***ML-MAJOR PROJECT – SEP***

***ML09B9***

THE MAJOR PROJECT FOR OUR GROUP WAS ON THE TOPIC “ TWITTER – GENDER -BIAS ‘ . WE’VE USED SKLEARN LIBRARIES TO PERFORM ENSEMBLE MACHINE LEARNING .

***QUESTIONS AND ANSWERS***

**Q1) WHAT ARE THE MOST COMMON EMOTIONS/WORDS USED BY MALES AND FEMALES ?**

ANS- From the histogram plotted it is clear that the most common word used by males is ‘just’ and by females it is ‘ love’ .

**Q2) WHAT IS THE GENDER CONFIDENCE OF MALES VS FEMALES ?**

ANS – For male it was found to be 0.91(approx.) and female it was found to be 0.93(approx.).

***SUMMARY OF THE PROJECT***

1. The project work was started by **cleaning the data**. Here we removed all the entries with null values and replaced white-spaces, non alphanumeric characters like '' and characters such as ! , @ , # , $. We dropped null rows as well to make our model more accurate.
2. The questions that were asked were correctly implemented along with accuracy prediction. The first question that was asked was **“What are the common emotions/words used by Males and Females”** and the second question that was asked was **“What is the gender confidence of Males vs Females”.**

**Feature selection and engineering** was implemented, where we removed the unnecessary columns namely **'\_golden'**,**'\_unit\_state'**,**'\_trusted\_judgments'**,**'\_last\_judgment\_at'**,**'gender\_gold'**,**'link\_color'**,**'profile\_yn\_gold'**,**'profileimage'**,**'sidebar\_color'**,**'tweet\_coord'**,**'tweet\_location'**,**'user\_timezone'** from the dataset to make the model more accurate and make it easier to perform calculations and predictions over the data.

1. The **independent and dependent variables** were identified for each of the question – for the **first question** **‘gender’** was dependent variable and the **‘words used’** are dependent variable , similarly for **second** one the dependent remained the same that is **‘gender’** and **‘gender confidence’** .
2. Actual implementation of Ensemble Machine Learning started with the use of three classification algorithms, Logistic Regression, KNN Classifier and Support Vector Machine algorithm.
3. The training data and test data were segregated and applied accordingly and results were realized using histogram .
4. The accuracy was predicted using all the three and average was found to be **66** for the prediction of the results of the first question and **38.14** for the prediction of of the results of the second question.
5. The best algorithm for predicting the results of the both the questions was Support Vector Machine algorithm .
6. After comparing the overall accuracy of the three applied models on the two prediction problems, we came to the conclusion that **Support Vector Machine** algorithm will be best suited to perform predictions over the given dataset.