**QUANG NHUT LY – 11055575**

**HEMANTH KOTHAPALLI – 11053675**

**MARKETING ANALYTICS PROJECT REPORT**

**OBJECTIVE:**

Utilizing RapidMiner, we aim to leverage advanced sentiment analysis techniques to analyze guest reviews at Disneyland California. The objective is to extract valuable insights into guest experiences, perceptions, and preferences by proficiently categorizing textual data into positive, negative, or neutral sentiments. Further review ratings will be leveraged to identify areas of excellence and improvement, providing actionable insights for marketing strategies. Staying attuned to evolving guest sentiments is essential for Disneyland California to maintain relevance and competitiveness in the market.

**DATASET:**

The data utilized in this study was sourced from the Kaggle website

(https://www.kaggle.com/datasets/arushchillar/disneyland-reviews). To prepare the dataset for sentiment analysis model development, a subset was extracted from the original dataset, which comprises more than 42,000 reviews spanning 3 Disneyland branches - Paris, California, and Hong Kong, as posted by visitors on Trip Advisor. Employing the sample function in R, 1000 observations were randomly sampled to represent negative reviews, 1000 for positive reviews, and another 1000 for neutral reviews. We focused solely on the 'Review\_Text' and 'Rating' fields of the Disney California branch, representing the textual content and rating respectively. Consequently, we categorized 1 and 2-star ratings as negative reviews, 3-star ratings as neutral, and 4 and 5-star ratings as positive reviews. A for loop was then employed to create individual text files for each review, resulting in 1000 files for each category label. These files were subsequently organized into separate folders named 'neg', 'neu', and 'pos', facilitating efficient data management and preprocessing.

**EXPERIMENT PROCESSES:**

1. **Text Processing:** In RapidMiner, text processing is crucial for preparing textual data for sentiment analysis. By performing these text processing steps, the textual data is transformed into a format that is suitable for sentiment analysis, allowing the model to effectively identify and classify sentiments expressed in the text. We processed text files containing positive, negative, and neutral reviews using the 'Process Documents from Files' operator. This converted each document into a TF-IDF vector:

* **Tokenize:** break down the text of each review into individual tokens.
* **Filter Stopwords (English):** remove English stopwords from the documents by eliminating tokens that matched those found in a predefined list of stopwords.
* **Transform Cases:** ensures consistency by converting text to lowercase.
* **Stem (Snowball):** Reduce words to their root form (e.g., "walk" and "walking" become "walk").

1. **Data Partition:** We utilized the 'Cross Validation' operator with 'number of folds' set to 10 to effectively split the dataset into 10 equal subsets (folds). During each iteration of the process, 9 folds were dedicated to training the model, leaving 1-fold aside for testing. This cycle repeated 10 times, ensuring that each fold served as the test set exactly once. This approach helps in mitigating the risk of overfitting and provides reliable insights into its ability to generalize to unseen data.
2. **Model Construction:**

* **Process Document from File:** To prepare the text data for analysis, we engaged in preprocessing, encompassing various tasks: tokenize, stopwords removal, lowcasing and stemming.
* **Select Attribute:** encompassing all attributes generated in the preprocessing stage. by setting the 'attribute filter type' parameter to 'all attributes.'
* **Set Role:** assigned roles to attributes via the Set Role operator, designating the sentiment label and feature attributes essential for sentiment analysis.
* **Cross Validation**: involved data splitting, model training, application, and performance assessment. Key parameters included setting the number of folds to 10 and selecting the automatic metric for sampling type. The k-NN operator is used during training, configuring the k value to 5. Additionally, we specified 'NumericalMeasures' in the 'measure types' parameter and employed 'Euclidean' for the 'numerical measure' parameter.

**Model evaluation:** During our model evaluation phase, we observed that the overall accuracy of our initial model stood at 52.10%, with individual accuracies for 'pos', 'neg', and 'neu' sentiments at 64.5%, 52.6%, and 39.2% respectively. Seeking improved performance, we opted to switch the classification algorithm to Naive Bayes. This adjustment resulted in the second model achieving a higher overall accuracy of 59.27%, with improved accuracies in predicting 'pos', 'neg', and 'neu' sentiments at 70.9%, 64%, and 42.9% respectively. In conclusion, while both models demonstrated modest overall accuracies, they exhibited better performance in predicting 'pos' and 'neg' sentiments compared to 'neu' sentiments.

**TEXT MINING FOR SENTIMENT ANALYSIS:**

Text mining for sentiment analysis involves cleaning and preprocessing text data to remove noise and tokenize it into individual words or phrases. Features are then extracted, such as word frequencies to represent the text numerically. Sentiment classification is performed using k-NN algorithms or other approaches to categorize text into positive, negative, or neutral sentiments. The results are analyzed to gain insights into customer opinions and market trends, empowering businesses to make informed decisions and enhance customer satisfaction.

**IMPORTANCE OF TEXT MINING IN MARKETING:**

Text mining in marketing also facilitates sentiment analysis, allowing businesses to gauge customer satisfaction levels and identify areas for improvement. By analyzing sentiment trends over time, marketing analysts can track the impact of marketing campaigns, product launches, and customer service initiatives. Additionally, text mining enables sentiment-based segmentation, where customers are grouped based on their sentiments and preferences, allowing for personalized marketing strategies and targeted messaging. Furthermore, text mining techniques can uncover emerging trends, customer pain points, and market sentiments, providing marketers with valuable insights for strategic planning and decision-making.

Top of Form

**Bottom of Form**

**Bottom of Form**

A diagram of a computer

Description automatically generated

Diagram 1 – Process - Experiment

A computer screen shot of a computer

Description automatically generated

Diagram 2 -Process – Document form Files

A screenshot of a computer

Description automatically generated

Diagram 3 – Cross Validation - Model 1

A screenshot of a computer

Description automatically generated

Diagram 4 – Cross Validation - Model 2

A screenshot of a graph

Description automatically generated

Diagram 5 – Result – Model 1

A screenshot of a graph

Description automatically generatedDiagram 6: Result – Model 2

A screenshot of a computer

Description automatically generated

Diagram 7 – R-studio Code 1

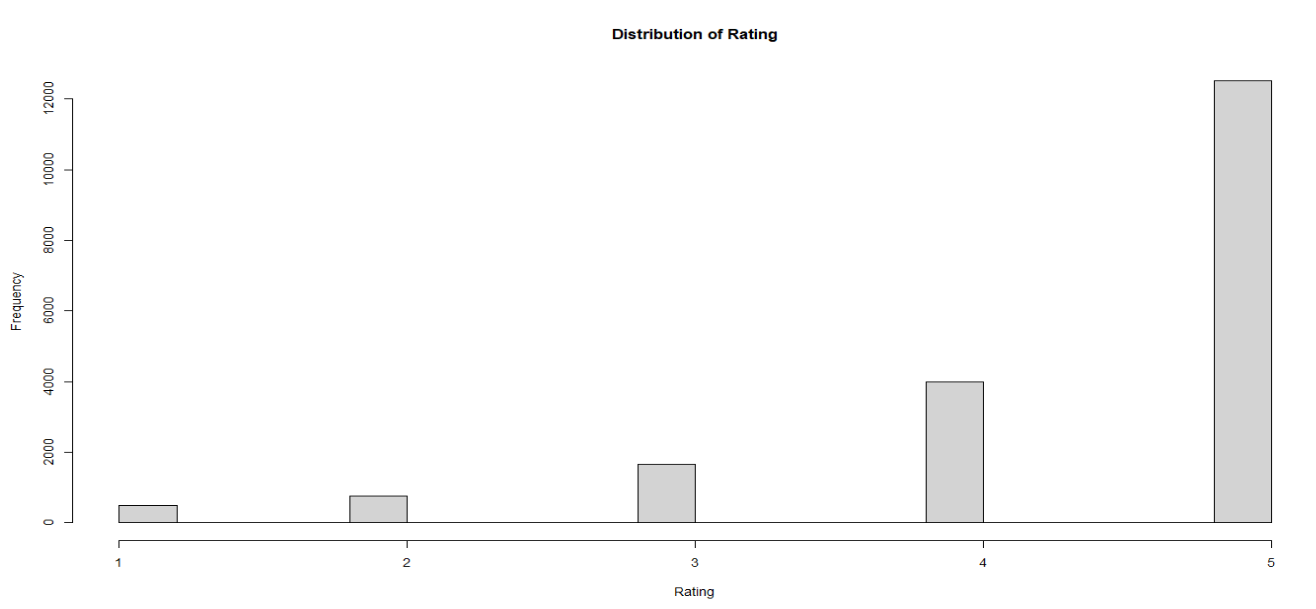


Diagram 8 – Distribution of Rating

A screenshot of a computer code

Description automatically generated

Diagram 9 – R-Studio Code 2

A screenshot of a computer code

Description automatically generated

Diagram 10 – R-Studio Code 3

A screenshot of a computer code

Description automatically generated

Diagram 11 – R-Studio Code 4