PART A.

Introduction

This report describes my approach to a text categorization project that included Natural Language Processing (NLP) techniques. The objective was to appropriately categorize messages by processing and evaluating their content. My technique included data preparation, exploratory data analysis, and model selection, resulting in the deployment of a machine learning model that made meaningful predictions.

Data Overview and Preprocessing

The collection included text documents, each with a title and content. My initial step was to combine these two fields into a single text feature for each document, which ensured a thorough examination. To standardize the dataset, I cleaned the text by removing HTML elements and punctuation and changing it all to lowercase. Recognizing the significance of linguistic context, I eliminated Greek stopwords to concentrate on the most important words.

Feature Deletion

After the text was cleaned, I created a single text feature that included both the document's title and content. By using this method, it was made sure that the model could take in all of the context that the text's body and title offered.

Greek stopwords were eliminated from the text in order to further improve the data. If stopwords are used excessively, they can lessen the significance of the qualities because they are frequently used and typically lack context. By eliminating these words using a predetermined list of Greek stopwords, I was able to focus on the texts' more insightful content.

Converting the cleaned and consolidated text data into a machine-learning-ready format was the next critical step. I transformed the text into a matrix of TF-IDF features using the TF-IDF Vectorizer, choosing parameters that maximized the text's representation in a high-dimensional feature space.

Model Building and Selection

I decided on a Linear Support Vector Machine (SVM) as the classification model because of its dependability and effectiveness when processing high-dimensional data. I trained the SVM model by dividing the dataset into training and validation sets, then fine-tuning it to achieve a satisfactory level of complexity and

performance.

I was able to evaluate the model with remarkable accuracy on the validation set. Precision, recall, and F1 scores for every class were among the comprehensive insights into the model's performance that the classification report offered. These measurements were very helpful in determining the model's areas of strength and improvement.

Alternative Models

Group Techniques: Take into consideration doing experiments using group techniques like Gradient Boosting Machines or Random Forests. Occasionally, these models are able to extract more complicated patterns from the data than a single SVM model.

Deep Learning: Examine deep learning models like as Long Short-Term Memory (LSTM) networks, which are ideal for sequential data such as text, or Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs). Text data's context and semantic linkages can be more intricately captured by these models.

Transfer Learning: Use models that have already been trained, such BERT or GPT, and refine them using the data you have. Transfer learning, which makes use of knowledge from larger datasets, can dramatically improve performance, particularly in situations where there is a shortage of training data.

Evaluation

The model's performance was evaluated using accuracy and a complete classification report that included information on precision, recall, and F1-scores across the various classes. This evaluation helped to fine-tune the model and technique, ensuring that the forecasts were robust and reliable.

Testing and Predictions

Finally, I applied the trained model to a previously unseen collection of test data, demonstrating its capacity to generalize and accurately classify new documents. The findings confirmed the model's usefulness and efficacy in a real-world situation, indicating a possible route for further research and implementation.

Final Thoughts

Using TF-IDF for feature extraction and preprocessing text data, I created and assessed a Linear Support Vector Machine (SVM) model for text classification in this research. The model showed encouraging results, successfully and accurately

classifying text into predetermined categories. Thorough data cleaning, eliminating Greek stopwords, and converting text data into a numerical format appropriate for machine learning were crucial project milestones. The evaluation metrics emphasized the SVM model's capacity to produce precise predictions, and it was selected because to its expertise with high-dimensional datasets.

Future Tasks

In the future, this project could be extended in a number of areas to improve performance and investigate new avenues:

Advanced Model Tuning: The performance of the model may be enhanced by investigating SVM hyperparameters in greater detail using more sophisticated grid search methods. Improvements may also be achieved by taking into account different loss functions and regularization techniques.

Ensemble and Deep Learning Models: Using ensemble techniques like Gradient Boosting or Random Forests in experiments may improve the robustness of the model. For capturing more intricate textual patterns and context, deep learning techniques—especially those that make use of pre-trained models like BERT or LSTM networks—represent an interesting new area.

Investigating methods for data augmentation: such as back-translation or synonym replacement, could improve the training dataset and, consequently, the model's capacity to generalize from sparse data.

PART B

Introduction

This research provides a comprehensive examination of bidding activity, drawing on a large dataset to investigate trends, device usage, regional distribution, bidding behavior, and ad performance indicators. I want to identify patterns and insights that will influence bidding strategies and improve ad effectiveness through thorough preparation and exploratory data analysis (EDA).

Data Overview and Pre-processing

The dataset loaded from 'analysis.csv' contains a variety of attributes, including bid identifiers, device information, country codes, domain details, and impression success flags. My first steps were converting date fields to datetime format for precise time series analysis and determining data quality by detecting missing values.

Exploratory Data Analysis (EDA)

- *Missing Value Analysis.*

I began by determining the percentage of missing values across columns, laying the groundwork for intelligent data cleaning and imputation procedures that would enable robust analysis.

- *Trends across time*

A group-by operation on the 'day' column revealed daily bidding activity and impression counts, providing insights into temporal trends that may impact bidding tactics.

- *Device Usage Analysis*

Device popularity was determined using value counts, which highlighted the presence of certain devices in bidding activities and might influence device-targeted advertising initiatives.

- *Geographic Distribution*

Analyzing the top ten nations by auction volume offered a geographical lens through which bidding tactics might be targeted for optimal success.

- *Bidding Practices*

I examined average bid values and success rates for various bidders and domains, finding trends that might indicate chances for optimization or successful bidding strategies.

- *User Consent's Effect on Impressions*

By classifying data according to user consent, it became clear how important it is to satisfy user preferences and legal obligations by taking impression rates into account.

- *Metrics for Ad Performance*

In order to determine which forms and dimensions result in the highest engagement and success rates, I looked into bids and impression rates across a range of media kinds and ad sizes.

Visualizations

To illustrate what I found, a number of plots were created, each designed to draw attention to a different feature of the dataset:

1. Average Bid Values: To show the possible influence of bid amount on ad visibility, a bar plot was created by comparing the average bid values for successful and failed impressions.

2. Success Rate Among Domains and Bidders: The top domains and bidders were displayed using horizontal bar plots based on success rate, indicating regions where bidding is highly successful or requires modification.

3. Impact of User Consent on Impressions: The relationship between user consent and impression rates was displayed as a bar plot, highlighting the importance of consent for the success of advertisements.

4. Bids and Impression Rates by Media Type and Ad Size: These charts provide media planners and ad designers insight into how various media kinds and ad sizes perform in terms of drawing bids and producing impressions.

5. Geographic and Device Distribution: To help with targeted strategies, additional plots were created to show how auctions were distributed around nations and what kinds of devices were most common.

6. UTM Campaign Analysis: I determined which UTM campaigns were the most successful by examining their total bids and impressions. This provided a foundation for improving marketing tactics.

7. Timing Analysis and Auction Length: Examining how long auctions last and when bids are placed in relation to the end of the auction provided information about how these variables affect success rates and could help bid timing techniques.

8. Device and OS Success Rates: To better target advertisements to certain technology usage patterns, a heatmap displaying success rates broken down

by device and operating system offered a detailed perspective on how technical settings impact ad performance.

Conclusions

The analysis revealed a number of significant variables that affect the outcome of bidding, including temporal trends, device preferences, regional patterns, and the critical function of user consent. These discoveries not only deepen our comprehension of the information but also provide practical advice for maximizing ad placements and bidding tactics. Utilizing these results going forward might greatly enhance ad effectiveness and engagement, resulting in better outcomes for advertisers and a more tailored experience for users.

Future Work

Future analyses could go more deeply into the interplay between variables, such as how certain device and operating system combinations affect the effectiveness of ads or how outside variables like the time of day and market circumstances affect bidding behavior. To further improve ad effectiveness, machine learning models may also be used to forecast bidding success and optimize bid amounts in real-time.