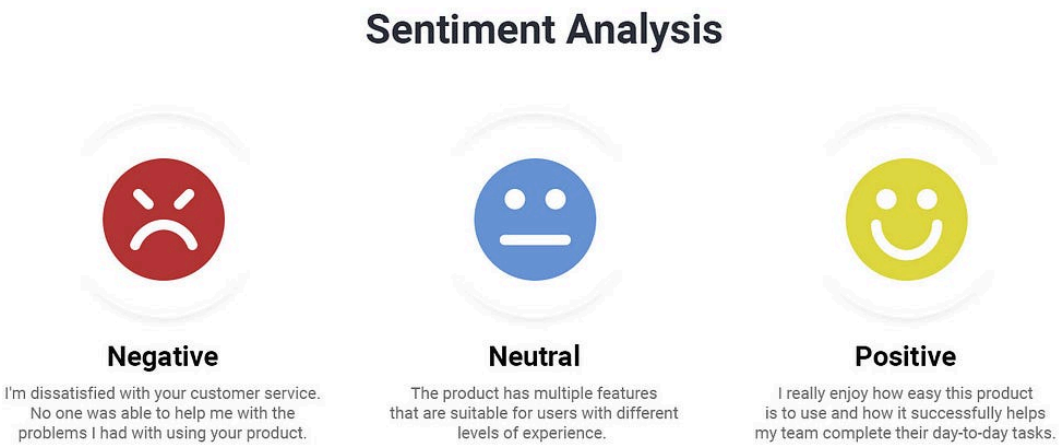
Course-End Project [Capstone Project 2]

Project Domain: E-commerce



Introduction to the Project:

The course-end project analyzes sentiments expressed in over 34,000 reviews for Amazon brand products within the e-commerce domain. The dataset contains attributes such as brand, categories, review titles, review text, and sentiment levels categorized into "Positive," "Negative," and "Neutral." The project aims to predict sentiment or satisfaction levels based on various features and review text.

Objectives of the Project:

- Understand the sentiment expressed in consumer reviews.
- Address class imbalance in sentiment categories.
- Implement classifiers and advanced techniques for sentiment analysis.
- Evaluate model performance using appropriate metrics.
- Compare traditional machine learning algorithms with neural network approaches.
- Explore topic modeling techniques for clustering similar reviews.

Project Tasks:

Week 1 & 2: Class Imbalance Problem

Exploratory Data Analysis (EDA):

During the first two weeks, the focus will be on understanding the characteristics of positive, negative, and neutral reviews within the dataset. Exploratory Data Analysis (EDA) techniques will be employed to visualize the distribution of sentiment categories and identify any patterns or trends. Additionally, the class imbalance issue will be addressed by examining the class counts to understand the distribution of sentiments and to determine the extent of class imbalance present in the dataset.

Feature Engineering:

Feature engineering plays a crucial role in building effective machine-learning models. In this phase, the reviews will be transformed into Tf-Idf (Term Frequency-Inverse Document Frequency) scores. Tf-Idf is a statistical measure that evaluates the importance of a word in a document relative to a collection of documents, and it will help in representing the review text as numerical features suitable for machine learning algorithms.

Classifier Selection:

To begin the modeling process, a multinomial Naive Bayes classifier will be implemented. Naive Bayes is a popular choice for text classification tasks due to its simplicity and efficiency. The classifier will be trained on the transformed features to predict the sentiment of the reviews. Challenges associated with class imbalance, such as biased predictions towards the majority class, will be recognized and considered during model evaluation.

Tackling Class Imbalance:

Addressing class imbalance is critical to ensure the model's effectiveness in predicting sentiments across all classes. Techniques such as oversampling (increasing the number of minority class samples) or under-sampling (reducing the number of majority class samples) will be applied to balance the class distribution. By alleviating class imbalance, the model's performance can be improved, leading to more accurate predictions for all sentiment categories.

Evaluation Metrics:

Model performance will be evaluated using a variety of evaluation metrics, including precision, recall, F1-score, and the Area Under the Receiver Operating Characteristic (ROC) curve (AUC-ROC). These metrics provide insights into the classifier's ability to correctly identify positive, negative, and neutral sentiments. Emphasis will be placed on the F1-score, which balances precision and recall, making it suitable for evaluating models in the presence of class imbalance.

Week 3 & 4: Model Selection and Advanced Techniques

Multi-class SVM and Neural Nets:

In the following weeks, more sophisticated models such as multi-class Support Vector Machines (SVM) and neural networks will be implemented. SVMs are powerful classifiers capable of handling multi-class classification tasks. Similarly, neural networks, particularly deep learning models, have shown promising results in text classification tasks due to their ability to capture complex relationships within the data.

Ensemble Techniques:

Ensemble methods, which combine multiple base classifiers to improve predictive performance, will be explored. Techniques such as XGBoost combined with oversampled multinomial Naive Bayes will be considered. Ensemble methods can effectively mitigate the effects of class imbalance and enhance the overall performance of the sentiment analysis model.

Feature Engineering:

Additional feature engineering will be performed to enhance the predictive power of the models. A sentiment score feature will be engineered and integrated into the models for performance comparison. This sentiment score will capture the overall sentiment expressed in the reviews and may provide valuable information to improve the accuracy of the predictions.

LSTM Implementation:

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network (RNN), will be applied to the sentiment analysis task. LSTMs are well-suited for processing sequential data such as text due to their ability to capture long-term dependencies. Parameter tuning will be conducted to optimize model performance, including adjustments to parameters such as the top word, embedding length, dropout, epochs, and layers.

Comparison:

The performance of the neural network models will be compared with traditional machine learning algorithms trained in the earlier stages of the project. This comparison will provide insights into the effectiveness of deep learning approaches for sentiment analysis tasks and may guide the selection of the final model architecture.

Optimization:

Model optimization techniques will be employed to fine-tune the parameters of the LSTM and other deep-learning models. Techniques such as Grid Search, Cross-Validation, and Random Search will be utilized to determine the optimal settings for the models. By optimizing model parameters, the performance of the sentiment analysis models can be further improved, leading to more accurate predictions.

Topic Modelling:

In addition to sentiment analysis, topic modeling techniques will be explored to cluster similar reviews based on different aspects such as device features, aesthetics, and performance. Techniques such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) will be used to identify latent topics within the reviews and group them accordingly.

Report on EDA:

The exploratory data analysis revealed insights into the distribution and characteristics of positive, negative, and neutral reviews. Class imbalance was addressed by examining class counts and applying oversampling or under-sampling techniques.

[Include pictures of the graphs generated during EDA]

Learning Outcomes:

This project gives practical experience in sentiment analysis, classification techniques, and advanced modeling approaches. Understanding of class imbalance handling, model evaluation, and optimization techniques.

Conclusion:

By systematically addressing the class imbalance problem, selecting appropriate classifiers, and exploring advanced modeling techniques, this project aims to develop an accurate and robust sentiment analysis model for e-commerce applications. Insights gained from the project will contribute to the understanding of sentiment analysis methodologies and their applications in real-world scenarios.

Citations:

- [List any relevant books, research papers, or websites used for the project]

Files to be Submitted:

Jupyter Notebook or Google Colab file
Project Summary Report (Word File) or ppt