Sentiment analysis on IMDB movie reviews

Project Team Members:

Sai Dheeraj Surampally (Team leader): A20545700
Sanjana Rayarala: A20548132
Narahari Rahul: A20545662

1. Project Proposal

- **Research Goal**
  The goal of this research is to perform sentiment analysis on IMDB movie reviews to develop an accurate and interpretable model that can effectively classify the sentiment (positive, negative) of user-generated text reviews. This analysis aims to provide valuable insights into the sentiment expressed by reviewers, enabling a deeper understanding of audience opinions on movies and their associated factors.
- **Research Questions**
  1. How accurately can movie reviews be classified as "positive" or "negative" based on their textual content?
  2. What text preprocessing techniques are most effective in preparing the text data for sentiment analysis?
  3. Are there specific keywords or phrases that are strongly associated with "positive" sentiment in movie reviews?
  4. What linguistic and textual patterns differentiate "positive" reviews from "negative" ones in the dataset?
  5. What is the impact of review length on sentiment classification accuracy? Is there an optimal review length for accurate sentiment analysis?
- **Proposed Methodology**
  1. Data Collection and Preparation:
     - Collecting the IMDB dataset of 50k movie reviews, including two columns: "review" and "Sentiment."
     - Then we preprocess the textual data by removing stopwords, special characters, and converting text to lowercase.
  2. Data Exploration:
     - Explore the dataset to understand its characteristics, such as review length distribution, sentiment distribution.
  3. Text Preprocessing:
     - Implement various text preprocessing techniques to prepare the text data for sentiment analysis. This includes techniques such as stemming, lemmatization, and removing HTML tags.
  4. Feature Extraction:
     - Create a document-term matrix or word embeddings to represent the textual data for modeling.
  5. Sentiment Analysis Model:
     - Train and evaluate a sentiment analysis model. Using Naive Bayes classifier.
     - Split the dataset into training and testing sets to assess model accuracy.
  6. Model Evaluation:
     - Evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

7. Identify Keywords and Phrases:
   - Use text analysis techniques to identify specific keywords and phrases strongly associated with positive sentiment.
8. Analyze Textual Patterns:
   - Investigate linguistic and textual patterns that differentiate positive reviews from negative ones.
9. Impact of Review Length:
   - Explore how the length of movie reviews impacts sentiment classification accuracy.
   - Analyze if there is an optimal review length for accurate sentiment analysis.
10. Visualization and Reporting:
   - Create visualizations such as word clouds, sentiment word lists, and temporal sentiment trends.
   -Prepare a comprehensive report summarizing the findings, insights, and recommendations.

2. Project Outline

- **Data sources:**
  (Kaggle) https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data
- **Dataset Description**
  Dataset Description: IMDB Dataset of 50K Movie Reviews(Kaggle)
  The "IMDB Dataset of 50K Movie Reviews" is a popular and widely used dataset for sentiment analysis and natural language processing tasks. This dataset is intended for binary sentiment classification, where the goal is to determine whether a movie review expresses a "positive" or "negative" sentiment. The dataset consists of two main columns:

  1. Review (Textual Data):
  - This column contains a collection of 49,582 movie reviews, each in textual format.
  - Movie reviews vary in length and content, reflecting the diversity of opinions and expressions found in user-generated reviews.
  - The reviews are written by users who have shared their thoughts and feelings about movies, making this dataset a valuable resource for sentiment analysis.

  2. Sentiment (Categorical Data):
  - This column provides sentiment labels for each movie review.
  - It contains two unique sentiment values: "positive" and "negative."
  - Sentiment labels are assigned based on the overall tone and opinion expressed in the corresponding review.
  - "positive" sentiment typically indicates a favorable opinion, while "negative" sentiment suggests an unfavorable or critical viewpoint.

  Key Information:
  - The dataset is frequently used for machine learning and natural language processing tasks, particularly for binary sentiment classification.
  - It serves as a benchmark dataset for testing and developing sentiment analysis algorithms and models.
  - The dataset's balance of positive and negative sentiments allows for reliable model evaluation and benchmarking.

- The textual data within the reviews provides diverse and authentic language expressions, making it suitable for text analysis and natural language processing.

Sample Entry:
- Review: "This movie was absolutely fantastic! I loved the acting and the storyline. A must-watch!"
- Sentiment: "positive"

- **Literature review and related work**
  Yu, Hong, et al. "Improved movie review sentiment classification using sentence-level lexicon polari." Expert Systems with Applications 36.2 (2009)

  The paper titled "Improved Movie Review Sentiment Classification using Sentence-level Lexicon Polari" by Yu, Hong, et al. (2009) presents an approach to sentiment analysis in the domain of movie reviews using a sentence-level lexicon-based method. Sentiment analysis, also known as opinion mining, is a burgeoning field of natural language processing and machine learning that aims to extract and classify subjective information from text, typically into positive or negative sentiments. In the context of this research, the authors focus on improving the accuracy of sentiment classification for movie reviews, a domain of high interest for both academia and industry.

  Joshi, Mahesh, et al. "Movie reviews and revenues: An experiment in text regression." Proceedings of the 5th international conference on Weblogs and Social Media. 2011.

  The influence of online user-generated content, such as movie reviews, on consumer decision-making has become a topic of significant interest in recent years. Joshi, Mahesh, et al. (2011) contribute to this body of research with their paper "Movie reviews and revenues: An experiment in text regression." The study explores the relationship between text-based movie reviews and box office revenues, shedding light on the potential impact of online sentiment analysis on the film industry.
  The paper by Joshi et al. (2011) adopts a text regression approach to address this gap in the literature. Text regression involves modeling the relationship between textual features (in this case, movie reviews) and a quantitative outcome variable (box office revenues). Text regression is a subfield of natural language processing (NLP) that seeks to make predictions based on text data (Hastie, Tibshirani, & Friedman, 2009). In this context, the authors aim to develop a predictive model that leverages the sentiment and content of online reviews to estimate box office revenues.

  Pak, A., & Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining.

  Pak and Paroubek's paper delves into the methodology used to tackle the challenges of Twitter sentiment analysis. They employ several strategies for data preprocessing, tokenization, and feature extraction, emphasizing the importance of handling URLs, hashtags, and mentions. The paper explores various classification algorithms, including Naive Bayes and Support Vector Machines, to assess their effectiveness in classifying tweets as positive, negative, or neutral.

- **Data processing and pipeline:**

  Data Cleaning:
  Text Preprocessing: Perform text preprocessing on the "review" column, which includes removing special characters, converting text to lowercase, and eliminating stopwords.
  Handling Missing Values: The dataset is complete and contains no missing values.

  Data Transformation:
  Text Tokenization: Tokenize the textual data in the "review" column. Tokenization is the process of splitting text into individual words or tokens. It helps prepare the text data for feature extraction.
  Feature Extraction: Create a document-term matrix or word embeddings to represent the textual data for modeling. A document-term matrix is created as a bag-of-words representation.

  Data Splitting:
  Split the dataset into training and testing sets. The training set is used to train the sentiment analysis model, while the testing set is used to evaluate its performance.

  Sentiment Analysis Model:
  Train and evaluate the sentiment analysis model. Naive Bayes classification model is used.

  Model Evaluation:
  Evaluate the model's performance using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

- **Data stylized facts:**
  Data Source: The dataset consists of 49,582 movie reviews collected from the IMDB platform. Each review is labeled with one of two sentiments: "positive" or "negative." This binary sentiment classification provides the basis for sentiment analysis.

  Data Distribution: The dataset contains a fairly balanced distribution of "positive" and "negative" sentiments, allowing for robust sentiment analysis and model training. Exploratory data analysis (EDA) reveals that the number of "positive" and "negative" reviews is roughly equal.

  Textual Format: The "review" column contains the textual content of the movie reviews. These reviews are in natural language and vary in length, reflecting the diversity of opinions and expressions among reviewers.

  Text Preprocessing: As part of the data preprocessing stage, text data has been cleaned and preprocessed. This includes the removal of stopwords, special characters, and the conversion of text to lowercase. These preprocessing steps are essential for preparing the text data for sentiment analysis.

  Feature Extraction: Extracts features from the text data for modeling. This typically involves creating a document-term matrix or using word embeddings to represent the reviews in a format suitable for machine learning algorithms.

Sentiment Analysis Model: Employs a sentiment analysis model, specifically a Naive Bayes classifier, to classify movie reviews as "positive" or "negative." This model is trained and evaluated to assess its accuracy and effectiveness in sentiment classification.

Performance Metrics: The model is evaluated using various performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provide insights into the model's effectiveness in correctly classifying sentiment.

Textual Patterns: Includes an analysis of textual patterns that differentiate "positive" reviews from "negative" ones. Investigating linguistic and textual patterns helps understand what aspects of reviews contribute to their sentiment.

Review Length Analysis: Explores the relationship between review length and sentiment. It examines how the length of movie reviews impacts sentiment classification accuracy. A boxplot is used for visualizing the distribution of review lengths for both "positive" and "negative" reviews.

Visualization and Reporting: Includes visualization components, such as word clouds and sentiment word lists, to provide a graphical representation of textual patterns and sentiment-related terms. These visualizations enhance the interpretability of the sentiment analysis results.

- **Model Selection:**
  Naive Bayes classifiers are simple and effective for text classification tasks. They work well with text data and can handle high-dimensional feature spaces efficiently.

- **Software packages, applications, libraries and associated tools**
  Softwares:
  R studio, R

  Libraries:
  `tm`,`tm.plugin.webmining`,`stringr`,`ggplot2`,`tidyverse`,`data.table`,`wordcloud2`,`tm.plugin.sentiment`,`SnowballC`,`quanteda`,`tm.plugin.lexicon`, `e1071`, `caret`

  Project Management & Source control:
  Kaggle Dataset, GitHub