**SENTIMENT ANALYSIS FOR IMDb MOVIE REVIEWS USING TEXTMINING**

**Description:**  
This project aims to use text mining techniques to analyse IMDb reviews to determine audience sentiment. By categorizing reviews as positive or negative, we aim to gain insights into how audiences perceive different movies. This analysis will offer valuable understanding for filmmakers, production companies, and marketers, helping them make informed decisions about film production, marketing strategies, and engaging with audiences.

**Dataset:**  
The dataset comprises IMDb reviews obtained from [Kaggle](https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews). Each review includes text feedback provided by users who have reviewed different movies. Important attributes in the dataset include the review text, movie title, user rating, and review date. Reviews are classified as positive or negative based on the sentiment expressed by the reviewer.

**Experiment Processes:**

**1.Text Processing:**

*Tokenization* is the process of dividing review text into individual words or phrases.

*Stop word removal* involves discarding common words such as "the" or "is" that have little impact on sentiment analysis.

*Transforming cases* includes converting words to either uppercase or lowercase to ensure uniformity.

*Stemming* reduces words to their base or root form, aiding in text standardization for analysis.

**2. Selecting Attributes:**

In RapidMiner, the "Select Attributes" operator is used for feature selection, with the goal of choosing a subset of relevant features from a dataset. Feature selection is crucial in machine learning and data mining tasks, as it helps improve model performance, reduce overfitting, and enhance interpretability. Users have the flexibility to manually specify which attributes to include or exclude from the analysis, leveraging domain knowledge or specific needs.

**3. Setting Role:**

Attributes labelled as "label" serve to annotate data and are commonly associated with output attributes in tasks such as text or image classification. By assigning roles to attributes within RapidMiner, you gain control over how your data is managed, allowing for the exclusion of specific attributes from modelling or the designation of certain attributes as different types of variables. This practice fosters consistency throughout the RapidMiner workflow. For example, when you designate an attribute as an output, RapidMiner identifies it as the target variable for predictive models. Clearly defining attribute roles enhances the interpretability of the process by specifying which attributes serve as inputs, outputs, or specialized variables, thereby facilitating comprehension and replication of your analyses by others.

**4.Model Construction:**

Splitting the dataset into training and testing portions aids in both training and assessing the sentiment analysis model. We utilize approaches such as k-Nearest Neighbours (k-NN), Logistic Regression, and others to train the dataset and predict sentiment. Afterwards, we apply the model and assess its performance through cross-validation using various criteria.

**5.Model Evaluation:**

Assessment Criteria: Calculating accuracy, precision, recall, and F1-score to assess the effectiveness of the sentiment analysis model.

**a. Experimental Results, Including Results from Cross Validation:**

Assessing Accuracy using k-NN method for different values of k:

*k = 10*

For 5-fold cross-validation using knn model of 10 nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a graph

Description automatically generated

For 5-fold cross-validation using knn model of *k = 50* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

For 5-fold cross-validation using knn model of *k = 200* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

For 5-fold cross-validation using knn model of *k = 500* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

For 5-fold cross-validation using knn model of *k = 1000* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a graph

Description automatically generated

For 5-fold cross-validation using knn model of *k = 5000* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a graph

Description automatically generated

For 5-fold cross-validation using knn model of *k = 7000* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

For 5-fold cross-validation using knn model of *k = 5000* nearest neighbours and numerical measure = Cosine Similarity

A screenshot of a graph

Description automatically generated

For 5-fold cross-validation using knn model of *k = 5000* nearest neighbours and numerical measure = Manhattan Distance

A screenshot of a graph

Description automatically generated

For 500-fold cross-validation using knn model of *k = 5000* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

**b. Conclusion and Interpretation**

*k-value*

Decreasing the k value heightens the model's responsiveness to local data variations. Smaller k values may establish intricate decision boundaries between classes, which could result in overfitting, particularly in datasets with noise. Conversely, increasing the k value smoothens decision boundaries and enhances the model's robustness against noise. Larger k values reduce the impact of individual data points, resulting in more stable predictions that generalize effectively to new data.

*Folds:*

Expanding the number of folds partitions our dataset into additional subsets for training and testing. Each fold involves training the model on a larger fraction of the data and conducting evaluations multiple times with distinct testing subsets. This approach can enhance performance assessments by subjecting the model to a wider array of data. However, increasing the number of folds also escalates computational demands, necessitating multiple training and evaluation cycles.

On the contrary, reducing the number of folds diminishes the subsets utilized for training and testing. With fewer folds, each model is trained on a smaller portion of the dataset, potentially resulting in less reliable evaluations, especially with limited data. Although fewer folds can accelerate model training and evaluation, performance estimates may exhibit greater variability and may not accurately reflect the model's actual performance on unseen data.

*Numerical Measures:*

Cosine similarity evaluates the likeness between two vectors in a multi-dimensional space by assessing the cosine of the angle between them, indicating their alignment.

Euclidean distance signifies the direct distance between two points in Euclidean space, computed as the square root of the sum of squared differences between corresponding elements of the vectors.

Manhattan distance, also known as city block or taxicab distance, determines the distance between two points by summing the absolute differences of their coordinates. Unlike Euclidean distance, it gauges distance along grid lines rather than in a straight line.

**Text Mining Used for Sentiment Analysis:**

Text mining plays a crucial role in sentiment analysis by extracting sentiments and insights from text data. In marketing, sentiment analysis helps businesses understand customer opinions, attitudes, and emotions towards products, services, or brands. By analyzing IMDb reviews, marketers can assess audience reactions to movies, identify trends, and evaluate the impact of marketing campaigns. Positive reviews can be leveraged for promotional activities, while negative reviews provide valuable feedback for product improvement and customer satisfaction. Ultimately, sentiment analysis empowers marketers to make data-driven decisions, improve customer experiences, and optimize marketing strategies for greater effectiveness.

By,

Arun Babu Indupalli

Vishnu Kacha