# **Movie Genre Classification Project Report**

**Introduction:**

The Movie Genre Classification project aims to categorize movies into various genres based on their titles and descriptions. The dataset includes information such as movie titles, genres, descriptions, and lengths. Genres range from common ones like drama and comedy to more specific ones like western, reality-TV, and fantasy.

**Data Preprocessing and Model Building:**

To prepare the data for classification, the document mentions the use of TfidfVectorizer. This technique converts movie descriptions into numerical features, making them suitable for machine learning models. To address the imbalance in genre representation, random oversampling is applied. This ensures that each genre has an equal influence on the model, improving its performance. The accuracy of the model is reported to have increased after oversampling, indicating better representation and learning.

**Word Cloud Analysis:**

A word cloud is generated to visually represent the most common words in movie descriptions. During preprocessing, non-alphanumeric characters are removed. The resulting word cloud offers an intuitive snapshot of frequently occurring words, providing a quick insight into the dataset.

**Top Trigrams Analysis:**

Understanding trigrams, or sets of three consecutive words, is crucial for identifying patterns or anomalies in the descriptions. A bar chart depicting the top 30 trigrams is included. This analysis helps uncover any irregularities such as unnecessary symbols, punctuation, or web links that may affect the model's performance.

**Model Evaluation:**

The accuracy of the model is reported to be 0.41. To delve deeper into its performance, precision, recall, and F1-score metrics are calculated for each genre. Genres like drama, documentary, comedy, action, thriller, and others are individually assessed. The classification report provides a detailed breakdown of how well the model performs for each genre, allowing for a nuanced understanding of its strengths and weaknesses.

**Conclusion:**

In conclusion, the Movie Genre Classification project employs sophisticated techniques like TfidfVectorizer and random oversampling to enhance the model's accuracy and address data imbalances. The inclusion of word cloud and trigram analyses adds depth to the understanding of the dataset, shedding light on common words and potential anomalies. The comprehensive evaluation through precision, recall, and F1-score metrics ensures a nuanced assessment of the model's performance across various genres.