Bheemeshwar Punna
Amisha Gadhia - 11053894

PROJECT REPORT: Sentiment Analysis on IMDb Movie Reviews:

Objective:

The primary objective of this project is to conduct sentiment analysis on IMDb movie reviews. Sentiment analysis, also known as text mining and/or opinion mining, involves identifying and extracting subjective information from textual data to determine the sentiment expressed, typically as positive or negative. In this project, we aim to analyze a dataset comprising 5000 IMDb movie reviews to classify them as either positive or negative sentiments.

Description of the Dataset:

The dataset utilized in this project consists of 5000 movie reviews extracted from IMDb. Each review is accompanied by its corresponding sentiment label, denoting whether the sentiment expressed in the review is positive or negative. The data is balanced, as positive folder contains, separate text file for each positive review, there are 250 such text files in positive folder. The negative folder has 250 text files, each containing negative reviews.

Experiment Processes:

1. Text Processing:

We started building the model with "Process Documents from files" operator to import the text files containing positive and negative reviews, generating each document as a vector output using TF-IDF vector creation setting. There are various parameters which were implemented during the processing phase such as:

Tokenization: Breaking down the text into individual tokens (words or phrases).

Removal of Stopwords: Eliminating common words that do not carry significant meaning (examples: the, and, is, are).

Stemming: Reducing the words to their base or root form to normalize the text (running to run).

Bag of Words Representation: Converting text data into numerical vectors representing the frequency of occurrence of words in each review.

After building the model to train and import the dataset, we moved on to the next phase called Data Partition in model building process.

2. Data Partition:

We have trained the model to perform cross-validation on the dataset in 10 folds to ensure robust model evaluation. Through this approach, we systematically validate our model across different subsets of the data, enhancing the performance of the model.

3. Model Construction:

- Text Data Retrieval: We utilized the "Process Documents from Files" operator to access the movie review data stored in text files. We have created two different labels, namely: pos (positive) and neg (negative). Each label has a directory path set to folders which has 250 text files having positive and negative reviews stored.
- Text Preprocessing: In the preprocessing phase, we prepared the text data for analysis through several essential tasks.
    - In the first step, tokenization divided the text into individual words or tokens, aiding in detailed analysis.
    - Moving forward, lowercasing the tokens to ensure the uniformity of the data.
    - Additionally, removing stopwords eliminated common, insignificant words.
    - In the last step, stemming normalized variables by reducing words to their base form.
- Attribute Selection: The "Select Attributes" operator facilitated the selection of relevant features crucial to sentiment analysis. We selected a combination of parameters in this operator: 1. Type: Include attributes. 2. Attribute filter type: all attributes. This ensures the inclusion of all attributes generated during the preprocessing stage.
- Cross-Validation: Cross-validation integrated data splitting, model training, application and performance evaluation.
    - We specified the 'Number of Folds' parameter to 10, which divides the data into ten subsets for robust training and testing.
    - In the Evaluation parameter, we selected metrics such as accuracy, precision, recall with a focus on precision as the default metric.
- Model Training Approach: For the classification task, we employed the k-NN (k-Nearest Neighbors) operator during the training phase, setting the k value to 5. The 'measure types' parameter was configured to "Numerical Measures" with "Cosine Similarity" parameter, ensuring an effective classification approach. Additionally, we configured the same parameters with another model of "Mixed Measures" with "Mixed Euclidean Distance".

4. Model Evaluation:

The results for the first model where the parameter was set to "Numerical Measures" with "Cosine Similarity" parameter: The overall accuracy of the model is 69% with the accuracy of "pos" and "neg" being 74.40% and 63.60% respectively. To increase the accuracy of the model, we changed the algorithm for the classification from k-NN to Support Vector Machine. The results for Support Vector Machine were satisfactory, better than the first model: The overall

accuracy of the model is 74.60% with the accuracy of "pos" and "neg" being 78% and 71.20% each.

Importance of Text Mining in Sentiment Analysis and Marketing:

Text mining is very important in sentiment analysis, especially in the context of marketing, for extracting valuable insights from textual data, thereby enabling businesses to understand customer sentiment, tailor marketing strategies, and make informed decisions: Here are few cases where text mining is crucial for business development:
1. Customer Feedback Analysis: Text mining enables businesses to analyze customer feedback from various sources, such as social media, product reviews, and surveys, to understand customer sentiment towards their products or services.
2. Market Research: Text mining allows marketers to extract valuable insights from unstructured textual data, aiding in market research, trend analysis and competitor marketing.

In essence, text mining is a game-changer in the world of marketing, transforming the once-daunting task of analyzing massive amounts of unstructured text data into a data-driven, insightful journey. It empowers businesses to craft responsive, optimized, and impactful marketing strategies that truly connect with their customers, driving success in an ever-evolving market landscape.

Conclusion:

This project demonstrates the application of sentiment analysis in analyzing IMDb movie reviews to classify sentiments as positive or negative. By utilizing the text mining techniques and algorithms such as k-NN and SVM (Support Vector Machine) with cross validation method, we have constructed and evaluated sentiment classification models, highlighting the importance of text mining in sentiment analysis and its relevance in marketing strategies.

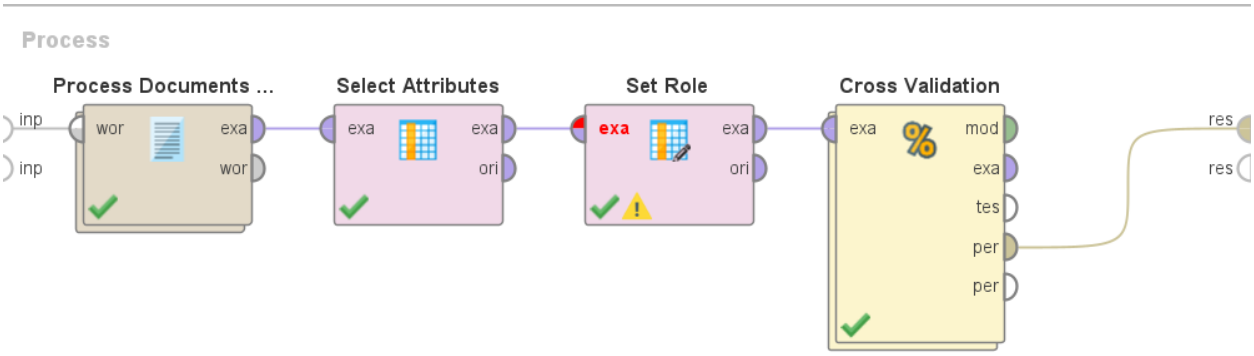

Figure 1: Process – Experiment:

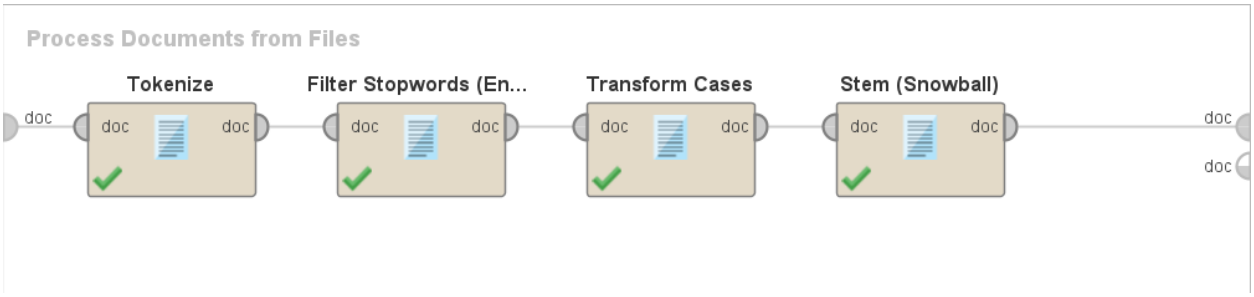

Figure 2: Process – Documents from Files:

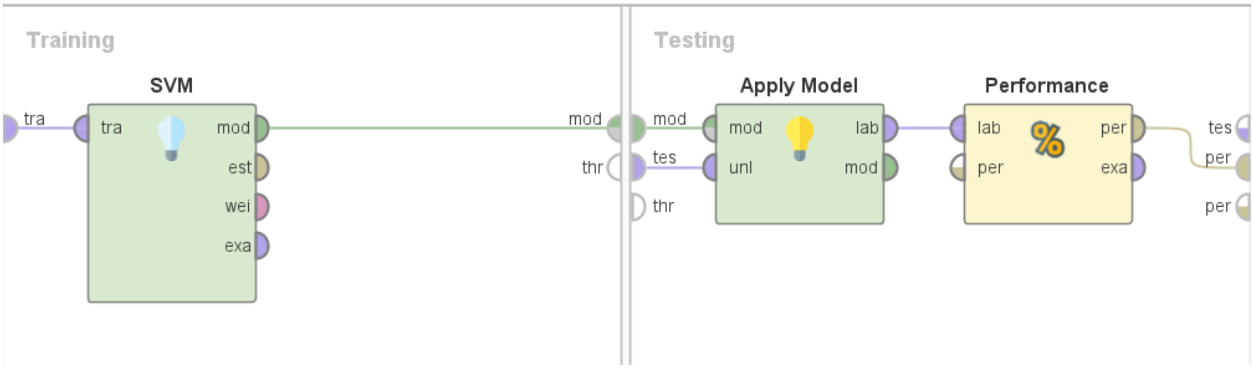

Figure 3: Process – SVM (Support Vector Machine)

accuracy: 74.60% +/- 6.33% (micro average: 74.60%)

	true pos	true neg	class precision
pred. pos	195	72	73.03%
pred. neg	55	178	76.39%
class recall	78.00%	71.20%	

Figure 4: Results – SVM Model

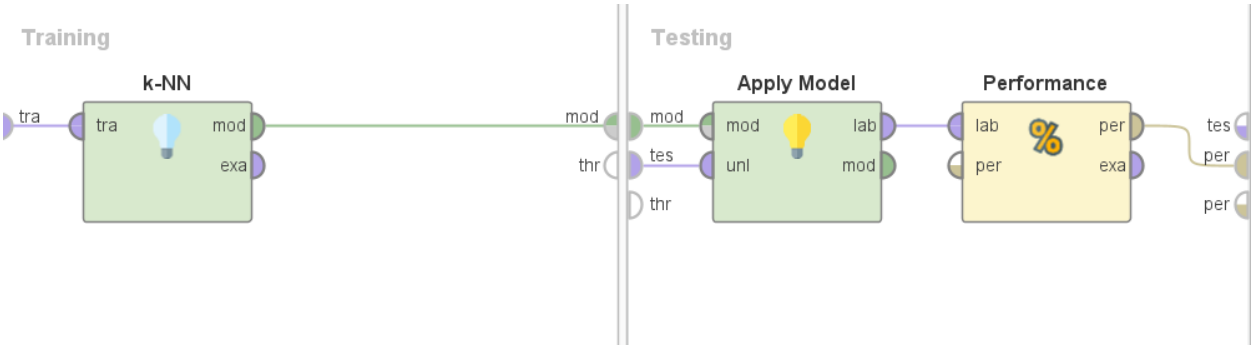

Figure 5: Process – k-NN Model (k=5)

accuracy: 69.00% +/- 6.48% (micro average: 69.00%)

	true pos	true neg	class precision
pred. pos	186	91	67.15%
pred. neg	64	159	71.30%
class recall	74.40%	63.60%	

Figure 6: Results – k-NN Model