Sentiment Analysis of Medium App Reviews from Google Play Store

Milestone 1: Project Initialization and Planning Phase

The "Project Initialization and Planning Phase" marks the project's outset, defining goals, scope, and stakeholders. This crucial phase establishes project parameters, identifies key team members, allocates resources, and outlines a realistic timeline. It also involves risk assessment and mitigation planning. Successful initiation sets the foundation for a well-organized and efficiently executed machine learning project, ensuring clarity, alignment, and proactive measures for potential challenges.

Activity 1: Define Problem Statement

Manual review of large-scale, user-generated app feedback is time-consuming and inconsistent. Product managers and developers struggle to extract meaningful insights due to the volume and unstructured nature of reviews, which results in delayed updates and uninformed decisions. This project aims to automate sentiment analysis on Medium app reviews from the Google Play Store to extract actionable insights.

Activity 2 : Project Proposal

The project proposes a sentiment classification system leveraging deep learning models. It will automate the categorization of user reviews into positive, neutral, or negative sentiments. A web UI built using Flask will provide real-time predictions and visual analytics. The outcome will enhance decision-making for product development and user engagement strategies.

Activity 3 : Project Planning

Initial project planning involved outlining key objectives, defining the scope of the sentiment analysis system, and identifying stakeholders including app developers and product managers. This phase included setting up a sprint schedule, allocating computational resources (GPU/CPU), and understanding the dataset's structure and quality. The team also formulated goals for preprocessing and model development, ensuring a structured workflow. Effective planning laid a solid foundation for building a robust, data-driven solution.

Milestone 2: Data Collection and Preprocessing Phase

The Data Collection and Preprocessing Phase involves executing a plan to gather relevant reviews data from Kaggle, ensuring data quality through verification and addressing missing values. Preprocessing tasks include cleaning, encoding, and organizing the dataset for subsequent exploratory analysis and machine learning model development.

Activity 1 : Data Collection Plan, Raw Data Sources Identified, Data Quality Report

The dataset for "Sentiment Analysis of Medium App Reviews" is sourced from Kaggle. It includes user-generated reviews along with metadata such as app version, review date, and developer replies. Data quality is ensured through systematic handling of missing values, removal of redundant fields, and text normalization techniques, establishing a reliable foundation for sentiment classification.

Activity 2 : Data Quality Report

Missing fields like `reviewCreatedVersion`, `replyContent`, and `repliedAt` were filled using mode. - Redundant columns like `reviewId`, `appVersion`, etc., were dropped. - Text issues like inconsistent casing, special characters, and stopwords were cleaned using regex and NLTK.

Activity 3 : Data Exploration and Preprocessing

Text normalization: lowercasing, punctuation removal - Stopword removal using NLTK - TF-IDF vectorization (with min_df=2 and ngram_range=(1,3)) to convert text into numerical format

---

Milestone 3: Model Development Phase

The Model Development Phase entails crafting a predictive model for loan approval. It encompasses strategic feature selection, evaluating and selecting models (Random Forest, Decision Tree, KNN, XGB), initiating training with code, and rigorously validating and assessing model performance for informed decision-making in the lending process.

Activity 1 : Model Selection Report

The Model Selection Report details the rationale behind choosing KNN, Naïve Bayes, Random Forest, Logistic Regression, and deep learning models (LSTM, BiLSTM) for sentiment classification. It considers each model's strengths in handling complex relationships, interpretability, adaptability, and overall predictive performance, ensuring an informed choice aligned with project objectives.

Activity 2 : Initial Model Training Code, Model Validation and Evaluation Report

The Initial Model Training Code employs selected algorithms on the app reviews dataset, setting the foundation for predictive modeling. The subsequent Model Validation and Evaluation Report rigorously assesses model performance, employing metrics like accuracy and precision to ensure reliability and effectiveness in predicting sentiment.

Milestone 4: Model Optimization and Tuning Phase

The Model Optimization and Tuning Phase focused on refining machine learning models for peak performance through systematic hyperparameter tuning and performance evaluation.

Activity 1: Hyperparameter Tuning

Logistic Regression was optimized by adjusting the regularization strength (C), solver type, and penalty function to balance model complexity and generalization. K-Nearest Neighbours (KNN) was fine-tuned by varying the number of neighbours, weight functions, and distance metrics to improve local pattern recognition. For Naive Bayes, var_smoothing was tuned to enhance stability in probabilistic predictions. Random Forest underwent tuning of key parameters such as the number of estimators, maximum tree depth, minimum samples for node splitting, and maximum features considered for splitting, resulting in improved ensemble learning performance.

Activity 2: Final Model Selection

After evaluating all models using validation accuracy and performance metrics, Random Forest was selected for final deployment. It outperformed others in terms of predictive accuracy, robustness, and interpretability. Its ability to handle feature interactions and reduce overfitting made it the most suitable model for the task.

Milestone 5: Project Final Submission

Activity 1 : Deployment

Flask-based web application built for sentiment prediction. - Users can input new reviews and get real-time sentiment feedback.

Conclusion This project successfully implemented an end-to-end sentiment analysis pipeline for Google Play Store reviews. The system effectively predicts user sentiment and aids app developers in understanding user satisfaction and feedback trends. The final deployed model is accessible via a user-friendly web interface.

GitHub Repository Link : https://github.com/Ashwin-S-Narayanan-2005/analysis-of-medium-app-reviews-from-google-play-store