Gender Prediction Using Twitter Data

John Manohar
Master's in computer science- Future Networked Systems
18314752
manoharj @tcd.ie

Khushboo Goyal
Master's in computer science- Future Networked Systems
1830052
kgoyal@tcd.ie

Shilpa Manda
Master's in Computer Science- Future Networked Systems
18316477
mandas@tcd.ie

1 INTRODUCTION

It is estimated that 2.5 quintillion bytes of data [1] is produced each day in the current scenario. Companies have started to embrace this data driven era by investing in analytical solutions for improving their productivity. There is a paradigm shift from generalized marketing to customer targeted marketing. Social media platforms allow business to target users based on market segments where gender is an important demographic segment. Twitter is a rich source of public data and profile information, but it does not store gender details. Hence, it is a challenge to identify the gender of users with the available information namely tweets and description. Our aim is to identify the most important features that influence the gender of a twitter profile using the public data available on twitter.

2 RELATED WORK

Notable work done in this field in [2] where gender was predicted on twitter using profile picture, screen name and profile description using machine learning algorithms namely decision tree, SVM and neural networks. They concluded that the analysis of a profile picture and name user can determine the gender of a Twitter' user with a minimum expenditure of resources and effectiveness than 82% [2]. Another paper by [3], generated name related features which they assessed using supervised and unsupervised learning algorithms. An unsupervised approach based on Fuzzy c-Means proved to be very suitable for this task, returning the correct gender for about 96% of the users [3].

3 METHODOLOGY

3.1 Data Collection

We captured popular 3600 male and female names each from the USA social security national names dataset of the year 2000 [4]. Next, we used the Twitter API to extract user profile information along with a maximum of 20 timeline tweets each.

3.2 Feature Generation

As a part of feature generation, we processed the text data i.e. tweets and description into numerical features namely number of characters, number of emojis, number of hashtags etc. which resulted in a total of 23 new features. In addition, we created flag variables for features having more than 10 categories.

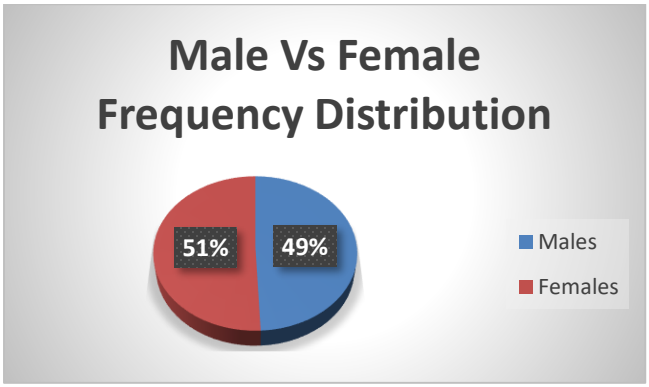

Figure 1: Male Vs Female Frequency Distribution

3.3 Data Preparation

We removed features which produced minimal variation and had maximum null values. Further, we formatted the collected and generated data into a consistent format for model building. We also created dummy variables for categorical features and then summarized the data to user level. Finally, we got 71 predictors and one target variable and the proportion of male and female was nearly equal (figure 1).

3.4 Model Building

Firstly, we divided the dataset into train and test data in 70:30 ratios respectively ensuring equal distribution of male and female. We built three base models using Logistic Regression, Support Vector Machine and Random Forest on the training data and evaluated each across the test data. The performance

evaluation metrics of these models are given in Table 1. In order to understand the most influential feature on the target variable, it is required for the model to exhibit feature weights. But, it is difficult to interpret feature weights of Kernel based SVM models, so we eliminated them for further model building.

Base Model	Accuracy (%)	Specificity	Sensitivity	AUC
Logistic Regression	63.96	65.98	62.25	0.64
RBF SVM	65.21	67.34	63.41	0.653
Sigmoid SVM	55.67	56.36	54.98	0.557
Polynomial SVM	63.23	67.34	63.41	0.653
Random Forest	65.26	65.74	64.77	0.653

Table 1: Base models with performance evaluation values

Base Model	Accuracy (%)	Specificity	Sensitivity	AUC
Logistic Regression	63.56	65.74	61.76	0.636
Random Forest	65.68	66.07	65.28	0.657

Table 2: Final models with performance evaluation values

Next, we applied recursive feature elimination on Logistic Regression and Random Forest algorithms to observe the change in accuracy with reduction in features. The plots for the same are given in Figure 3 and Figure 4. Based on the graph, we decided to use 31 features for Logistic Regression and 42 for Random Forest.

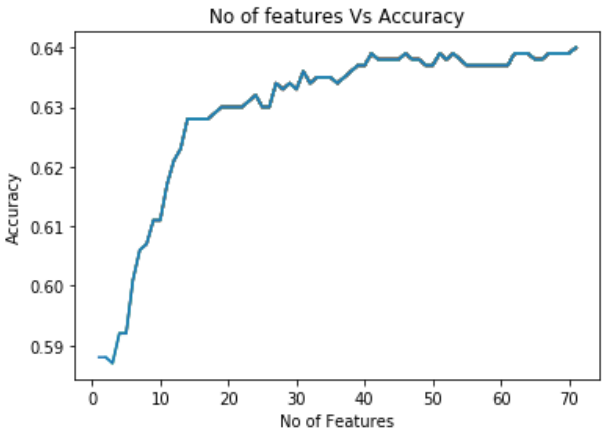

Figure 3: RFE for Logistic Regression

After this, we used the grid search cross validation for Logistic Regression to obtain the optimal value of inverse regularization strength(C) as 1. Similarly, for Random Forest, we got the below hyper parameters:
number of trees: 999
max features per tree: sqrt
evaluation criteria: entropy
minimum sample split: 2

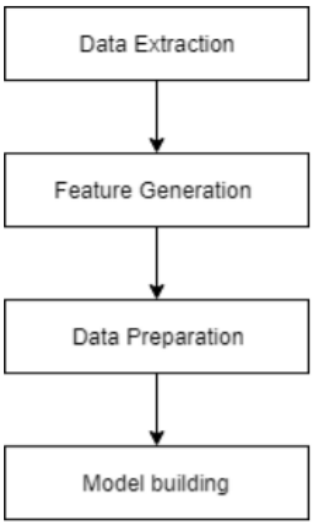

Figure 2: Methodology

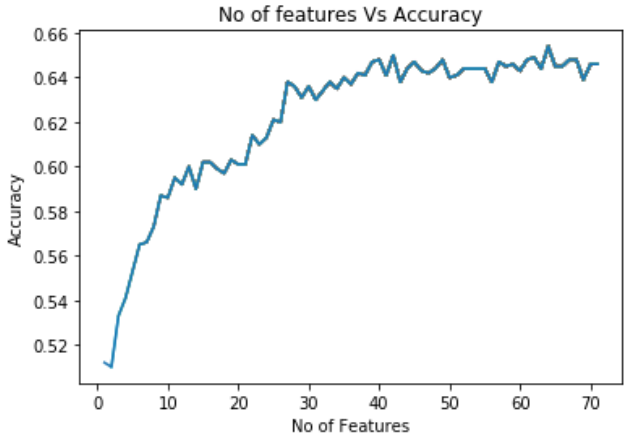

Figure 4: RFE for Random Forest

We built the final model with the above obtained parameters. Table 2 shows the final model with the corresponding performance evaluation metrics. From the results we concluded that Random Forest has yielded the highest accuracy.

4 RESULTS AND DISCUSSION

We used python as the main language for processing the data. Analysis were performed on 8 GB of RAM and Intel core i5 CPU.

From our final Random Forrest model we obtained an accuracy of 66%, sensitivity of 66%, and specificity of 65%. From which we understand that the model is equally capable of identifying between male and female categories. The confusion matrix is shown in figure 6. The model's ability to distinguish male and female was further evaluated using ROCAUC metric(figure7).

Thus, with Random Forest we achieved the highest performance evaluation metrics amongst all. Random Forrest yields the feature importance of each variable in predicting the gender. Figure 5 contains feature importance of the top 15 predictors. From the results obtained we understand that features such as average number of digits in a user's tweet, the number of tweets issued by the user, and the average retweet count are the top 3 predictors for gender. There is very slight difference in the feature importance weights among the top 15 features, from which we infer that the variables are equally important.

From the bar plot it is conspicuous that features created from the individual tweet text data, and the features which are related to the user's popularity (Eg. followers count) and activity (Eg. avg_perc_of_digits) are predictive. We could also conclude that tweet textual data is a very important component for gender identification and exploiting them further could yield better results. Also for better performance we are required to use more advanced features such as profile picture that has been use in paper [2] and features created from profile names[3] with complex algorithms like Multinomial Naive Bayes.

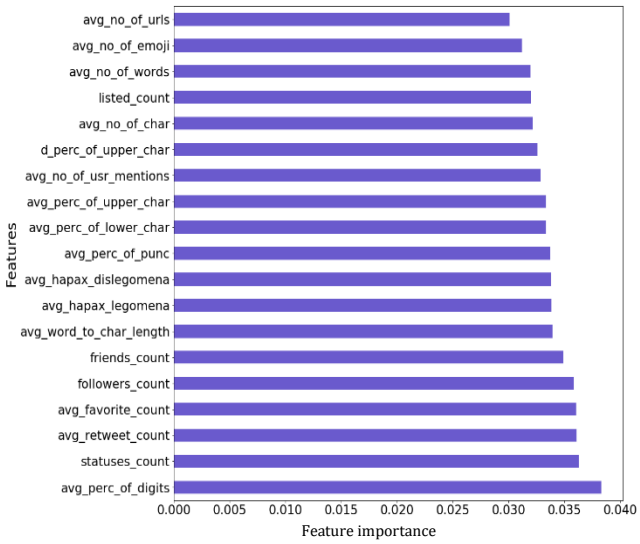

Figure 5: Top predictors in Random Forest.

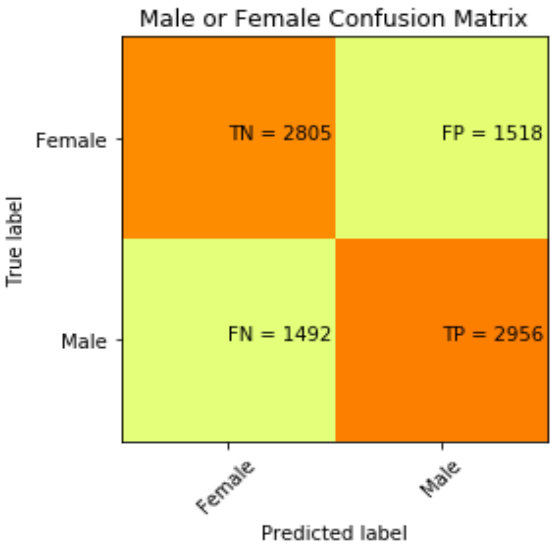

Figure 6: Confusion Matrix for Random forest

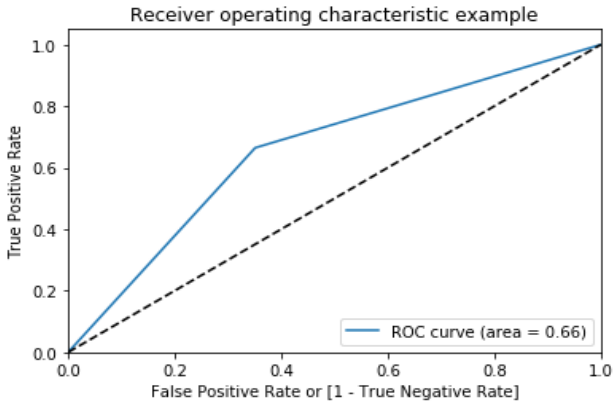

Figure: 7 ROCAUC for Random forest

5 LIMITATION AND OUTLOOK

We built the model based on the popular user names collected from the US Social Security website, hence, this model may not produce expected results with the names of other regions.
From our analysis we understood the importance of textual data of tweets and in future, we would like to extend our model to generate more features from user tweets using Natural Language Processing (NLP) techniques such as n-gram to build model.

REFERENCES

[1] Bernard Marr, https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#34a5bacf60ba

[2] Daniela Fernández; Daniela Moctezuma; Oscar S. Siordia; Features combination for gender recognition on Twitter users, IEEE International autumn meeting on Power, Electronics and Computing, Nov 2016

[3] Marco Vicente ; Fernando Batista ; Joao Paulo Carvalho; Twitter gender classification using user unstructured information; 2015 IEEE Conference on Fuzzy Systems.

[4] https://www.ssa.gov/oact/babynames/limits.html

[5]https://www.researchgate.net/publication/293794120_Gender_Classification_of_Twitter_Data_Based_on_Textual_Meta-Attributes_Extraction