Twitter User Gender Classification

Cleaning Dataset:

* Unnecessary columns like, '\_unit\_id', '\_last\_judgment\_at', 'user\_timezone', 'tweet\_coord', 'tweet\_count', 'tweet\_created', 'tweet\_id', 'tweet\_location', 'profileimage', 'created' were dropped.
* Rows with unknown gender and no gender were removed.
* Profile attributes- 'profile\_yn', 'profile\_yn:confidence', 'profile\_yn\_gold' were removed as they were unavailable.
* Rows with confidence of labeling gender<100% were removed.

Manipulating Text Data:

* Text was normalized-(everything was converted to lower case, and URLs, special characters and double spaces were removed.
* The most common words which were meaningless in terms of sentiment (called stopwords) were removed.

Lemmatization:

* Words which expressed same positivity were reduced to their roots using Porter algorithm.
* Two tokenizers, a regular one and one that performs steaming, were used to break down the tweets into individual words.

Exploratory Data Analysis:

The answers to the following questions were explored:

1. What are the most common emotions/words used by Males and Females?

Most common words used by:

1. Females- im, like, get
2. Males- like, get, im
3. Brands- weather, get, updates
4. Which are the most frequently used link colors by Males and Females?

Most frequently used link colors by:

1. Males- 0084B4, 009999, 3B94D9
2. Females- 0084B4, 9266CC, F5ABB5

Visualization:

1. A countplot was created to visualize the amount of each Gender.
2. A bar plot was created to visualize the amount of favorites and retweets.
3. A bar plot was created to visualize colors attributes.

Classification models with Tweet-text only:

* Independent variables- Text, Description.
* Dependent variable- Gender.

Firstly, the categorical labels were converted into numerical ones and it was encoded using LabelEncoder. The data was split into train and test.

* Logistic Regression Model:
* Accuracy obtained- 0.6001931434089812
* Random Forest:
* Accuracy obtained- 0.5676001931434089
* SVM:
* Accuracy obtained- 0.5982617093191694
* Best Accuracy: Logistic Regression Model

Classification models with content of Description added to text:

To increase the accuracy further the ‘description’ was concatenated with ‘text’ and training dataset was re-created.

* Logistic Regression Model:
* Accuracy obtained- 0.6815548044422984
* Random Forest:
* Accuracy obtained- 0.6448575567358764
* SVM:
* Accuracy obtained- 0.6868662481892805
* Best Accuracy: SVM

Ensemble Modelling:

Ensemble technique was used to take advantage of all the three models.

* Accuracy obtained: 0.6895219700627716

Conclusion:

The results show that the **Tweet text** yields a moderate accuracy, but with the content from the **Description**, the performance of classification models significantly improves yielding much better accuracy.

Implementing Ensemble Modelling slightly increases the accuracy.