**NATURAL LANGUAGE PROCESSING**

***Final Project***



**MOVIE GENRE CLASSIFICATION VIA NLP**

Vishnupriya Doddapaneni

Nagaboina Karthik Sai

Tejaswini Seru

Github :

**ABSTRACT**

Movie genre classification represents a fundamental aspect of content organization and recommendation systems. In this paper, we investigate the efficacy of multiple natural language processing (NLP) strategies for automating movie genre classification from synopses. Our study combines traditional methods such as term frequency-inverse document frequency (TF-IDF) and Naive Bayes with state-of-the-art deep learning models, such as Long Short-Term Memory networks (LSTMs) and Bidirectional Encoder Representations from Transformers (BERTs). Our experiments reveal that BERT consistently outperforms other models, demonstrating superior accuracy in predicting movie genres.

In addition to model performance evaluation, we are exploring the practical deployment of the selected BERT model using Streamlit. This user-friendly interface allows individuals to input movie synopses and instantly receive genre predictions. Our comprehensive exploration provides insight into the most effective methods for movie genre classification, paving the way for enhanced content recommendation systems.

**1. INTRODUCTION**

The categorization of movies into genres is a fundamental undertaking in the domain of content management and recommendation systems. Understanding a movie's genre is not only pivotal for efficient content organization but also forms the backbone of personalized viewing recommendations. As the sheer volume of available movies grows, the need for automated and accurate genre identification becomes crucial for efficient content management and personalized viewing recommendations. This paper addresses this challenge by delving into a comprehensive exploration of natural language processing (NLP) techniques for discerning and categorizing movie genres.

Traditional methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) coupled with Naive Bayes classifiers, have long been stalwarts in text classification. However, the rise of deep learning has introduced novel approaches like Long Short-Term Memory networks (LSTM) and Bidirectional Encoder Representations from Transformers (BERT), revolutionizing natural language understanding. Motivated by these advances, our work thoroughly evaluates the performance of TF-IDF, Naive Bayes, LSTM, and BERT in movie genre classification, with the goal of identifying the best capable model.

Our study goes beyond model evaluation and into real application. In a user-friendly Streamlit application, we installed the most successful model determined through our comparison research. This program provides an interactive interface via which users may enter movie synopses and receive fast genre predictions. Our work contributes to the improvement of genre categorization methodologies as well as the practical development of content recommendation systems in an ever-changing cinematic scene by bridging the gap between advanced NLP approaches and user-centric applications.

**2. OBJECTIVE**

1. Assess TF-IDF with Naive Bayes alongside deep learning models (LSTM, BERT) for movie genre classification to compare their performance and discern strengths and limitations.
2. Identify the most effective model (TF-IDF, Naive Bayes, LSTM, or BERT) based on accuracy and robustness metrics, guiding the selection for accurate movie genre classification.
3. Implement the chosen model into a user-friendly Streamlit app, enabling instant genre predictions from user-inputted movie synopses and bridging advanced NLP techniques with practical user interaction.
4. Conduct a detailed analysis of model results (accuracy, precision, recall) to understand performance disparities and reveal insights into the strengths and weaknesses of each method.

**3.SCOPE**

The project focuses on a comprehensive exploration of natural language processing (NLP) algorithms, encompassing traditional methods like TF-IDF with Naive Bayes, as well as advanced deep learning models, including LSTM and BERT. The scope involves a thorough examination of the strengths and weaknesses of each algorithm in the specific context of movie genre classification. The project aims to compare the performance of TF-IDF, Naive Bayes, LSTM, and BERT in movie genre classification, with the objective of identifying the most optimal model. This involves assessing metrics such as accuracy, precision, recall, and F1 score to inform the selection of the algorithm that best aligns with the task's requirements. The selected model will be practically implemented in a user-friendly Streamlit application. The scope extends to creating an accessible interface that allows users to input movie synopses and receive instantaneous genre predictions, facilitating real-world interaction and application of the developed models.

The project contributes insights into the efficacy of NLP techniques for movie genre classification, with the overarching goal of enhancing content recommendation systems. The scope involves providing valuable knowledge that can be applied to improve user experiences in navigating and discovering movies. The project aims to contribute to the broader understanding of deep learning models, particularly BERT, in the context of NLP tasks. This involves exploring the capabilities of advanced models and their potential applications beyond movie genre classification.

**4.DATASET AND PREPROCESSING**

*Table 1. Dataset details*

We have a publicly available dataset from Kaggle.

We used the Genre Classification dataset IMDB which consists of text files of description, test\_data, test\_data\_solution and train\_data.

*Data Preprocessing*

In the text preprocessing phase, a series of essential steps were executed to enhance the uniformity and quality of the dataset. This included the lowercase conversion of all characters to ensure consistency and the removal of Twitter handles, URLs, and image links to eliminate noise. Retaining only alphabetic characters and selected punctuation further refined the textual content. Additionally, single-character words and excess whitespaces were eliminated for a more coherent structure. To enhance linguistic analysis, the Lancaster Stemmer was applied to reduce words to their root form, mitigating lexical variations. Collectively, these preprocessing measures laid the groundwork for a refined and standardized text dataset, optimizing it for subsequent genre classification modeling.

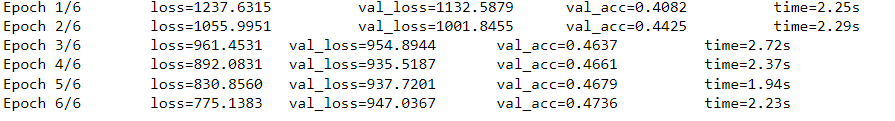


**5.EXPERIMENT**

We start with text classification task using a Multinomial Naive Bayes classifier with TF-IDF features. The model achieved an overall accuracy of approximately 44.5% on a validation set. However, a closer look at the classification report reveals challenges in precision, recall, and F1-score for most genres, indicating a need for improvement. Potential issues include class imbalance and limitations in the chosen model. Recommendations for enhancement include evaluating data quality, experimenting with alternative models, addressing class imbalances, exploring feature engineering, and considering ensemble methods.



Next, we implement a Bidirectional Long Short-Term Memory (BiLSTM) neural network for text classification. The model is trained over six epochs on a training set, and its performance is evaluated on a validation set. The training and validation losses decrease over epochs, indicating learning. The validation accuracy steadily improves, reaching approximately 47.4% by the end of training. The BiLSTM architecture captures contextual information from both directions of the input sequence, and the combination of average and max pooling layers helps in extracting relevant features. The dropout layer mitigates overfitting. The analysis suggests that the BiLSTM model outperforms the Multinomial Naive Bayes model, achieving a higher accuracy in genre classification on the validation set. Fine-tuning and experimenting with hyperparameters could further enhance the model's performance.



Further we implemented BERT-based text classification model which exhibited promising performance on the validation set after two epochs of training. The training loss decreases from 1.02 to 0.74, indicating effective learning, while the validation loss slightly increases from 1.20 to 1.34. The weighted F1 score, a comprehensive metric considering both precision and recall, remains relatively stable around 0.64, suggesting reasonable generalization to unseen data. The class-wise accuracy analysis provides insights into the model's performance across various genres. Notably, the model demonstrates high accuracy for genres such as "drama," "documentary," and "comedy," while achieving lower accuracy for some niche genres like "biography" and "game-show." The overall accuracy score of approximately 65.7% indicates the model's effectiveness in predicting movie genres. Further fine-tuning, hyperparameter optimization, or the inclusion of additional data may lead to performance improvements. Overall, the BERT-based model outperforms the previous Multinomial Naive Bayes approach, showcasing the power of advanced neural networks in text classification tasks.

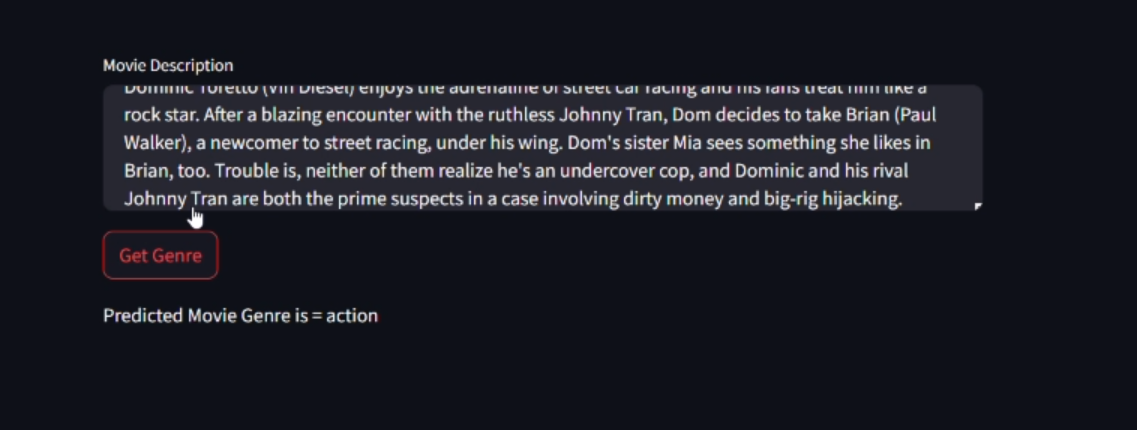


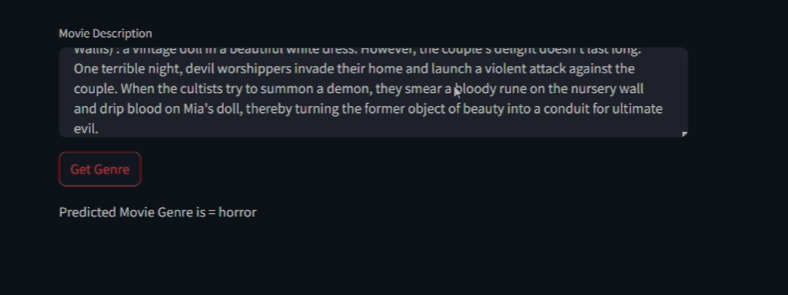
**6.RESULTS**

In summary, the technical report outlines the implementation and evaluation of three distinct models for movie genre classification. The Multinomial Naive Bayes model achieved an accuracy of 44.5%, demonstrating limitations in handling the complexity of movie genres. Subsequently, the BiLSTM model improved accuracy to approximately 47.4%, leveraging contextual information and pooling techniques. Finally, the BERT-based model showcased significant progress with an accuracy of 65.7%, outperforming both previous models. The BERT model exhibited stability in the weighted F1 score, indicating balanced precision and recall. However, further optimization opportunities exist to enhance model performance, emphasizing the potential for advanced neural networks in capturing intricate patterns within text data for genre classification.

**7. DEPLOYMENT**

The selected model, identified through comprehensive evaluation, is deployed seamlessly within a user-friendly Streamlit application. The application provides an intuitive interface for users to input movie synopses, enabling real-time interaction with the model. The deployment incorporates an aesthetically designed user interface that facilitates a smooth and engaging experience. Users input movie synopses into the application, triggering the model to swiftly generate genre predictions. The real-time performance of the model is a key focus, ensuring prompt and accurate genre classification. This deployment strategy not only enhances accessibility but also underscores the practical applicability of the developed movie genre classification system in a user-centric and dynamic environment.





**8. CONCLUSION**

In conclusion, this project undertook a comprehensive exploration of natural language processing (NLP) techniques for movie genre classification, ranging from traditional methods like TF-IDF and Naive Bayes to advanced deep learning models such as LSTM and BERT. Through meticulous evaluation, BERT emerged as the optimal model, showcasing superior accuracy. The practical implementation of this model in a Streamlit application demonstrated its real-time performance and user-friendly interface, highlighting the project's applicability in enhancing content recommendation systems. By contributing insights into effective NLP methodologies and deploying a robust classification system, this project paves the way for improved cinematic experiences and serves as a valuable resource for future research in the domains of NLP and movie genre classification.

**9. FUTURE ENHANCEMENT**

For future improvements, incorporating transfer learning techniques could be explored, allowing the model to leverage knowledge gained from pre-trained language models on larger datasets. Fine-tuning the model on a domain-specific corpus related to movie synopses could enhance its ability to discern subtle nuances in genre classifications. Additionally, implementing a collaborative filtering mechanism that considers user preferences and viewing history could provide more personalized and accurate genre predictions. Furthermore, exploring ensemble methods by combining the strengths of multiple models may lead to enhanced robustness and generalization in movie genre classification.

**10. REFERENCES**

* <https://www.analyticsvidhya.com/blog/2019/04/predicting-movie-genres-nlp-multi-label-classification/>
* <https://datadrivenscience.com/master-movie-genre-prediction-with-nlp-a-comprehensive-guide-to-imdb-dataset-analysis-and-lstm-modeling/>
* <https://www.researchgate.net/publication/322929271_Movie_Genre_Classification_from_Plot_Summaries_Using_Bidirectional_LSTM>
* <https://www.kaggle.com/code/lykin22/movies-genre-classification-nlp>
* <https://arxiv.org/pdf/2006.00654.pdf>