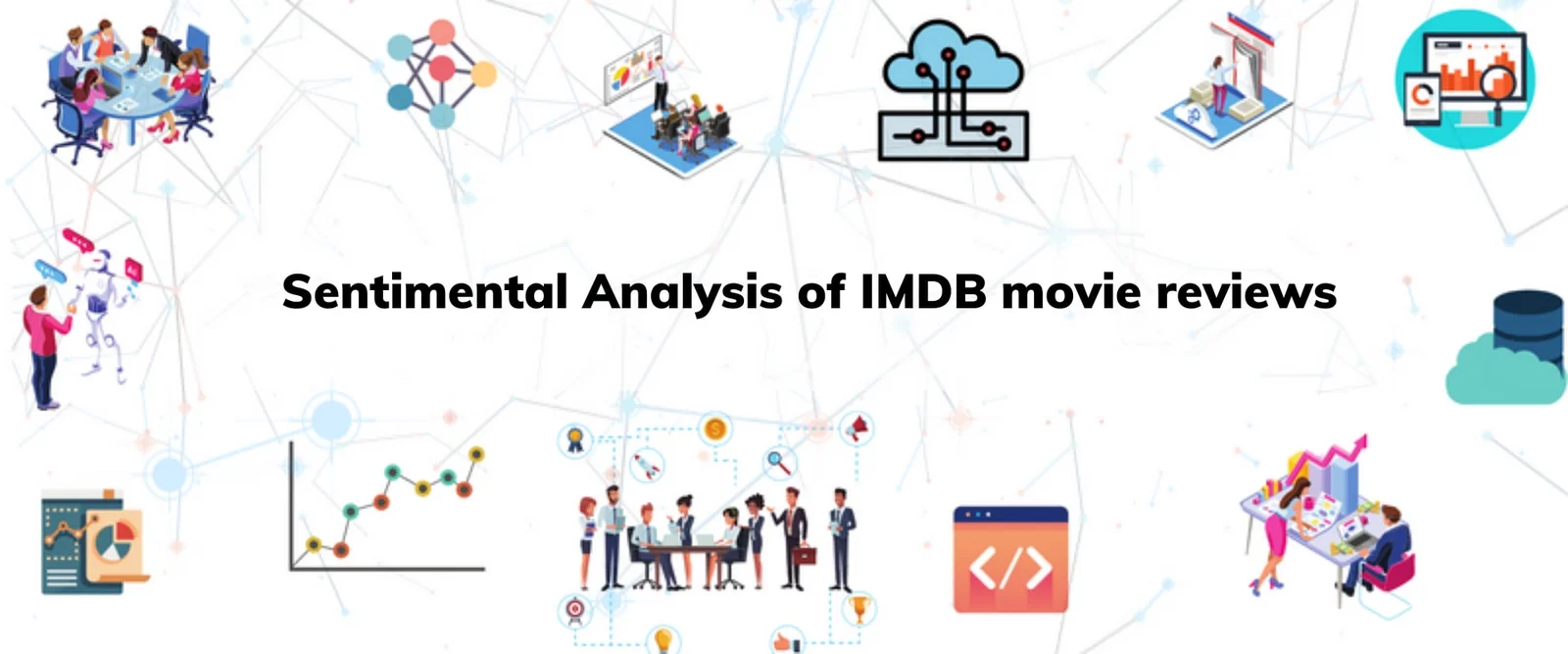
**Text Mining**

**Description:**  
To build a robust sentiment analysis model in RapidMiner with a range of segmentation and classification operators to accurately classify textual input into positive and negative attitudes. The model's accuracy and reliability should be continuously increased through iterative training and testing methods until it operates at a level that either meets or exceeds predetermined satisfaction levels.

**Dataset:**  
The IMDb reviews dataset is made up of Kaggle reviews. Each review contains text feedback from users who have written reviews for various movies. Key characteristics of the IMDb reviews dataset include: Review text Movie title User rating Date of review Reviews are categorized as positive and negative based on how the reviewer feels about the movie.

**Experiment Processes:**

1. **Text Processing:**

*Tokenization:* is used to divide the words or phrases into individual words.

*Stop word removal:* includes getting rid of terms that are often used but don't really affect sentiment analysis, like "the" or "is".

*Transforming cases:* involves changing words to lowercase or capital to maintain consistency.

*Stemming:* cuts words down to their most basic or root form, which helps standardize text for analysis.

**2. Selecting Attributes:**

The “Select Attributes” operator in RapidMiner is used to select a subset of appropriate features from a set of data. In predictive modeling, feature selection is crucial because it enhances interpretability, reduces overfitting, and optimizes model performance. We can manually select which attributes are included or excluded from the analysis.

**3. Setting Role:**

"Label" attributes are frequently connected to output attributes in order to make data annotation easier for tasks like categorizing texts. You may manage your data in RapidMiner by assigning roles to attributes. This enables you to omit some attributes from modeling or to categorize them as variables of a different kind. This instruction supports similarity throughout the whole RapidMiner workflow. For instance, RapidMiner identifies an attribute as the target variable for predictive models when you define it as an output. By identifying what features are utilized as inputs or outputs, you may improve the process's interpretability and make it easier for others to understand.

**4.Model Construction:**

The sentiment analysis model is trained and evaluated more effectively when the dataset is divided into training and testing sets. To train the dataset and predict sentiment, we employ techniques like Logistic Regression and k-Nearest Neighbors (k-NN), among others. Next, we put the model to use and use a variety of criteria to cross-validate the model's performance.

**5.Model Evaluation:**

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

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

*k = 5*

For 3-fold cross-validation using knn model of 5 nearest neighbors and numerical measure = Euclidean Distance

A screenshot of a computer

Description automatically generated

For 20-fold cross-validation using knn model of *k = 5* nearest neighbors and numerical measure = Euclidean Distance

A screenshot of a graph

Description automatically generated

For 20-fold cross-validation using knn model of *k = 5* nearest neighbors and numerical measure = Manhattan Distance

A screenshot of a computer

Description automatically generated

For 20-fold cross-validation using knn model of *k = 5* nearest neighbors and numerical measure = Cosine Similarity

A screenshot of a graph

Description automatically generated

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

A screenshot of a graph

Description automatically generated

For 200-fold cross-validation using knn model of *k = 500* nearest neighbours and numerical measure = Euclidean Distance

A screenshot of a graph

Description automatically generated

Using SVM model

A screenshot of a graph

Description automatically generated

**b. Conclusion and Interpretation**

When we lower the K value in the model, it becomes better at reacting quickly to changes in nearby data. But with smaller "k" values, the model might start making very detailed decisions between different groups, which can lead to overfitting. On the other hand, if we increase the "k" value, the model becomes better. When "k" is bigger, individual bits of data have less impact, and the model's predictions become more stable, which helps when dealing with new data's confusion matrix.

Our dataset is divided into additional subsets for training and testing by increasing the number of folds. Each increase includes preparing the model on a bigger part of the information and leading assessments on different occasions with unmistakable testing subsets. This approach can upgrade execution evaluations by exposing the model to a more extensive cluster of information. In any case, expanding the quantity of folds additionally heightens computational requests, requiring various preparation and assessment cycles.

In reality, when we decrease the number of folds, we're using fewer subsets to train and test the model. Each model gets trained on a smaller part of the data with fewer folds, which might make the evaluations

The cosine similarity measures the degree of similarity between two vectors. It accomplishes this by examining the cosine of the angle formed by them, which provides information about their arrangement.

Euclidean distance measures the straight-line distance between two points. It's calculated as the square root of the sum of squared differences between corresponding elements of the vectors.

Manhattan distance, decides the distance between two focuses by adding the outright contrasts of their directions. Dissimilar to Euclidean distance, it measures distance along grid lines than that of the straight line.

**Sentiment Analysis in Text Mining:**

Text mining is really important in sentiment analysis. It helps by pulling out opinions and insights from written text. In marketing, sentiment analysis is very important. It helps companies understand what customers think and feel about their products, services, or brands. For example, by looking at IMDb reviews, marketers can see how people react to movies, spot trends, and see if their advertising is working. Good reviews can be used in promotions, while bad ones can help improve products and keep customers happy. Overall, sentiment analysis helps marketers make data-driven decisions, improve customer experiences, and make their marketing strategies even better.

By,

Srikar Varma Nadimpalli

Ruthwick Reddy Podduturi