# *GENDER PREDICTION ON TWITTER*

# 

# Introduction

The success of many marketing company such as amazon, flipkart mainly depends on their ability to provide personalized service. Behavior targeting, recommended system is an example of that. According to recent study [1], the advertising systems based on demographic attributes such as gender, age and prior searches are getting more success than elementary advertisement system. For personalized service, age and Gender plays a key role in marketing sector. In many cases we do not have this personalized information available explicitly, so predicting that information from other attribute will play key role in this sector. Nowadays, social media has become immensely popular, the ability to predict personal attribute such as gender, age using their social media information will be useful for marketing purpose and predicting user behavior. In this project, i was provided with twitter dataset with 20,000 instances and 26 features and my purpose is to find relevant pattern in those features to predict user gender.

# Prior Work

The majority of existing model are either focusing on color and text data to predict gender. Following is the brief description of existing papers on this topic:

*[2]* In this paper author has used following features to train a model: tweet count, link color, background color, perc emoji, per\_ioc, user location ,retweet count etc. and trained model on SVM using RBF kernel.

*[3]*In this paper author has used mainly color layout feature. He converted color rbg format to HSV (Hue, Saturation, Values) sorted them and converted back to RBG. Author trained his features on neural network, NB, Decision Tree and used 10 fold cross validation method to validate it.

*[4]*In this paper author has used image, text, and username as features and created separate classifier for each text, username, image and stacked all classifier via boosting

*[5]*In this paper author has assembled text processing and image processing technique to enhance the twitter user gender prediction.

In this project, I followed two approaches. First approach is using color codes such as link color ,sidebar color along with different continuous features such as favorite number ,tweet count ,retweet count for the prediction. Second approach to predict gender is using text feature such as description and text.

# Design and Implementation

* + **Understanding Dataset and Feature selection:** The given dataset contains 20,512 instances and 26 features. Out of 26 features 7 features are continuous and everything else is discrete. For now, I have considered color-hex-code as discrete variable but later in preprocessing stage i have converted color code into RBG format and RBG format is continuous. As i am doing classification on the following dataset so, it won’t be possible to use Unit\_id and tweet\_ID to get some meaningful pattern/information because Unit\_ID and TweetID could be different for each user which makes number of classes equal to number of tuples in prediction.

**Continuous feature**: Further to get into more detail about each feature, I checked spread of the data for each continues feature. From below fig 1.1 we can say that favorite number(fav\_number), retweet count, and tweet count has maximum standard deviation and spread. Feature such as \_trusted\_judgment, profile\_yn\_confidence has almost constant values all over. So later in the project I am focusing on favorite number(fav\_number), retweet count, tweet\_count as it describes data more.

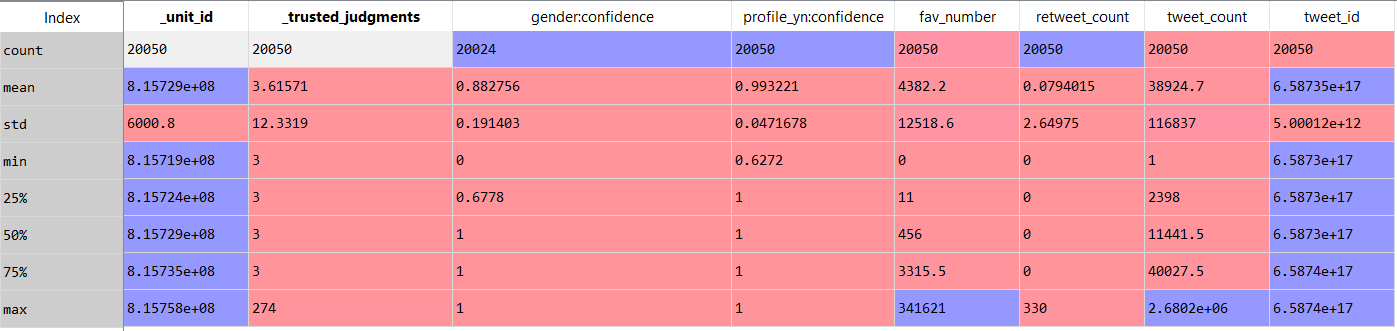


Fig 1.1 Spread of the continuous feature

**Discrete feature**: Form following fig 1.2, i observed that for stated discrete/binary features: ‘unit\_state’,’profile\_yn’,’\_golden’ values are almost constant for all instances. Example feature ‘golden’ has false value all over. So, it’s hard to find specific pattern using this feature, however I am considered these features while creating basic model on raw data, later in project we are not more focused on this feature.

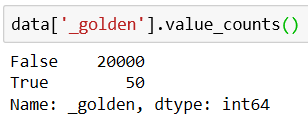
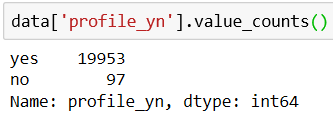
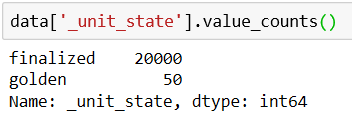


Fig 1.2

**Color Codes and Text Data**: #0084B4 and #C0DEED is Twitter default link and sidebar color. and 50% of user in the given dataset has set link and sidebar color to default. To find some useful pattern I am converting hex to RBG. Tweet and description are also considered as an important features to predict gender.

**Missing values and Distribution of the dataset:** gender\_gold and profile\_yn\_gold has 99.75% missing values And Dependent variable gender has only 1% missing value, moreover it has 5% unknown gender values. And following histogram is the distribution of the dataset based on dependent variable ‘gender’.

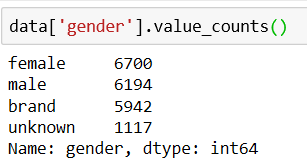
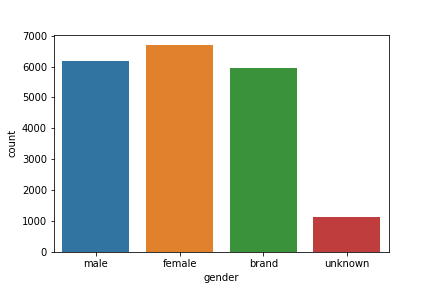
 

Fig1.3

Later in the project, I applied LDA dimensionality reduction technique to the Model 1: with feature Color Code (Link Color and Sidebar Color), fav\_number, retweet count, tweet\_count to check the improvement in performance.

* + **Preprocessing-**

Data Preprocessing is the important step of evaluation of this project. Initially I removed the instances with missing and ‘unknown’ gender. Then I removed the following feature ‘gender\_gold’,’profile\_yn\_gold’,’tweet\_coord’, as 99.75% values are missing for this feature. I also removed ‘unit\_id’ ,’tweetID’ features as its values could be different for each instance which makes number of tuples equal to number of classes. Later I did preprocessing based on model i build, and feature selected for that model. In this project, I have built one model for color codes and features which has more data spread. Also built separate model on text data (text and description)

**Model 1 preprocessing**: Model 1 is built on following features: Color Code (Link Color and Sidebar Color), fav\_number, retweet count, tweet\_count, because those continuous features has highest data spread and other feature values are almost constant as we have seen in above section. Here in preprocessing, initially I have set gender confidence level to >=0.7 and remove instances with non-hex Color Code. And converted color code into RBG pattern. Rest of the features does not have any missing values. Further I divided the features into independent features and dependent feature (‘gender’). And standardize all in-dependent variable.

**Model 2 preprocessing**: Model two is built on following features: text and description. Initially I have set gender confidence level to >=0.7. Further I combined text and description data into one feature and named new feature as TextandDescription. In new TextandDescription feature there are no missing values. After that I used Natural language toolkit to clean the text data. In the cleaning process I tokenized tweet using regular expression while tokenizing regular expression will remove punctuations and unnecessary characters, converted all token into lower text, lemmatize each word using wordnet\_lemmatizer, removed stopword from the tweet and finally converted cleaned text into tweet again. After that I split the data into training and testing set with test set size=25%

* + **Modeling and Analysis:**

**Model 1: Parameter/Feature:** Color Code (Link Color and Sidebar Color), fav\_number, retweet count, tweet\_count. Initially divided the dataset into training and testing set. Further, checked distribution of the data in training and testing set to make sure that distribution is same as dataset distribution. Further in a process, Used naïve Bayes, Decision Tree, Random Forest, KNN machine learning algorithms. Decision Tree and Random Forest got accuracy within range of [54-56] %. Next step i did is dimensionality reduction using LDA and tried those ML algorithm again but there was no improvement in performance. To check the importance of each feature using RamdomForest, I used RandomForest-estimator.feature\_importance command. Here i analyzed that following feature has much more importance than other: Link Color and Sidebar Color, fav\_number and tweet count. These feature has significant role in decision process. So, selected only those important features and did parameter tuning using grid-search-cv on random forest. Below table will show models i tried, with feature and optimized parameter and its accuracy.

**Model 1 Finding:**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Features | Optimized Parameter Value | Current Accuracy |
| KNN | Link\_Color,Sidebar\_color,Fav\_number , tweet\_count | K=30 | 49% |
| Decision Tree | Link\_Color,Sidebar\_color,Fav\_number , tweet\_count | Max\_depth=10 | 52-24% |
| Random Forest | Link\_Color,Sidebar\_color,Fav\_number , tweet\_count | n\_estimators=1000, max\_depth=15 | 56-62% |
| Naïve Bayes | Link\_Color,Sidebar\_color,Fav\_number , tweet\_count | Default | 43-47% |

For random forest I did cross validation on training set to check the average accuracy and it was same as unseen dataset accuracy. Fig 1.4 shows confusion matrix and classification report for random forest.

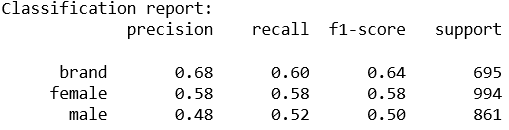
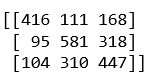


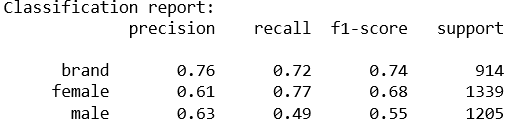
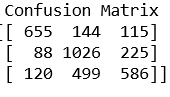
Fig:1.4

**Model 2:** Parameter/Feature**:** text, Description. Here, after preprocessing on text data and Description data I have divided the dataset into training and testing set, used count vectorized function to make tf-IDF on both train data and test data. Ran random forest, Naïve Bayes (MultinomialNB). On text data. Observation is Naïve Bayes (MultinomialNB) is working fine. Then I did parameter tuning using grid-search-cv on Naïve Bayes. Below is chart of accuracies on different algorithm technique.

**Model2 Findings:**

|  |  |  |  |
| --- | --- | --- | --- |
| Method | Features | Optimized Parameter Value | Current Accuracy |
| Naïve Byes (MultinomialNB) | Text and Description | Alpha=1.010 | 66% |
| Decision Tree | Text and Description | Default | 56% |
| Random Forest | Text and Description | Default | 59% |

Below is the confusion matrix for Naïve bayes and its classification report.



**Conclusion:** Overall, using text data on naïve bayes we can achieve good prediction result. However, features: [color code, Favorite number, tweet count] are also powerful predictor because only using those few feature we are getting good prediction results.

**Reference:**

**[1]** Search Engine Watch Journal, Behavioral Targeting and Contextual Advertising,

http://www.searchenginejournal.com/?p=836

**[2]**Language independent Gender Prediction on twitter[NikolaLjubeˇ si´ c, DarjaFiˇse, TomaˇzErjavec]

**[3]**Language independent Gender Classification on twitter[Jalal S. Alowibdi1, 2, Ugo A. Buy1 and Philip Yu1,]

**[4]**cross domain gender prediction on twitter [Mohsen Sayyadiharikande, Giovanni Luca Ciampaglia2]

**[5]**Twitter user Gender inference using combined analysis of image and text processing[Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma]