Minor Project Documentation

Q1. How was cleaning/EDA performed?

a. Firstly, the columns with language related values were dropped as they require Natural Language Processing (NLP), which is not of focus in our project.

b. Then feature selection was performed to drop the columns which wouldn't play their part in influencing the target feature at all. The columns which had even a slight chance of influencing the target feature were kept.

c. Then the columns which had very poor quality of data were dropped.

d. The target feature, that is *'gender'* had some missing values in it. On performing EDA, it was observed that the rows with missing values matched with the rows which had the value 'no' for the column 'profile_yn'. These rows were 97 in number and hence made a very small fraction of the total Data that we had (20050 rows). So, rather than bringing incorrect data into picture by filling the missing values, we decided to drop these rows. After dropping these rows, the column 'profile_yn' was rendered useless for our machine learning model with all the values filled as 'yes'. Hence, it was dropped.

e. Missing values were filled with **user defined values** for rest of the columns.

f. The data for the columns *'_last_judgment_at'* and *'created'* was simplified (modified) to improve its quality, by keeping only the date and discarding the time.

Q2. What were the independent and dependent features?

The Independent Features after performing EDA and Cleaning were:

['_golden', '_unit_state', '_trusted_judgments', '_last_judgment_at', 'gender:confidence', 'profile_yn:confidence', 'created', 'fav_number', 'gender_gold', 'link_color', 'profile_yn_gold', 'retweet_count', 'sidebar_color', 'tweet_count', 'user_timezone']

The Dependent Feature was the target feature – *'gender'*

Q3. Why and how was selection/engineering/scaling performed?

Feature selection and feature engineering were performed at the same time EDA and Cleaning were being performed to only keep the data which is relevant for training our model.

Label Encoding was used to convert the categorical data in the columns to numerical data as Machine Learning models work only with numerical data.

The dataset had to be balanced before training our model, otherwise it would've introduced bias in our model. **RandomUnderSampler** function of the package **imblearn** was used for the purpose.

The data had to be normalized before training our model. This improves the accuracy of the model and decreases the time consumed. For the purpose, **MinMaxScalar** function of the package **sklearn** was used.

Q4. Which activation function was chosen and why?

'relu' was used for the Input and the Hidden Layers as the dataset we have is non-linear in nature for which, *relu* is a good activation feature. For the Output Layer, **'softmax'** was used as we had a categorical data, at the Output Layer we work with probabilities of outcomes. For working with probabilities, *sofmax* is a good activation function.

Q5. Which optimizer was chosen and why?

'adam' is an adaptive optimizer, i.e., it adjusts the learning rate automatically. It also has advantages above other adaptive optimizers like '*Adadelta'* and '*RMSprop'*. So, there were mainly 3 reasons for using the *adam* optimizer:

- **Adam is the best among the adaptive optimizers** in most of the cases.
- **Good with sparse data**: the adaptive learning rate is perfect for this type of datasets.
- There is no need to focus on the value of learning rate.

Q6. Which neural network was used and why?

The problem statement for this project was classification of gender. **For Classification and Regression problems, Artificial Neural Networks (ANN) are used.** So, in our project, *ANN* was used.