CS-412 Term Project

Bülay Ecem Çakallı

30788

I had to do this project on my own because there were not people on auditorium and the people, I met had already formed their group. I tried to find someone until the end of round 1 however it did not work very well in the end. The first two rounds I sent are not correct, since I could not find a teammate, I couldn't do much for implementing a code. Therefore, I put new JSONS for these rounds. Since I don't have a teammate

Task-1

The goal of Task 1 is to classify Instagram influencers into specific categories based on the profile data provided. We've worked on understanding the given data, building a classification model, and making predictions for the sample inputs.

Feature Extraction

This section focuses on loading and inspecting data from a CSV file (my_training.csv) containing user information and categories. I used the example code on the courses google drive for this seciton and made some changes on that provided code such as adding more metada columns. Since, adding metadata columns like follower_count, following_count, is_business_account, is_private, and is_verified improves the model by providing user-level features that complement text-based data. This multi-modal approach enhances performance and generalization by combining textual and non-textual insights for robust classification.

Model and Training

I split the dataset into training and validation sets using train_test_split to evaluate the model's performance on unseen data. I initialized and trained a Logistic Regression model using the combined TF-IDF and metadata features. I calculated the training accuracy and made predictions on the validation set to compute validation accuracy. I also generated a detailed classification report to assess performance metrics like precision, recall, and F1-score. I chose Logistic Regression because at first I also tried Random Forest but the validation results were better in Logistic Regresion.

Task-2

In this task, the goal was to build a regression model to predict the "like count" for social media posts. I tried to make this task by using features extracted from captions, hashtags, and other metadata, then save the predictions in a JSON format where keys are post IDs and values are predicted like counts. I tried to use Logistic Regresion, however I gave an error about y_test being empty therefore I could not create a JSON file for regression. In the first 2 rounds I did create some of them but they were rather wrong. I deleted the code because of that. Now I do not have a code that generates the regression.