Minor Project – Twitter User Gender Classification

1) How cleaning/EDA was performed

Cleaning was done by dropping the nan values from the dataset and did label encoding on few columns.

EDA was performed using matplotlib.pyplot and seaborn.
By using the text and description column, I visualized the first 20 most words used by male, female and brand. By plotting bar graphs using seaborn, found out that:

- more males are included in the gold standard for the model compared to females and brands.
- females have favorited the most number of tweets than males and brands.
- brands have retweeted the most when compared to male and female.
- brands have posted the most number of tweets than males and females.
- There are almost equal number of male, female and brand.

Then plotted a heatmap on the correlation matrix of the dataset.

2) Your independent and dependent feature

Independent features - _unit_id, _golden, _trusted_judgments, gender:confidence, fav_number, link_color, retweet_count, sidebar_color, tweet_count.

Dependent feature - gender.

3) Why and how selection/engineering/scaling were performed

'_unit_id', '_golden', '_trusted_judgments', 'gender', 'gender:confidence', 'description', 'fav_number', 'link_color', 'retweet_count', 'sidebar_color', 'text', 'tweet_count', 'Tweets', 'Description' are the features selected from the domain knowledge I have.

For scaling I have used Standard Scaler for scaling the data to unit variance so that the artificial neural network can perform better.

4) Which activation function was chosen and why?

Relu is the activation function used because it does not activate all the neurons at the same time and it showed better results compared to other activation functions.

5) Which optimizer was chosen and why?

‘adam’ is the optimizer that is chosen because it gave better result than other optimizers.

6) Which neural network and why? Describe how your neural structuring?

Artificial Neural Network is used because it is a classification problem.
This neural network has one input layer with 9 neurons and a hidden layer with 8 neurons and an output layer with 3 neurons.