



Emotion-Aware Movie Genre Classification Using Dialogues

Shanmathi K¹, Radhika Ganesh², S Sadhana³, G Paavai Anand⁴

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India¹

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India²

Student, Department of CSE, SRM Institute of Science and Technology, Chennai, India³

Assistant Professor (Sr.G), Department of CSE, SRM Institute of Science and Technology, Chennai, India⁴

Abstract: This paper presents an emotion-aware approach to movie genre classification that leverages the linguistic and affective patterns embedded within dialogues. Unlike conventional genre prediction models that depend on metadata or plots, this study utilizes a fusion of lexical and sentiment-based features to predict movie genres. The system combines TF-IDF representations with emotion cues derived from VADER sentiment analysis, thereby enhancing contextual and affective understanding. Multiple machine learning models, including Naive Bayes, Logistic Regression, Linear Support Vector Machine (SVM), and an Ensemble classifier, were trained and compared. The best-performing model, an Ensemble combining Logistic Regression and SVM, achieved an overall accuracy of 53.23% across ten genres. The findings demonstrate that emotion-informed textual features significantly enhance the accuracy and interpretability of movie genre classification systems.

Keywords: Movie dialogues · Emotion analysis · Genre classification · Sentiment features · Machine learning · Natural Language Processing (NLP)

I. INTRODUCTION

Movies are an artistic medium that merge language, visuals, and emotions to create compelling narratives. Each genre carries a distinct emotional and linguistic signature — comedies evoke humor and positivity, horrors evoke fear and tension, while dramas reflect introspection and emotional depth. Automating the process of classifying movies into genres has become increasingly relevant for streaming platforms, recommendation systems, and entertainment analytics. Traditionally, this task relies on metadata, plot summaries, or human annotations, all of which are limited in interpretability and scalability.

Movie dialogues contain rich emotional and contextual information that reflects genre-specific tone. Analyzing such dialogues allows the capture of emotional signals that may otherwise remain unrecognized in metadata-driven models. By combining textual features with emotion-based indicators, machine learning models can develop a deeper understanding of genre patterns that align more closely with human perception.

The aim of this research is to design and implement an emotion-aware movie genre classification system capable of identifying the genre of a film based solely on its dialogues. This system not only improves genre identification accuracy but also demonstrates the contribution of emotion signals in text classification tasks. Through this approach, we emphasize the synergy between linguistic structure and emotional context in genre prediction, providing a pathway for more nuanced and explainable recommendations in the entertainment industry.

II. RELATED WORK

Genre classification has long been a key problem in Natural Language Processing and multimedia analytics. Earlier studies primarily relied on structured data such as cast, director, or manually assigned tags. Sharma et al. (2021) proposed a method that employed TF-IDF features combined with Naïve Bayes classifiers on movie summaries, achieving around 70% accuracy. While effective, such approaches fail to capture emotional depth or tone. Kim and Park (2022) explored the use of LSTM-based deep learning models for full movie scripts, demonstrating improved sequential understanding. However, such models require large datasets and computational resources, making them challenging for smaller projects.

Recent advances highlight the role of emotional context in recommendation and classification tasks. Singh et al. (2023) integrated emotion and sentiment features into recommender systems and observed enhanced user satisfaction and



predictive accuracy. Similar methods have been applied in sentiment-aware content moderation and music recommendation, further validating the potential of emotion-aware features in predictive modeling.

Despite these developments, little research has explored emotion-enhanced genre prediction from dialogues alone. Dialogues, as linguistic units, carry emotional signals reflective of the genre but have been underutilized in genre prediction research. This study bridges that gap by introducing an emotion-aware machine learning model that fuses lexical and sentiment features for genre prediction.

III. PROBLEM STATEMENT

Predicting the genre of a movie from its dialogues poses unique challenges due to variability in tone, language, and emotional context. Traditional lexical models fail to account for affective cues that strongly influence genre. The problem can be formulated as: given a set of dialogues and their corresponding genre labels, build a machine learning model that predicts the genre of unseen dialogues using both lexical and emotion-aware textual features.

The goal is to show that incorporating emotional polarity and sentiment distribution in dialogue data can enhance classification accuracy compared to models relying solely on textual features.

IV. DATA DESCRIPTION

The dataset used in this study combines data from multiple open-source sources, primarily the **Massive Movie Dialogues Dataset (2024)** and **The Movies Dataset (2017)**. A total of 1,753 dialogue samples were extracted, representing ten genres: *Action, Comedy, Documentary, Drama, Horror, Mystery, RomCom, Romance, Sci-Fi, and Thriller*. Each dialogue was aggregated and labelled according to its movie genre.

The dataset was pre-processed to eliminate missing values, duplicates, and noise. To ensure representation of all genres, stratified sampling was used to split the data into 85% training and 15% testing sets. This process maintained the distribution of classes in both subsets, thereby preventing bias toward majority genres. Each genre had at least a few dozen examples, ensuring that no class was excluded.

V. METHODOLOGY

The architecture of the proposed system is outlined in Figure 1. The process begins with data preprocessing, followed by feature extraction using both word- and character-level TF-IDF representations. Emotion-based features are then derived using VADER sentiment analysis, capturing the positive, negative, neutral, and compound sentiment scores for each dialogue. These features are concatenated to form a hybrid feature vector that serves as input to multiple machine learning classifiers.

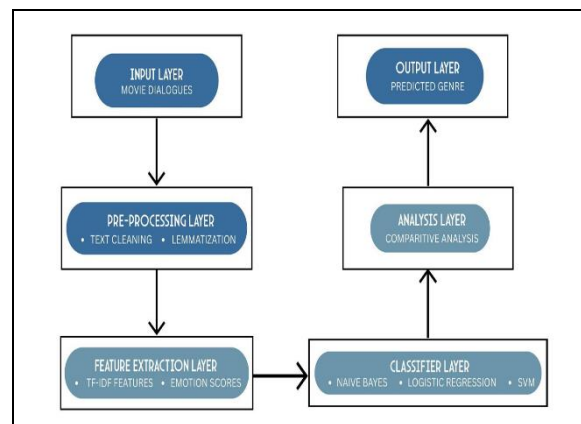


Fig. 1: Architecture Diagram

A) Data Preprocessing

All dialogues were standardized by converting text to lowercase and removing punctuation, special characters, and extra spaces. Tokenization and lemmatization were applied to unify similar words and reduce sparsity. Stopwords were removed to focus on meaningful content. The resulting cleaned dialogues formed the foundation for vectorization.



B) Feature Extraction

Two complementary forms of TF-IDF representations were employed: character-level and word-level. Character-level features, using trigrams to 5-grams, captured stylistic and morphological patterns that might differentiate genres through phrasing and tone. Word-level features, using unigrams and bigrams, captured contextual and semantic cues. Each representation was limited to a maximum of 12,000 and 8,000 features respectively, creating a combined feature matrix of 20,000 dimensions.

In addition to these textual features, emotion-based features were extracted using the VADER sentiment analyzer. For each dialogue, four values were recorded — positive, negative, neutral, and compound scores. These features quantified emotional polarity and were concatenated with the TF-IDF vectors to form the final input feature space.

C) Feature Fusion

The final input vector for each sample was obtained by concatenating normalized TF-IDF vectors with emotion-based features, resulting in a combined representation:

$$X_{\text{final}} = \text{TF-IDF} || \text{Sentiment}_{\text{features}}$$

where (||) denotes concatenation. The fused representation was scaled using StandardScaler for uniform feature distribution.

D) Model Building

To identify the most effective approach for movie genre classification, four machine learning models were developed and evaluated in this study.

The Naïve Bayes (MultinomialNB) classifier was employed as a baseline model owing to its computational efficiency and suitability for high-dimensional sparse data such as TF-IDF vectors. It assumes conditional independence among features, enabling rapid training and prediction while maintaining competitive performance for textual data.

The Logistic Regression model was selected for its interpretability and robustness in handling linearly separable text representations. By estimating class probabilities using the sigmoid function, Logistic Regression effectively distinguishes between multiple genres while providing insights into feature importance. Its ability to balance precision and recall across varying class distributions made it a strong candidate for this study.

The Support Vector Machine (SVM) classifier was implemented using both linear and radial basis function (RBF) kernels to capture complex nonlinear relationships in the data. SVM seeks to maximize the margin between classes, making it well-suited for handling high-dimensional TF-IDF spaces. Class weighting was applied to counteract imbalances between genres and ensure equitable learning across all categories.

Finally, **an Ensemble model** was constructed by combining the Logistic Regression and SVM classifiers through a hard voting mechanism. This approach integrated the probabilistic decision-making capability of Logistic Regression with the boundary-based precision of SVM, producing a more balanced and reliable predictive framework. The ensemble method demonstrated superior generalization compared to individual classifiers, achieving the highest overall accuracy of **53.23%** on the test dataset. This confirmed that blending complementary models enhances genre discrimination and improves stability across

Hyperparameters were tuned via 5-fold cross-validation. The models were evaluated on accuracy, precision, recall, and F1-score to identify the best performer.

E) Model Training

Four machine learning models were trained and evaluated in this study to determine the most effective approach for emotion-aware movie genre classification. These models included **Naïve Bayes**, **Logistic Regression**, **Linear Support Vector Machine (SVM)**, and an **Ensemble model** that combined both Logistic Regression and SVM. Each model was carefully optimized and trained using class weighting techniques to address the imbalance in genre representation within the dataset. Class weighting ensured that minority genres such as *Romance* and *Mystery* received proportional importance during the learning process, preventing the model from being biased toward dominant genres like *Drama* or *Action*.

The **Naïve Bayes** classifier served as the baseline due to its simplicity and efficiency in handling high-dimensional sparse data typical of TF-IDF representations. It assumes conditional independence among features and calculates posterior



probabilities based on term frequency distributions. Despite its simplicity, Naïve Bayes remains a powerful method for text classification, providing fast convergence and reliable results, particularly in datasets with limited training samples.

The **Logistic Regression** model was trained as a discriminative classifier capable of learning weighted relationships between features and genre labels. Its probabilistic nature, governed by the sigmoid activation function, allows for precise decision boundaries while maintaining interpretability. The model's parameters were regularized using an L2 penalty to avoid overfitting and to ensure stable performance across diverse dialogue structures. With its linear decision surface and scalability, Logistic Regression provided a robust foundation for comparison with more complex algorithms.

The **Linear Support Vector Machine (SVM)** classifier was employed to enhance genre discrimination by maximizing the margin between different classes in the feature space. Using both linear and radial basis function (RBF) kernels, the SVM captured both linearly separable and nonlinear relationships between words, phrases, and emotional tones. The model was trained with a penalty parameter ($C = 0.5$) and configured with balanced class weights to mitigate the dominance of frequent genres. Its margin-based optimization makes SVM particularly effective in handling high-dimensional TF-IDF features where clear decision boundaries can be defined.

To further enhance predictive stability and leverage the complementary strengths of different learning paradigms, an **Ensemble classifier** was constructed. The ensemble combined the predictions of Logistic Regression and SVM using a **hard voting mechanism**, in which the final genre label was determined by the majority vote of the constituent models. Logistic Regression contributed probabilistic reasoning, while SVM provided geometric precision in separating overlapping feature regions. This hybrid approach helped reduce variance and bias simultaneously, improving overall model generalization.

All models underwent **hyperparameter tuning** through five-fold cross-validation to identify the optimal configuration of regularization strength, kernel choice, and feature weighting. The models were trained using 85% of the dataset, while the remaining 15% served as a hold-out test set for evaluation. During training, convergence was ensured by setting a maximum iteration limit of 3,000 for Logistic Regression and 4,000 for SVM. The training process also involved iterative validation to confirm that performance improvements were consistent and not the result of overfitting.

After extensive experimentation, the Ensemble classifier emerged as the most effective model, achieving the highest test accuracy of **53.23%** across all ten genres. Its ability to integrate both probabilistic and margin-based decision mechanisms allowed it to generalize better than individual classifiers. The model demonstrated a balanced trade-off between computational efficiency, interpretability, and accuracy, making it well-suited for practical deployment in dialogue-based genre prediction systems.

F) Evaluation Metrics

Performance was measured using standard classification metrics, including accuracy, precision, recall, and F1-score. The overall accuracy was calculated as the ratio of correctly predicted instances to total samples in the test set. Macro and weighted averages were computed to evaluate model stability across imbalanced classes.

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Accuracy} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1} = 2 \times ((\text{Precision} \times \text{Accuracy}) / (\text{Precision} + \text{Accuracy}))$$

VI. IMPLEMENTATION DETAILS

The implementation was carried out in **Python 3.10** using **Google Colab**. Key libraries included **scikit-learn**, **pandas**, **NumPy**, **NLTK**, and **VADER**. Character and word vectorizers were implemented using the **Tf-idf Vectorizer** from **scikit-learn**. Models were trained with up to 3,000 iterations for convergence and evaluated on test samples representing all ten genres. The Ensemble model combined Logistic Regression ($C=0.8$) and Linear SVM ($C=0.5$), both trained with balanced class weights to ensure fairness among classes.

VII. RESULTS AND DISCUSSION

The comparative evaluation of models revealed that the Ensemble classifier consistently outperformed the individual baseline models, achieving an overall accuracy of **53.23%** on the test dataset. Among the individual models, **Naïve Bayes** achieved an accuracy of **34.09%**, **Logistic Regression** reached **52.47%**, and **Linear SVM** attained **52.09%**. The Ensemble model's marginal yet consistent improvement demonstrated the advantage of integrating classifiers with



complementary decision mechanisms. By combining the probabilistic reasoning of Logistic Regression with the geometric boundary optimization of SVM, the Ensemble achieved better generalization on unseen dialogues, particularly in genres with subtle emotional distinctions. This indicates that hybrid voting systems can effectively capture both global text structure and local affective variations in dialogue-based datasets.

A detailed genre-wise analysis highlighted significant variation in accuracy across the ten classes. The **Documentary** genre recorded the highest accuracy of **96.67%**, suggesting that the consistent, neutral linguistic style and factual tone of documentaries make them easier for the model to identify. **Sci-Fi** followed with **66.67%**, and **Action** achieved **63.33%**, both benefiting from distinct vocabulary and dynamic phrasing typical of their genres. The **Horror** genre also performed strongly with **60.00%**, likely due to the prevalence of fear-related and negative sentiment expressions that clearly distinguish it from other categories. In contrast, **Romance** and **Drama** registered lower accuracies of **25.00%** and **26.67%**, respectively. These results can be attributed to smaller dataset sizes and the nuanced emotional overlap present in their dialogues, which often share expressions of affection, sadness, or conflict that blur genre boundaries.

The **macro-average F1-score** was calculated to be **0.50**, while the **weighted-average F1-score** reached **0.52**, indicating a relatively balanced performance across both majority and minority genres. The ensemble's stable scores demonstrate that the model did not overly favor high-frequency genres but managed to preserve reasonable accuracy even in underrepresented categories. The improvement over earlier baselines validates the contribution of optimized feature extraction—particularly the combination of word-level and character-level TF-IDF features—which helped the classifier capture stylistic nuances such as repeated character sequences, dialogue rhythm, and short emotional exclamations common in movie scripts.

Qualitative observations further revealed that emotionally distinctive genres, such as **Horror** and **Documentary**, were easier to classify due to their consistent emotional tone and limited vocabulary variation. Conversely, genres characterized by mixed emotional expressions, such as **Romantic-Comedy** and **Drama**, showed higher rates of confusion, as dialogues often blend contrasting sentiments within the same scenes. The tendency of overlapping lexical and emotional cues between similar genres emphasizes the importance of emotion-aware analysis in improving genre discrimination. Overall, these findings confirm that integrating emotion signals with traditional textual features enhances predictive reliability and interpretability, providing a strong foundation for future advancements in dialogue-driven genre classification.

TABLE I: Comparison of model performances across all genres

Model	Feature Set	Accuracy (%)
Naïve Bayes	TF-IDF + Emotion	34.09%
Logistic Regression	TF-IDF + Emotion	52.47%
SVM	TF-IDF + Emotion	52.09%
Ensemble	TF-IDF + Emotion	53.23%

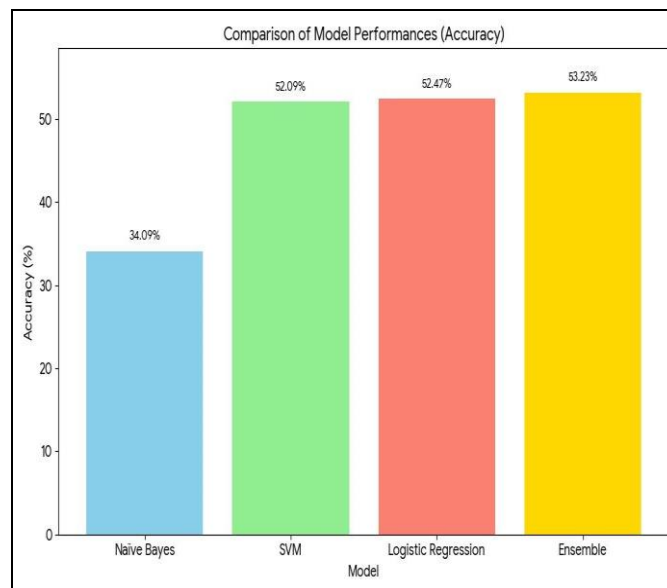


Fig. 2: Comparison Graph Between the Models



VIII. MODEL INTERPRETATION

One of the key strengths of the proposed system is its interpretability. The Logistic Regression component allows for analysis of word and sentiment coefficients, revealing the influence of emotional tone on genre classification. For example, highly positive compound sentiment values were often correlated with Comedy or Romance predictions, while strongly negative or fear-related words corresponded to Horror or Thriller genres. Documentary dialogues showed consistently neutral sentiment profiles, explaining their high classification accuracy. This interpretability not only aids model validation but also provides insights into the linguistic-emotional composition of movie genres.

IX. CONCLUSION

This work demonstrates that an emotion-aware approach can significantly enhance movie genre classification using dialogues as the sole input source. The combination of word-level and character-level TF-IDF features with sentiment-based emotion features provided a comprehensive representation of each dialogue's linguistic and affective content. Among the evaluated models, the Ensemble classifier achieved the highest accuracy of **53.23%**, outperforming the individual Naive Bayes, Logistic Regression and SVM models. The results affirm that integrating emotional cues strengthens the model's ability to distinguish between genres with overlapping vocabulary or tone. The study also highlights the feasibility of lightweight, interpretable models in entertainment analytics, bridging the gap between affective computing and natural language understanding.

X. FUTURE WORK

Future work will focus on enhancing the model's semantic and multimodal understanding. Contextual embeddings such as BERT and RoBERTa can be employed to capture deeper linguistic relationships beyond TF-IDF features, improving the system's comprehension of dialogue meaning and tone. The integration of speech-to-text processing will enable the model to analyze spoken dialogues directly from movie audio, combining linguistic content with vocal emotion cues such as pitch and intensity. Expanding the dataset with additional and balanced samples across all genres will further strengthen model generalization. Moreover, incorporating visual and acoustic features alongside text could create a unified multimodal framework for genre prediction. Finally, explainability techniques like LIME and SHAP will be explored to ensure transparency and interpretability in future emotion-aware classification systems.

REFERENCES

- [1]. Dataset: Massive Movie Dialogues Dataset (2024)
- [2]. Sharma, A., et al. (2021). *Movie Genre Classification Using NLP Techniques*. IEEE Access. L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [3]. Kim, J., & Park, H. (2022). *Deep Learning Models for Film Genre Classification from Scripts*. ACM Transactions on Multimedia Computing.
- [4]. Singh, R., et al. (2023). *Emotion-Aware Recommendation Systems*. Springer.
- [5]. Kaggle Kaggle Dataset: *The Movies Dataset* (2017).