emotions from poems is therefore not only a technical problem but also an interdisciplinary challenge spanning computational linguistics, digital humanities, and affective computing.

In recent years, sentiment and emotion analysis have become central research topics in NLP, driven by applications in opinion mining, mental health analysis, and creative content understanding [1], [2], [3]. While substantial progress has been achieved for domains such as social media, product reviews, and news articles, poetic text remains relatively underexplored due to its stylistic deviation from standard language models and limited annotated datasets [4]. This gap is even more pronounced for low-resource languages, such as Bangla, where computational literary resources remain scarce.

Early approaches to emotion and theme detection relied heavily on lexicon-based methods, which map words to predefined affective categories [5]. Although interpretable, these methods often fail

David C. Wyld et al. (Eds): NLPTT – 2026
pp. 01-18, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160701

to capture contextual meaning and figurative language commonly found in poetry. Consequently, machine learning–based approaches, particularly those utilizing vector-space representations such as Bag-of-Words [6], [7] and TF-IDF [8], [9], have gained popularity for modeling latent semantic structures in text. Classical classifiers, including Support Vector Machines (SVM) and Random Forests (RF), have demonstrated strong performance in sentiment and emotion classification tasks, especially when training data is limited [10], [11], [12], [13].

More recently, deep learning architectures, such as Convolutional Neural Networks (CNNs) and BERT-based models, have been applied to emotion and theme detection with notable success [14], [15]. These models are capable of learning hierarchical representations of text but often require large-scale labelled datasets and extensive computational resources. For literary text, where datasets are typically small and stylistically diverse, deep models may not consistently outperform classical techniques. Hybrid approaches - where neural models are trained on engineered features such as TF-IDF have therefore emerged as a promising compromise between expressive power and data efficiency.

From a multilingual perspective, cross-lingual and bilingual literary analysis remains an open research problem. Bangla (Bengali), spoken by over 230 million people worldwide, possesses a rich poetic tradition characterized by strong emotional expressiveness and distinctive lexical patterns. However, most existing computational studies on Bangla focus on sentiment analysis in social media or news text rather than poetry [16]. Comparative studies that evaluate the same methodological framework across English and Bangla poetic corpora are notably rare.

This paper addresses these gaps by presenting a comparative machine learning framework for theme and emotion extraction from English and Bangla poetry. Using a unified TF-IDF feature representation, we evaluate multiple classical classifiers - SVM, k-Nearest Neighbors (KNN), Decision Trees (DT), and Random Forests (RF), alongside a custom lightweight MLP- trained directly on TF-IDF vectors. By conducting controlled experiments across two languages, this study aims to (i) assess the effectiveness of traditional versus neural models for poetic text, (ii) analyze cross-lingual performance differences, and (iii) contribute empirical insights into computational approaches for literary analysis in both high and low-resource settings. Statistical significance analysis has been done to figure out the differences between models.

2. LITERATURE REVIEW

The field of computational poetry analysis, particularly for emotion and theme extraction, has seen significant methodological evolution over the past several years. In contrast to traditional NLP tasks focused on social media or product reviews, poetry presents unique challenges: figurative language, stylistic abstraction, and culturally embedded metaphors complicate direct application of standard models. As such, recent research has explored both classical machine learning techniques and advanced neural architectures to address these challenges in literary text.

2.1. Natural Language Processing for Poetry

Poetry presents distinct challenges for natural language processing systems. Kao and Jurafsky conducted comprehensive computational analysis of poetic texts, revealing that poetry employs distinct linguistic patterns compared to prose, including higher rates of imagery, sound devices, and semantic complexity [17]. These characteristics necessitate specialized approaches to feature extraction and model design.

The work on automatic generation of poetry using Bi-LSTM network - demonstrated the complexity of capturing poetic structure and meaning computationally [18]. Their research highlighted that successful computational poetry analysis requires attention to both semantic content and formal properties such as meter, rhyme, and figurative language.

2.2. Emotion Detection in Text

Emotion detection in text has evolved significantly over the past two decades. Early work by Pang and Lee established fundamental approaches to sentiment analysis using machine learning classifiers [19]. Their pioneering work on movie reviews demonstrated that machine learning methods could effectively capture subjective information from text, laying groundwork for more nuanced emotion detection tasks. Early methods using sentiment dictionaries (e.g., VADER) provided computational efficiency but frequently failed to parse context, sarcasm, or archaic poetic diction [20]. Supervised machine learning offered a significant leap forward; comparative studies indicate that SVMs and TF-IDF consistently outperform lexicon approaches by learning domain-specific emotional associations [21]. The NRC Emotion Lexicon, developed by [22] that have become foundational resources for emotion detection research – comprises associations between words and eight basic emotions. These lexical resources provide crucial features for machine learning models attempting to understand emotional content in literary texts.

Traditional machine learning approaches have demonstrated considerable success in emotion classification tasks. Alm et al. [23] pioneered the application of supervised learning to emotion detection in text, using support vector machines and decision trees to classify sentences from children's stories into emotional categories. Their work established that even relatively simple feature extraction methods, when combined with robust classifiers, could achieve meaningful performance on emotion detection tasks. The application of ensemble methods has shown particular promise as random forest algorithm, which combines multiple decision trees to improve classification accuracy and reduce overfitting, has been successfully applied to various text classification problems [11], [12], [14].

2.3. Theme Extraction from Literature Texts

Thematic analysis of literary texts has benefited from advances in topic modeling and text classification. It involves distinguishing between unsupervised discovery and supervised retrieval. Latent Dirichlet Allocation (LDA) has long been the standard for "distant reading" and topic discovery. However, its application to poetry is often criticized for generating "opaque" topics that conflate thematic subject matter with recurring motifs, as well as for its instability when applied to short texts like individual poems [24]. Conversely, supervised learning approaches provide greater precision for categorization tasks. Research demonstrates that when predefined themes exist (e.g., Love, Nature, War), classifiers such SVM yield higher accuracy with Linear [25]. Ref. [26] examines the potential of context-dependent language models, specifically BERT, for understanding poetic words and shows that a BERT-based model outperforms a traditional support vector machine (SVM) based model.

An innovative AI framework was utilized using machine learning algorithms, particularly SVM and feature extraction techniques (TF-IDF and Doc2Vec) to objectively classify poems by their stylistic and thematic attributes [27]. A novel SaSa algorithm was proposed for classification of Koshur poems into five different genres, which achieved an 88% accuracy, outperforming the best performing traditional model, SVM, which achieved 76%. Overall many algorithms and methodologies were used to increase classification accuracy, but few to none found which actually allows an interactive view for the general enthusiastic population to explore. Yang et. al. applied deep learning methods to large-scale analysis of literary themes, demonstrating that

computational approaches could reveal patterns using a Duelist Algorithm-optimized Bi-directional Long Short-Term Memory (DAO-BiLSTM) model. His work has achieved an accuracy of 96.24%, showing superiority over previous works.

Recent work has explored the application of neural approaches to theme detection. However, as noted by researchers working with specialized literary domains, deep learning models often require substantial training data and may not always outperform traditional methods on smaller, domain-specific datasets [28].

2.4. Feature Extraction Methods

2.4.1. TF-IDF and Traditional Text Representation

Term Frequency-Inverse Document Frequency remains one of the most robust and interpretable methods for text representation. Salton and Buckley's foundational work on TF-IDF established its effectiveness for information retrieval and text classification tasks [29]. The method's ability to identify distinctive terms while reducing the influence of common words makes it particularly suitable for capturing thematic content in poetry, where specific word choices often carry significant semantic weight.
Jones provided theoretical justification for inverse document frequency weighting, demonstrating its statistical foundations [30]. For poetry analysis, TF-IDF offers the advantage of transparency - researchers can examine which terms the model considers most significant for classification, facilitating interpretation of results in literary terms.

2.4.2. Word Embeddings and Contextual Representations

The introduction of word embeddings revolutionized text representation. Mikolov et al.'s Word2Vec approach demonstrated that distributional semantics could capture meaningful semantic relationships [31]. These dense vector representations have shown promise for capturing nuanced meanings in literary texts, though their effectiveness varies depending on training data characteristics.

More recently, transformer-based models such as BERT have achieved state-of-the-art performance on numerous NLP tasks [32]. Devlin et al. demonstrated that pre-trained language models could be fine-tuned for specific tasks with relatively modest amounts of labeled data. However, the application of BERT to specialized domains like poetry has shown mixed results. The model's pre-training on contemporary web text may limit its effectiveness on poetic language, which employs distinct linguistic conventions and may include archaic vocabulary [33].

2.5. Limitations in Literature

Current approaches to computational poetry analysis face several limitations. The scarcity of large, well-annotated poetry datasets constrains the application of data-hungry deep learning approaches. Cultural and linguistic specificity of poetic traditions means that models trained on one language or tradition may not transfer effectively to others.

The interpretability-performance trade-off remains an important consideration. While transformer-based models like BERT achieve strong performance on many NLP tasks, their application to poetry has shown inconsistent results. Traditional methods with carefully engineered features may offer better performance and interpretability for specialized domains with limited training data.

3. METHODOLOGY

3.1. Data Collection and Organization

The English poem dataset utilized in this research (taken from Kaggle [34]) was meticulously curated and systematically structured into a hierarchical folder organization to facilitate supervised learning and classification tasks. Two primary classification objectives were established. The emotion classification task encompassed twelve emotional categories: anger, depression, despair, fear, funny, happiness, hate, joy, lonely, love, lust, and sympathy. In parallel, the theme classification task incorporated twenty-four thematic categories: respect, believe, teacher, money, crazy, sea, passion, marriage, snake, dream, poem, remember, father, mother, work, together, star, peace, friend, power, greed, food, birth, and river. Dataset details are mentioned in Table 1.

Table 1. English poem class distribution in theme and emotion corpus.

Theme	Number of Poems	Theme	Number of Poems		Emotion	Number of Poems
birth	100	power	100		joy	100
crazy	100	poem	100		hate	100
dream	100	snake	100		happiness	100
food	100	river	100		funny	100
greed	100	father	99		lonely	100
friend	100	believe	99		love	100
mother	100	remember	99		lust	100
marriage	100	money	99		sympathy	100
together	100	star	99		despair	99
teacher	100	respect	99		fear	99
passion	100	work	99		depression	99
peace	100	sea	98		anger	98

The Bengali Poems dataset has been taken from ref. [35] for Theme analysis, no reliable dataset was found for emotion analysis from poem. The theme classification dataset was organized using a hierarchical, poet-based folder structure, where each subfolder corresponded to a particular thematic category within Bangla literary traditions. In total, the thematic corpus encompassed 14 major categories. Details class distribution is shown in Table 2.

Table 2. Bangla poem class distribution in Theme corpus.

Theme Category	Number of Poems	Theme Category	Number of Poems
প্রেমমূলক	1,356	ছড়া	323
চিন্তামূলক	1,316	স্বদেশমূলক	249
মানবতাবাদী	723	নীতিমূলক	180
সনেট	468	হাস্যরসাত্মক	127
ভক্তিমূলক	394	কাহিনীকাব্য	57
রূপক	392	শোকমূলক	55
প্রকৃতিমূলক	347	গীতিগাথা	31

3.2. Data Processing and Feature Extraction

The English dataset was pre-processed through a structured and automated data-loading pipeline. The pre-processing stage commenced with data cleaning - which standardized the raw text by converting all characters to lowercase, removing newline characters, and eliminating redundant whitespace. Following data loading and pre-processing, a data balancing step was implemented to address the inherent class imbalance within both emotion and theme classification datasets – using resampling process of the minority classes. Resulting in a total of 24,000 poems in total for theme dataset.

The Bangla theme dataset was stored in a hierarchical directory structure based on poet and poem folders. Each poem contained two text files: one for the poem's content (.txt) and another for its thematic class label (CLASS.txt). A custom loading function was implemented to traverse the directory, read the UTF-8 encoded files, and construct a structured dataset for supervised training. To handle inconsistencies in Bangla digital text, a data cleaning procedure was followed to systematically remove extraneous characters - including Bangla digits (০–৯), punctuation marks (। ,; :? ! - …), and typographic noise (e.g., ‘’“” and whitespace anomalies). The initial theme dataset comprised 6018 poems, distributed unevenly across 14 thematic categories. A similar English poem like resampling method was applied to upsample the minority classes in Bangla theme dataset.

Finally, the text data from these balanced datasets were vectorized using the custom defined TF–IDF vectorizer, generating the feature matrix. For English poem dataset – vectorizer was configured to extract up to 3,000 of the most informative features, considering both unigrams and bigrams, while excluding common English stop words. To reduce noise and improve generalization, terms appearing in more than 90% of documents or in fewer than two documents are discarded. All text is converted to lowercase to ensure case insensitivity, and no accent stripping is applied, thereby preserving the original character accents in the input text. For Bangla theme data vectorization - a predefined list of common Bangla stop words (এবং, ও, থেকে, জন্য, হলো) is supplied to exclude high-frequency, low-semantic-value terms from the feature space. The vectorizer is configured to extract up to 3,000 features, incorporating both unigrams and bigrams to capture contextual information. To reduce noise and improve model robustness, terms occurring in more than 90% of documents or fewer than two documents are filtered out. Additionally, a Unicode-specific token pattern ([\u0980-\u09ff]) is applied to ensure that only valid Bangla script tokens are considered during tokenization, making the vectorization process linguistically appropriate for Bangla corpora. For both languages, corresponding labels were encoded using the label encoder to produce numerical target arrays.

3.3. Model Description and Training

3.3.1. Multilayer Perceptron (MLP) Network

Initially upsampled TF-IDF feature matrix and corresponding class labels were partitioned into training and testing subsets using a 70:30 split, with stratification applied to preserve the original class distribution and a fixed random seed to ensure reproducibility. It then defines a feedforward neural network for multi-class text classification using a sequential architecture. The model consists of multiple fully connected hidden layers with 512, 256, and 128 neurons for English poems, 512, 256, 128, 64 neurons for Bangla theme respectively, each employing ReLU activation, batch normalization to stabilize and accelerate training, and dropout regularization to mitigate overfitting. The output layer uses a softmax activation function to produce class probability distributions across all target categories. Finally, models are compiled with the Adam

optimizer and a learning rate of 0.001, using sparse categorical cross-entropy as the loss function to accommodate integer-encoded labels, and accuracy as the primary evaluation metric.

3.3.2. Machine Learning Models – SVM, KNN, DT, RF

Multiple supervised machine learning models were employed to evaluate and compare classification performance under a unified experimental framework. Model evaluation was conducted using stratified five-fold cross-validation, ensuring that class proportions were preserved across all folds while random shuffling with a fixed seed enhanced reproducibility.

A linear Support Vector Machine (SVM) classifier was implemented using a linear kernel, with a regularization parameter $C = 1.0$. To address potential class imbalance, balanced class weights were applied, and the maximum number of optimization iterations was set to 5,000 to ensure convergence. In addition, a k-Nearest Neighbors (kNN) classifier was used with $k = 5$, employing cosine distance as the similarity metric, which is particularly suitable for high-dimensional sparse representations such as TF-IDF vectors. Distance-based weighting was applied so that closer neighbors contributed more strongly to the classification decision.

Tree-based models were also considered. A Decision Tree classifier was configured with a maximum depth of 50 to control model complexity, while balanced class weights were used to mitigate bias toward majority classes. Furthermore, an ensemble-based Random Forest classifier was constructed using 300 decision trees, with no explicit depth limitation, allowing the model to capture complex, non-linear decision boundaries. Balanced class weighting and parallel computation across all available CPU cores were employed to improve robustness and computational efficiency.

Model performance was assessed using multiple evaluation metrics, including accuracy, macro-averaged precision, recall, and F1-score, providing a comprehensive evaluation that accounts for both overall performance and class-level balance. Cross-validation results were obtained through a standardized evaluation function, ensuring consistency and comparability across all models.

4. RESULT AND DISCUSSION

4.1. English Poems – Theme Result Analysis

The performance of the simple MLP model on the English poem's theme classification task demonstrates a strong and generally balanced capability to distinguish among a diverse set of classes. For theme classification - MLP achieved an accuracy of 0.81, with macro-averaged precision, recall, and F1-score all reaching 0.81-0.82, indicating consistent performance across classes and limited bias toward any particular theme.

Several themes exhibit particularly strong classification performance. Categories such as money, snake, remember, believe, andwork achieved high precision and recall values (F1-scores above 0.89), suggesting that these themes are characterized by distinctive lexical and semantic patterns that are effectively captured by the TF-IDF–based representation and the learning model. For instance, the "money" theme attained an F1-score of 0.95, reflecting both high precision (0.94) and recall (0.97), while "snake" and "remember" also demonstrated similarly robust results, highlighting the model's ability to identify well-defined thematic cues in poetic text. Detailed classification report is shown in Table 3 and confusion matrix is shown in
Figure 1.

Table 3. Classification report for English poem – theme classification for MLP.

Class	Precision	Recall	F1-score	Support
believe	0.93	0.90	0.92	30
birth	0.55	0.70	0.62	30
crazy	0.80	0.93	0.86	30
dream	0.72	0.77	0.74	30
father	0.62	0.83	0.71	30
food	0.96	0.80	0.87	30
friend	0.77	0.77	0.77	30
greed	0.85	0.77	0.81	30
marriage	1.00	0.63	0.78	30
money	0.94	0.97	0.95	30
mother	0.86	0.80	0.83	30
passion	0.92	0.73	0.81	30
peace	0.81	0.73	0.77	30
poem	0.72	0.70	0.71	30
power	0.78	0.70	0.74	30
remember	0.96	0.90	0.93	30
respect	0.66	0.97	0.78	30
river	0.79	0.90	0.84	30
sea	0.79	0.90	0.84	30
snake	0.96	0.90	0.93	30
star	0.81	0.87	0.84	30
teacher	0.86	0.80	0.83	30
together	0.72	0.60	0.65	30
work	0.96	0.83	0.89	30
accuracy			0.81	720
macro avg	0.82	0.81	0.81	720
weighted avg	0.82	0.81	0.81	720

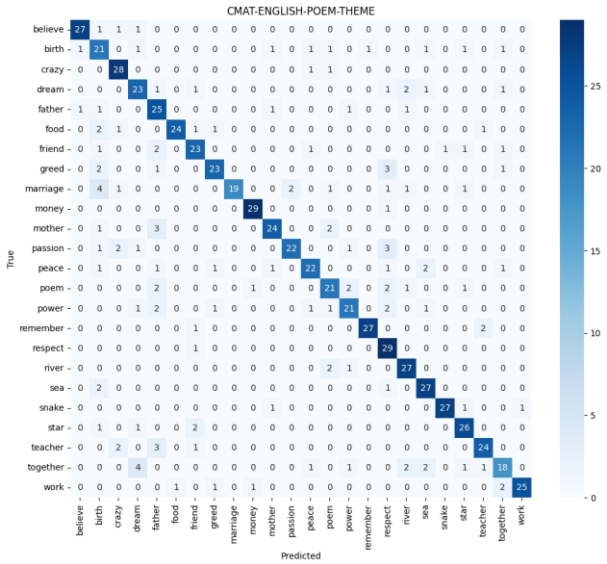

Figure 1. Confusion matrix of English theme classification using MLP.

The performance of classical machine learning models on theme classification demonstrates that ensemble and linear models achieve the highest predictive accuracy, while distance and tree-based approaches show comparatively lower performance. Among the evaluated models, Random Forest achieved the highest mean accuracy (0.855) and macro F1-score (0.853), closely followed by linear SVM with a mean accuracy of 0.852 and F1-score of 0.851. Both models also exhibit stable performance across folds, with low standard deviations, indicating consistent generalization.

kNN performed moderately, achieving a mean accuracy of 0.793 and macro F1-score of 0.789, reflecting slightly lower effectiveness in handling the high-dimensional TF-IDF features. Decision Tree exhibited the lowest overall accuracy (0.754) and F1-score (0.779), though its precision remains relatively high (0.840), suggesting the model is more precise for certain themes but prone to under-classifying others. Detailed result is shown in Table 4.

Overall, the results indicate that ensemble-based methods such as Random Forest and linear models like SVM are well-suited for English poem theme classification, providing both high accuracy and balanced macro-level performance across multiple thematic categories. These findings highlight the advantage of combining robust feature representations with models capable of capturing complex decision boundaries in multi-class poetic datasets.

Table 4. Classification result summary of four ML models for English theme.

Model	**Accuracy (mean)**	**F1macro (mean)**	**Precisionmacro (mean)**	**Recall macro(mean)**
SVM	0.852083	0.850642	0.857832	0.852083
KNN	0.7925	0.788708	0.799533	0.7925
DecisionTree	0.75375	0.778619	0.839681	0.75375
RandomForest	0.854583	0.853172	0.865639	0.854583

4.2. English Poem – Emotion Result Analysis

The emotion classification results for English poems indicate that the similar simple MLP can achieve strong and well-balanced performance across a diverse set of affective categories. An overall accuracy of 0.82 was obtained, with macro-averaged precision, recall, and F1-score all reaching approximately 0.82-0.83, reflecting consistent predictive behavior across emotions and minimal dominance by any single class.

High classification performance is observed for emotions with clear and distinctive linguistic expressions, such as anger, fear, and lust, all of which achieve F1-scores close to or above 0.89. These emotions often involve explicit affective vocabulary and strong sentiment cues, enabling the model to capture them effectively. Similarly, love and happiness demonstrate stable performance, indicating reliable detection of positive emotional states in poetic language. Detailed classification result is shown in Table 5 and confusion matrix is shown in Figure 2.

Table 5. Classification report for English poem – emotion classification for MLP.

Class	**Precision**	**Recall**	**F1-score**	**Support**
anger	0.96	0.87	0.91	30
depression	0.74	0.87	0.80	30
despair	0.75	0.80	0.77	30
fear	0.96	0.83	0.89	30
funny	0.77	0.90	0.83	30
happiness	0.83	0.83	0.83	30

hate	0.88	0.77	0.82	30
joy	0.65	0.67	0.66	30
lonely	0.81	0.73	0.77	30
love	0.89	0.83	0.86	30
lust	0.93	0.87	0.90	30
sympathy	0.74	0.87	0.80	30
accuracy			0.82	360
macro avg	0.83	0.82	0.82	360
weighted avg	0.83	0.82	0.82	360

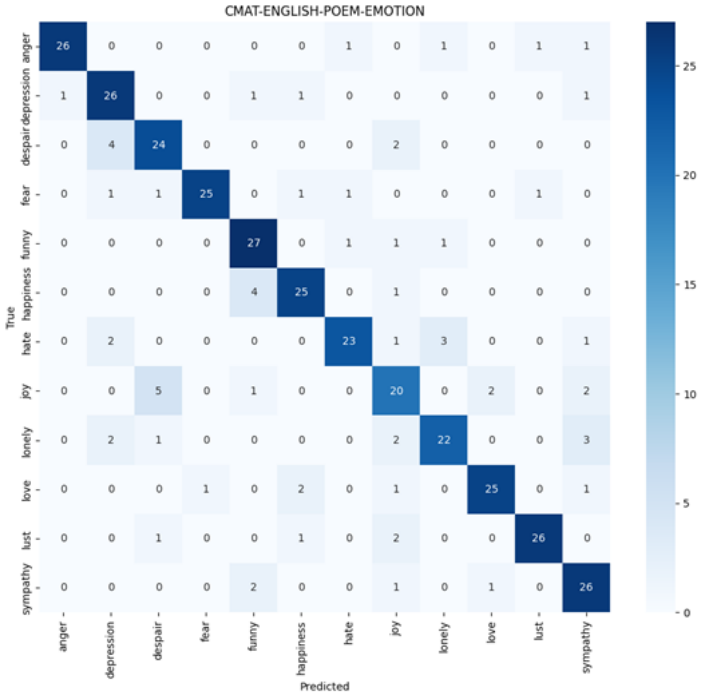

Figure 2. Confusion matrix for English emotion classification using MLP.

The evaluation of ML models for emotion recognition shows similar trend like theme classification. Linear SVM attained the highest mean accuracy (0.887) and macro F1-score (0.887), closely matched by Random Forest, which achieved a mean accuracy of 0.885 and F1-score of 0.886. Both models also demonstrated low to moderate standard deviations across folds, indicating reliable generalization.

kNN achieved a mean accuracy of 0.825 and F1-score of 0.823, reflecting slightly lower effectiveness in distinguishing high-dimensional, sparse TF-IDF features for nuanced emotional categories. Decision Tree achieved intermediate performance with an accuracy of 0.843 and F1-score of 0.843, showing relatively good precision (0.859) but slightly lower recall, indicating some under-classification in certain emotion categories. Overall, the results suggest that SVM and Random Forest are the most effective models for English emotion recognition in poetry, combining both high accuracy and balanced macro-level performance across diverse emotion classes. This underscores the suitability of these models for capturing subtle affective cues in textual data. Detailed result is shown in
Table 6.

Table 6. Classification result summary of four ML models for English emotion.

Model	Accuracy (mean)	F1macro (mean)	Precisionmacro (mean)	Recallmacro (mean)
SVM	0.886667	0.886822	0.893595	0.886667
KNN	0.825	0.823488	0.834343	0.825
DecisionTree	0.8425	0.842939	0.859016	0.8425
RandomForest	0.885	0.886124	0.897243	0.885

4.3. Bangla Poems – Theme Result Analysis

The Bangla theme classification results demonstrate very strong overall performance of the trained MLP model, substantially outperforming the English theme classification task. The model achieves an overall accuracy of 93%, with macro-averaged precision, recall, and F1-score all equal to 0.93, indicating highly consistent and balanced performance across all thematic categories. Several Bangla poetic themes exhibit near-perfect classification performance. Categories such as কাহিনীকাব্য (narrative poetry), গীতিগাথা (ballad), শোকমূলক (elegiac), and হাস্যরসাত্মক (humorous) achieve perfect precision, recall, and F1-scores of 1.00. This suggests that these themes possess highly distinctive structural, stylistic, and lexical characteristics, which are effectively captured by the TF-IDF representation and learning model. Similarly, নীতিমূলক (didactic) and স্বদেশমূলক (patriotic) poems show exceptionally high performance, with F1-scores of 0.99 and 0.98, respectively.

Strong performance is also observed for themes such as ছড়া (rhymed verse), ভক্তিমূলক (devotional), সনেট (sonnet), and রূপক (allegorical), all achieving F1-scores above 0.92. These results indicate that the model can successfully distinguish both structural poetic forms and content-driven themes within Bangla literature. Relatively lower performance is noted for চিন্তামূলক (philosophical) and প্রেমমূলক (romantic) themes, each achieving F1-scores of approximately 0.71. These themes often involve abstract concepts and emotionally nuanced language that overlaps with other categories, such as humanistic or reflective poetry, thereby increasing classification ambiguity. Nonetheless, the precision and recall values for these classes remain well balanced, suggesting stable rather than erratic model behavior. Detailed classification result is shown in Table 7 and confusion matrix is shown in Figure 3.

Table 7. Classification report for Bangla poem – theme classification for MLP.

Class	Precision	Recall	F1-score	Support
কাহিনীকাব্য	1.00	1.00	1.00	407
গীতিগাথা	1.00	1.00	1.00	407
চিন্তামূলক	0.71	0.71	0.71	407
ছড়া	0.98	0.97	0.97	407
নীতিমূলক	0.99	1.00	0.99	407
প্রকৃতিমূলক	0.93	0.92	0.92	407
প্রেমমূলক	0.71	0.72	0.71	407
ভক্তিমূলক	0.96	0.94	0.95	407
মানবতাবাদী	0.90	0.88	0.89	407
রূপক	0.94	0.90	0.92	406
শোকমূলক	1.00	1.00	1.00	407
সনেট	0.92	0.94	0.93	406
স্বদেশমূলক	0.96	1.00	0.98	407
হাস্যরসাত্মক	1.00	1.00	1.00	407

accuracy			0.93	5696
macro avg	0.93	0.93	0.93	5696
weighted avg	0.93	0.93	0.93	5696

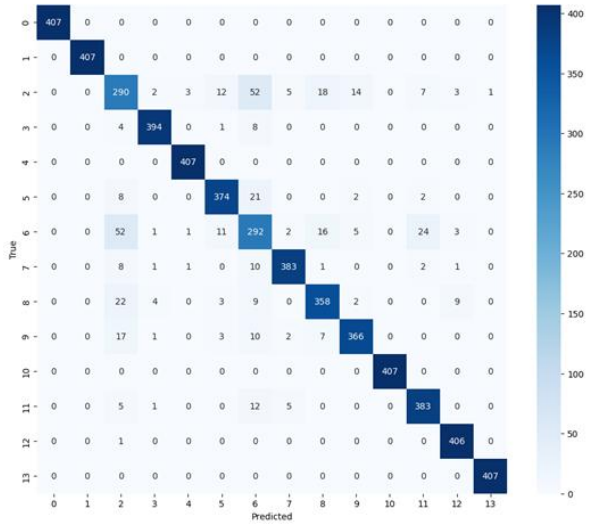

Figure 3. Confusion matrix for Bangla theme classification using MLP.

The performance of classical machine learning models on Bangla poem's emotion recognition demonstrates that ensemble methods, particularly Random Forest, achieve the highest predictive accuracy and balanced class level performance. Random Forest attained the highest mean accuracy (0.942) and macro F1-score (0.942), with very low standard deviation, indicating both superior performance and excellent stability across folds. Linear SVM also performed strongly, achieving a mean accuracy of 0.914 and F1-score of 0.910, closely trailing Random Forest and demonstrating reliable classification across diverse emotional categories.

K-Nearest Neighbors achieved slightly lower performance, with a mean accuracy of 0.903 and F1-score of 0.892, reflecting moderate effectiveness in distinguishing subtle emotional cues in high-dimensional TF-IDF representations. Decision Tree exhibited the lowest performance, with an accuracy of 0.788 and F1-score of 0.798, suggesting reasonable precision (0.838) but lower recall, indicating under-classification for certain emotion classes. Detailed result is shown in Table 8.

Overall, these results highlight that ensemble-based models such as Random Forest are particularly well-suited for Bangla them recognition, likely benefiting from the clear lexical and semantic separability of thematic expressions in Bangla poetry. The combination of high accuracy, balanced F1-scores, and low variability underscores the robustness and reliability of the proposed machine learning approach for fine-grained affective analysis in Bangla literary texts.

Table 8. Classification result summary of four ML models for Bangla theme.

Model	**Accuracy (mean)**	**F1macro (mean)**	**Precisionmacro (mean)**	**Recallmacro (mean)**
SVM	0.913875	0.910079	0.90974	0.913877
KNN	0.902549	0.892363	0.897947	0.902549
DecisionTree	0.787717	0.798359	0.838038	0.787724
RandomForest	0.942478	0.942435	0.94299	0.942483

4.4. Statistical Significance Analysis

4.4.1. English poem – statistical analysis

To assess the statistical significance of performance differences among the evaluated models for English theme and emotion classification, multiple non-parametric statistical tests were conducted using macro-averaged F1-score as the primary evaluation metric.

A Wilcoxon signed-rank test was first applied to compare the performance of the linear SVM and Random Forest classifiers. The test yielded a test statistic of 6 with a p-value of 0.8125, indicating no statistically significant difference between these two models at conventional significance levels. This suggests that, despite observable differences in mean F1-scores, the variation across folds is not sufficient to conclusively favour one model over the other based solely on this pairwise comparison.

In contrast, the Friedman test, which evaluates differences across multiple related models, revealed a statistically significant result ($x^2 = 13.56,\ p = 0.0036$). This outcome confirms that at least one model among SVM, kNN, Decision Tree, and Random Forest exhibits significantly different performance, justifying further pairwise and model-level comparisons. To provide a more robust estimation of model performance, bootstrap resampling was employed to compute confidence intervals for the macro F1-scores. The MLP-based model achieved a mean F1-score of approximately 0.810, with a 95% confidence interval ranging from 0.790 to 0.830. In comparison, the Random Forest model demonstrated a higher mean F1-score of approximately 0.853, with a tighter confidence interval between 0.837 and 0.864. The non-overlapping nature of these confidence intervals provides additional evidence of the superior average performance of the Random Forest model.

Furthermore, bootstrap analysis of the mean difference between MLP and Random Forest models indicates a consistently negative difference in F1-score (mean difference ≈ -0.054), with the confidence interval entirely below zero. This confirms that the Random Forest model significantly outperforms the MLP-based approach for English theme classification. Overall, the statistical analysis demonstrates that while some traditional models exhibit comparable performance, ensemble-based methods such as Random Forest achieve statistically and practically superior results for English poem theme classification.

A similar statistical evaluation framework was applied to the English emotion classification task. Using macro-averaged F1-score as the evaluation metric, a Wilcoxon signed-rank test comparing SVM and Random Forest models produced a test statistic of 7 with a p-value of 1.0. This result indicates no statistically significant difference between these two models under the Wilcoxon test, suggesting comparable fold-wise performance. However, the Friedman test across all classical machine learning models (SVM, kNN, Decision Tree, and Random Forest) yielded a statistically significant result ($x^2 = 13.56,\ p = 0.0036$). This confirms the presence of meaningful performance differences among the evaluated classifiers and supports the need for model-level comparison beyond pairwise testing.

Bootstrap-based confidence interval analysis further clarifies these differences. The MLP model achieved a mean macro F1-score of approximately 0.820, with a 95% confidence interval ranging from 0.800 to 0.840. In contrast, the Random Forest model achieved a substantially higher mean

F1-score of approximately 0.886, with a confidence interval spanning from 0.850 to 0.911. The wider separation between these intervals indicates a clear and consistent advantage of the Random Forest model. Additionally, bootstrap analysis of the mean F1-score difference between MLP and Random Forest models shows a negative mean difference of approximately -0.091, with the confidence interval entirely below zero. This provides strong statistical evidence that the Random Forest model significantly outperforms the MLP model for English emotion classification.

In summary, the statistical analyses for both English theme and emotion classification tasks consistently indicate that Random Forest models offer superior and statistically significant performance advantages, particularly when compared against neural models, while differences among some traditional classifiers remain statistically indistinguishable.

4.4.2. Bangla poem – statistical analysis

A similar comprehensive statistical analysis was conducted to examine performance differences among the evaluated models for Bangla poem theme classification, using macro-averaged F1-score as the primary evaluation metric. The Wilcoxon signed-rank test was used to perform a pairwise comparison between the SVM and Random Forest models. The test produced a statistic of 0 with a p-value of 0.0625, which is marginally above the conventional significance threshold. This indicates that, while the Random Forest model consistently outperformed the SVM across folds, the observed difference does not reach statistical significance at the 0.05 level, suggesting broadly comparable performance between these two strong classifiers.

In contrast, the Friedman test, which evaluates performance differences across multiple related models, yielded a statistically significant result ($x^2 = 15.0$, $p = 0.0018$). This confirms the presence of significant performance variation among the classical machine learning models considered, namely SVM, kNN, Decision Tree, and Random Forest, and justifies further model-level comparisons. To quantify performance stability and uncertainty, bootstrap resampling was applied to estimate confidence intervals for the macro F1-scores. The MLP-based model achieved a mean F1-score of approximately 0.930, with a 95% confidence interval ranging from 0.910 to 0.950. The Random Forest model demonstrated slightly higher performance, with a mean F1-score of approximately 0.942 and a narrower confidence interval between 0.937 and 0.950, indicating both superior average performance and greater stability.

Further insight is provided by bootstrap analysis of the mean difference in F1-score between the MLP and Random Forest models. The estimated mean difference is negative (approx. -0.020), with the corresponding confidence interval entirely below zero. This result provides statistical evidence that the Random Forest model achieves significantly higher performance than the MLP model for Bangla theme classification, although the absolute magnitude of the improvement is modest.

4.5. Discussion

This study evaluated the effectiveness of classical machine learning and neural network models for fine-grained thematic and emotional analysis of English and Bangla poetry. Both qualitative classification metrics and rigorous statistical analyses were employed to assess model performance, providing a comprehensive view of the applicability, robustness, and limitations of these approaches across languages and tasks.

For English poems, Random Forest and linear SVM consistently outperformed other models in both theme and emotion classification tasks. In theme classification, Random Forest achieved the

highest mean macro F1-score (0.853), closely followed by SVM (0.851), while KNN and Decision Tree lagged behind (F1-scores 0.789 and 0.779, respectively). Emotion classification revealed a similar pattern, with Random Forest and SVM yielding F1-scores above 0.885, and KNN and Decision Tree performing moderately (F1-scores 0.823 and 0.843, respectively). These results indicate that ensemble and linear models are particularly effective in handling high-dimensional, sparse TF-IDF representations, capturing both discriminative lexical patterns for thematic classification and subtle affective cues for emotion recognition.

Bangla poetry, in contrast, showed markedly higher classification performance. For theme classification, Random Forest achieved a mean F1-score of 0.942, with MLP and SVM also performing strongly (0.930 and 0.910, respectively). Class-level performance was exceptionally high, with several themes, including কাহিনীকাব্য, গীতিগাথা, শোকমূলক, and হাস্যরসাত্মক, achieving perfect precision, recall, and F1-score.

Though a nearly similar trend in experiments and parameters were maintained for both English and Bangla poem analysis, due to the uneven dataset size and properties, mentioned results are not directly comparable. Results should be interpreted as relatively comparative.

5. CONCLUSION

This study presents a comprehensive evaluation of machine learning approaches for thematic and emotional analysis of English and Bangla poetry, integrating quantitative performance metrics, class-level analysis, and statistical validation. For English poetry, Random Forest and linear SVM consistently achieved the highest accuracy and macro F1-scores for both theme and emotion classification, although abstract and overlapping categories such as birth or joy posed greater challenges. In contrast, Bangla poetry demonstrated near-perfect classification performance, with Random Forest and other models achieving exceptionally high accuracy and F1-scores for themes, reflecting the strong lexical and structural separability of Bangla poetic expressions. Statistical analyses, including Friedman tests, Wilcoxon signed-rank tests, and bootstrap confidence intervals, confirmed the significance of observed performance differences, highlighting the robustness and reliability of ensemble-based approaches such as Random Forest. Cross-linguistic comparison further revealed that Bangla poetry's explicit thematic markers enable higher model discriminability than English poetry, though ensemble models remain consistently superior across tasks and languages.

Overall, the results demonstrate the effectiveness of classical machine learning and neural approaches for fine-grained analysis of poetry. The findings provide strong evidence for the feasibility of automated, scalable computational analysis in literary research, enabling objective insights into thematic and emotional patterns. Future work may focus on extending these approaches to more heterogeneous and metaphorically complex corpora, integrating semantic embeddings and hybrid models to further enhance classification of nuanced poetic expressions.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008, doi: 10.1561/1500000011.

[2] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective Computing and Sentiment Analysis," pp. 1–10, 2017, doi: 10.1007/978-3-319-55394-8_1.

[3] R. A. García-Hernández *et al.*, "A Systematic Literature Review of Modalities, Trends, and Limitations in Emotion Recognition, Affective Computing, and Sentiment Analysis," *Applied Sciences 2024, Vol. 14,* vol. 14, no. 16, Aug. 2024, doi: 10.3390/APP14167165.

[4] S. M. Mohammad, "Sentiment Analysis: Detecting Valence, Emotions, and Other Affectual States from Text," *Emotion Measurement*, pp. 201–237, Jan. 2016, doi: 10.1016/B978-0-08-100508-8.00009-6.

[5] C. Strapparava and A. Valitutti, "WordNet Affect: an Affective Extension of WordNet," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2004. Accessed: Jan. 22, 2026. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2004/pdf/369.pdf

[6] N. T. Renukadevi, S. Nanthitha, R. T. Karthika, and S. Shobika, "Sentimental Analysis with Continuous Bag of Words for Book Reviews," in *3rd International Conference on Innovative Mechanisms for Industry Applications, ICIMIA 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1193–1198. doi: 10.1109/ICIMIA60377.2023.10426572.

[7] S. Lee, "Social Media Text Sentiment Classification Based on Bag-of-Words Models with Multiple Machine Learning Algorithms," *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology, AINIT 2024*, pp. 1775–1779, 2024, doi: 10.1109/AINIT61980.2024.10581434.

[8] B. Dutta, M. Neog, and N. Baruah, "Unveilling the Effectiveness of LSTM with TF-IDF in Assamese Poem Classification:A Comparative Analysis with Conventional and Deep Learning Approaches," *2024 15th International Conference on Computing Communication and Networking Technologies, ICCCNT 2024*, 2024, doi: 10.1109/ICCCNT61001.2024.10725384.

[9] H. Liu, X. Chen, and X. Liu, "A Study of the Application of Weight Distributing Method Combining Sentiment Dictionary and TF-IDF for Text Sentiment Analysis," *IEEE Access*, vol. 10, pp. 32280–32289, 2022, doi: 10.1109/ACCESS.2022.3160172.

[10] A. P. Jain and P. Dandannavar, "Application of machine learning techniques to sentiment analysis," in *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2016*, Institute of Electrical and Electronics Engineers Inc., Apr. 2017, pp. 628–632. doi: 10.1109/ICATCCT.2016.7912076.

[11] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis," *Procedia Comput. Sci.*, vol. 127, pp. 511–520, Jan. 2018, doi: 10.1016/J.PROCS.2018.01.150.

[12] P. Karthika, R. Murugeswari, and R. Manoranjithem, "Sentiment Analysis of Social Media Network Using Random Forest Algorithm," *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2019*, Apr. 2019, doi: 10.1109/INCOS45849.2019.8951367.

[13] A. S. Zharmagambetov and A. A. Pak, "Sentiment analysis of a document using deep learning approach and decision trees," *Proceedings of the 2015 12th International Conference on Electronics Computer and Computation, ICECCO 2015*, Feb. 2016, doi: 10.1109/ICECCO.2015.7416902.

[14] J. Yu and C. Qi, "Machine Learning-Based Sentiment Analysis in English Literature: Using Deep Learning Models to Analyze Emotional and Thematic Content in Texts," *IEEE Access*, vol. 13, pp. 65997–66008, 2025, doi: 10.1109/ACCESS.2025.3553386.

[15] J. F. Ruma, J. Sultana, S. Akter, J. J. Laboni, and R. M. Rahman, "Deep Learning and BERT-Based Models for Poetry Analysis: A Study on Chronological and Subject-Wise Categorization of Tagore's Poetic Works," *SN Computer Science 2025 6:7*, vol. 6, no. 7, pp. 904-, Oct. 2025, doi: 10.1007/S42979-025-04408-0.

[16] N. Banik, M. H. H. Rahman, S. Chakraborty, H. Seddiqui, and M. A. Azim, "Survey on Text-Based Sentiment Analysis of Bengali Language," in *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*, Institute of Electrical and Electronics Engineers Inc., May 2019. doi: 10.1109/ICASERT.2019.8934481.

[17] J. Kao and D. Jurafsky, "A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry," in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, 2012, pp. 8–17. Accessed: Jan. 22, 2026. [Online]. Available: https://aclanthology.org/W12-2502/

[18] T. Zhang, J. Guo, and Y. Zhang, "Research on Poetry Generation Based on Bi-LSTM Networks and Attention Mechanism," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, May 2023, pp. 6–13. doi: 10.1145/3613917.3613926.

[19] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, Jan. 2008, doi: 10.1561/1500000011.

[20] "Lexicon-Based Approach in Sentiment Analysis | MiaRec." Accessed: Jan. 10, 2026. [Online]. Available: https://blog.miarec.com/lexicon-based-vs-ml-based-sentiment-analysis
[21] R. Srivastava, P. K. Bharti, and P. Verma, "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 3, pp. 71–77, Mar. 2022, doi: 10.14569/IJACSA.2022.0130312.
[22] S. M. Mohammad, "Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, Association for Computational Linguistics (ACL), 2018, pp. 174–184. doi: 10.18653/V1/P18-1017.
[23] C. O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-based Emotion Prediction," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 579–586. Accessed: Jan. 22, 2026. [Online]. Available: https://aclanthology.org/H05-1073/
[24] A. Kaur and J. R. Wallace, "Moving Beyond LDA: A Comparison of Unsupervised Topic Modelling Techniques for Qualitative Data Analysis of Online Communities," vol. 1, Dec. 2024, Accessed: Jan. 10, 2026. [Online]. Available: https://arxiv.org/pdf/2412.14486v1
[25] N. Jamal, M. Mohd, and S. A. Noah, "Poetry Classification Using Support Vector Machines," *Journal of Computer Science*, vol. 8, no. 9, pp. 1441–1446, Aug. 2012, doi: 10.3844/JCSSP.2012.1441.1446.
[26] K. Choi, "Computational Thematic Analysis of Poetry via Bimodal Large Language Models," *Proceedings of the Association for Information Science and Technology*, vol. 60, no. 1, pp. 538–542, Oct. 2023, doi: 10.1002/PRA2.812.
[27] L. Yang, G. Wang, H. Wang, L. Yang, G. Wang, and H. Wang, "Reimagining Literary Analysis: Utilizing Artificial Intelligence to Classify Modernist French Poetry," *Information 2024, Vol. 15,* vol. 15, no. 2, Jan. 2024, doi: 10.3390/INFO15020070.
[28] A. Piper, "Think Small: On Literary Modeling," *PMLA*, vol. 132, no. 3, pp. 651–658, May 2017, doi: 10.1632/PMLA.2017.132.3.651.
[29] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, Jan. 1988, doi: 10.1016/0306-4573(88)90021-0.
[30] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval | Document retrieval systems," in *Document retrieval systems*, 1988, pp. 132–142. Accessed: Jan. 22, 2026. [Online]. Available: https://dl.acm.org/doi/10.5555/106765.106782
[31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, Jan. 2013, Accessed: Jan. 22, 2026. [Online]. Available: https://arxiv.org/pdf/1301.3781
[32] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, vol. 1, pp. 4171–4186, Oct. 2018, Accessed: Jan. 22, 2026. [Online]. Available: https://arxiv.org/pdf/1810.04805
[33] L. Martin *et al.*, "CamemBERT: a Tasty French Language Model," *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 7203–7219, May 2020, doi: 10.18653/v1/2020.acl-main.645.
[34] michaelarman, "Poems Dataset (NLP)." Accessed: Jan. 24, 2026. [Online]. Available: https://www.kaggle.com/datasets/michaelarman/poemsdataset/data
[35] M. K. B. Shuhan, R. Dey, S. Saha, M. S. U. Anjum, and T. S. Zaman, "A Stylometric Dataset for Bengali Poems," in *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, Association for Computing Machinery, Dec. 2022, pp. 176–180. doi: 10.1145/3582768.3582788.

AUTHORS

ZINIA RAHMAN (Student Member, IEEE), received the B.Sc. degree (Hons.) in computer science and engineering from Pabna University of Science and Technology Bangladesh, in 2017. She is currently pursuing M.Sc. degree in Control Science and engineering under School of automation with Southeast University, Nanjing, China.

Her research interest including computer vision, artificial intelligence and machine learning.

ZHENG WANG (Senior Member, IEEE) is currently working as a professor at school of automation department, Southeast University, Nanjing, China. He received Ph.D. in Control Theory and Control Engineering, in 2002 at Southeast University, China. His research area is Control of Radio Telescopes; Optimization and Control of inventory and supply chain systems; Manufacturing systems analysis, optimization and control; Optimization theory and optimal control; Discrete event dynamic systems; Design optimization of complex products and engineering systems.

REFAT KHAN PATHAN received his B.Sc. degree (Hons.) in computer science and engineering from BGC Trust University Bangladesh, and M.Sc. in Computer Science (By Research) from Sunway University Malaysia. He is currently pursuing his PhD in Computing from Same University. His research interest includes image processing, trustworthy AI, Agentic AI, data mining and security.