AML Project Deliverable #1 - Project Proposal

Group 7: Shivank Agrawal, Ruoheng Du, Yiwen Ge, Yandong Xiang

The rapid rise of misinformation and inappropriate content on platforms like Twitter has driven significant research into identifying and preventing these harmful activities. Our project aims to apply machine learning techniques to detect suspicious accounts, with a focus on distinguishing between human-operated accounts and bots. Bots are automated Twitter accounts that spread misinformation or promote content without human oversight. The goal is to classify these accounts, framing the problem as a binary classification task where accounts are labeled as either human or bot based on specific features.

For this project, we are utilizing a dataset[1] from Kaggle that contains information on 37,438 different Twitter user accounts. The dataset consists of 20 features describing various features of each account. These features include account metadata such as creation date, number of followers, number of tweets, verification status and more. Some of the features, like the account description and profile image URL, provide additional context about the user, while others, such as the number of friends and favorites, offer behavioral insights.

We will begin by doing exploratory data analysis: determining label distribution and strategizing how to effectively handle missing data. Since we are solving a binary classification problem and the target variable is imbalanced, where approximately 66% of the data points are marked as human. We will apply stratified data splitting technique to split the training and testing dataset. In terms of the evaluation metric, we will employ the F-1 score to avoid imbalanced predictions. For the model to train with the description column (a text column), we can either drop the description column or we can use Word2Vec to encode the description column. This may create an opportunity to implement NLP techniques on the description column to create richer features.

We will use SVM and Logistic Regression as baseline models, we will seek to improve the model performances with Random Forest Classifier and XGBoost Classifier. When training, we use randomized cross validation to search a wider space of hyperparameters. Finally, we will also run the data through deep learning techniques and compare how the tree based and baseline models compare.

[1] https://www.kaggle.com/code/davidmartngutirrez/bots-accounts-eda/input