ML1819 Research Assignment 2

Team 32

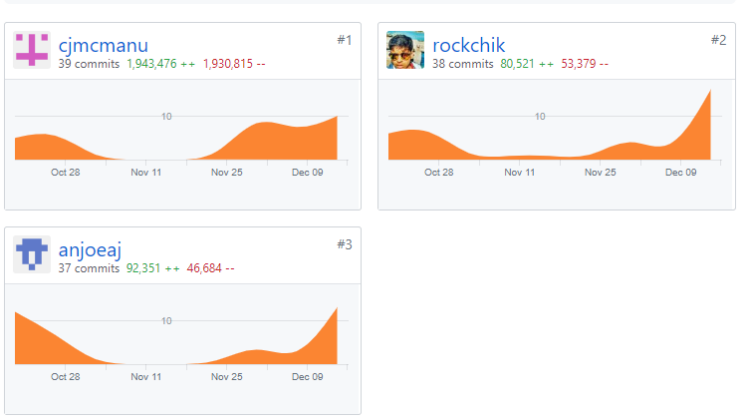
Task:
How well can the gender of Twitter users be predicted? (107)

Team Members:
Cathal J. McManus (11402988)
Avi Garg (18315721)
Anjoe Jacob (18313644)

Contributions:
Cathal: Data Preprocessing, Logistic Regression, Report
Avi: Data Preprocessing, XGBoost, Report
Anjoe: Data Preprocessing, Logistic Regression, Report

Word Count: 1498

Repository:
https://github.com/anjoeaj/ML-1819--task-107--team-32
https://github.com/anjoeaj/ML-1819--task-107--team-32/graphs/contributors



Twitter Gender Classification Based on Categorical and Numerical, and Textual Data

Cathal J. McManus
Trinity College Dublin
Ireland
cjmcmanu@tcd.ie

Avi Garg
Trinity College Dublin
Ireland
gargav@tcd.ie

Anjoe Jacob
Trinity College Dublin
Ireland
anjacob@tcd.ie

1 INTRODUCTION

When registering, a Twitter user only needs to provide a name, handle, and phone number or email. Sharing further information, such as bio, location, or a website is optional. Lastly, the user can share a profile photo, header photo, and choose a theme colour. Interpreting this data is problematic due to the limited information available. To gain further insight into the person behind the Twitter account, additional textual information provided by the user can be analysed. Textual information appears in the form of the profile bio and the content of their tweets.

This paper assesses the potential for this information to be processed and analysed to indicate the gender of a Twitter user. To evaluate this question, information drawn from Twitter profiles and a sample tweet will be processed and analysed using logistic regression and XGBoost.

2 RELATED WORK

Burger et al. [1] predicted Twitter user gender based on limited features from user profiles and sample tweets. Linear SVM, Naïve Bayes, and Balanced Winnow2 were tested with Balanced Winnow2 performing best. Features were tested in heterogeneous sets to simulate varying test circumstances. Their method relied heavily on word analysis against a prepared dictionary of significant terms and on the real name of the user, something not included in our chosen dataset.

To answer this same question, Sayyadiharikandeh et al. [2] obtained 96% accuracy through a stacked classification framework using individual classifiers of tweet text, profile picture, and username that forms the first layer. The username classifier produces a predicted gender, and in the case of text and image, a measure of confidence. The last layer, a meta classifier takes this input and predicts the gender. The use of Computer Vision algorithms was a novel approach in inferring the gender of Twitter users.

Vicente et al. [3] achieved higher accuracy, 97%, using an unstructured dataset of 242,000 twitter users. For pre-processing, features like username and screen-name were compared with a gender-based dictionary. Supervised and unsupervised learning techniques were applied to evaluate the performance. Multinomial Naive Bayes method achieved the highest accuracy.

3 METHODOLOGY

3.1 Acquiring data

Our dataset, created by Figure Eight, was acquired from Kaggle [4, 5]. It contains 20,000 rows of information from Twitter profiles including sample tweets.

3.2 Pre-Processing

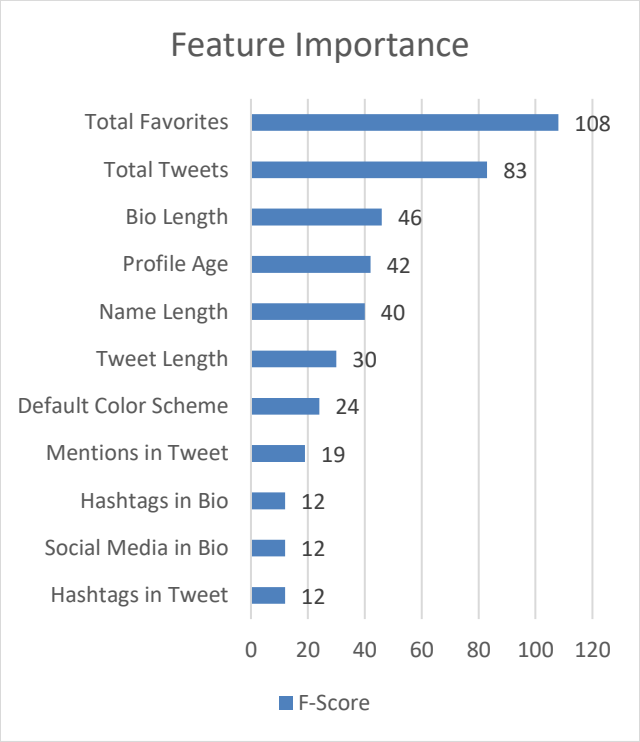
The labelled data contains the gender values ‘Male’, ‘Female’, ‘Brand’, and ‘Unknown’. Only ‘Male’ and ‘Female’ are considered i.e., 65% of the data. The remaining data is composed of about 52% female data and 48% of male data. To normalize numerical features, vector normalization was implemented. Where possible, categorical data was converted into Boolean. For other categorical features, metrics to facilitate comparison were identified. In addition to the textual information, many attributes of the data were formed like number of hashtags, total tweets, length of bio, number of hyperlinks, etc. The importance of these is calculated using the F-score metric and the top five features are considered for this task. The ranking of features based on F-Score can be seen in Fig.1. An outline of our finalized feature list can be seen in Table-1. Initial tests were carried out based on the statistical features only.

Table 1: Features extracted from a Twitter user profile and one sample tweet

Feature	Description
Gender (target variable)	Whether the twitter user is male or female
Profile Age	The number of years since the Twitter profile was created.
Total Favourites	Number of tweets favorited by the user
Total Tweets	Number of tweets sent by the user
Bio Length	The length (in characters) of the user's profile description
Hashtags in Bio	The number of hashtags in the user's profile description
Social Media in Bio	Whether the user has linked another social media platform
Name Length	The length (in characters) of the user's name
Default Color Scheme	Whether the user uses the default color scheme
Tweet Length	The length (in characters) of the sample tweet
Mentions in Tweet	Whether another user is mentioned (using '@')
Hashtags in Tweet	Whether a hashtag was used in the sample tweet

Next, the textual content of the user's profile and a sample tweet were merged and cleaned. Non-alpha characters and words with less than three characters were removed. The hashtag symbol was removed but the hashtag word was kept. Mentions, using the '@' symbol were removed. Lemmetizer [6] was then used to remove inflectional endings so only the base form of each word remained. For example, 'playing' became 'play'. This prevented variations of the same word from complicating results. Bag-of-words was used to count the frequency of words used in the text. Bag-of-words creates a vector representing the most prevalent words in the text within specified limits. Words with a frequency of higher than ninety percent were ignored. Similarly, single occurrence words were ignored [7]. In this case, we analysed the three-thousand most frequent words. Word-clouds of the highest occurring words in tweets, bio, and hashtags are visualized in Fig.6, Fig.7, and Fig.8 respectively.

Figure 1: F Score measurement of features



3.3 Choosing Models

3.3.1 Logistic Regression

Since our prediction is binary, we chose Logistic Regression as the baseline model. The parameters modified were 'C', the inverse of regularization strength, 'penalty', the norm used in the penalization, 'solver', the algorithm used for the optimization problem, and the max number to iterations was modified to allow convergence. Combinations of these were tested to identify the optimal values which are shown in Table-2.

Table 2: Optimum Logistic Regression Parameter Values

Parameters	Values
C	0.2
Penalty	L2
Solver	Liblinear

3.3.2 Extreme Gradient Boost

For comparison, we also employed the state-of-the-art XGBoost algorithm. XGBoost is a gradient boosting decision trees method [8]. The model can be defined by metrics like booster type, step

size, depth of trees, etc. The tuned parameters are shown in Table-3.

Table 3: Optimum XG Boost Feature Values

Parameters	Values
Evaluation Metric	accuracy
Objective	Binary:logistic
Colsample_bytree	0.6
Step Size (eta)	0.01
Tree Depth	8
Child Weight	1
Split Value	1
Subsample	0.5

3.4 Optimization

Initial tests were carried out based on the statistical features of the data to identify the best features for gender classification using F-Score. The importance of a feature is measured by its F-Score. In tree boosting, it is calculated as the number of times a value splits. Fig.1 shows the importance of each feature in the prediction. The features which contributed the least to the prediction were: hashtags in a tweet or bio and the linking of other social media accounts in the user bio. Conversely, the numbers of favourites and tweets, with scores of 108 and 83 respectively, played a significant role in the predictions made by the model. Feature selection was carried out based on these F-Scores. Only top five features were selected. These features were then combined into a new feature dataset which also contained the bag of words vectors for the text data in the user bio and sample tweet.

We tested multiple variations of the newly conceived bag-of-words feature vector. Starting with a maximum word count of 500, we measured the performance, followed by incrementing it to 1,000, 2,000 and 3,000. The threshold was set to 3,000 since this provided the best results while not being too time-consuming. Having established a benchmark using default hyperparameters, optimization was carried out to improve predictions. Parameters, with a range of values, were identified for testing and arranged in a grid. Scikit-Learn's Grid Search iterated over combinations of these values to identify a 'best match'. These values were then used in the final version of each model. For example, the parameters listed in Table-2.

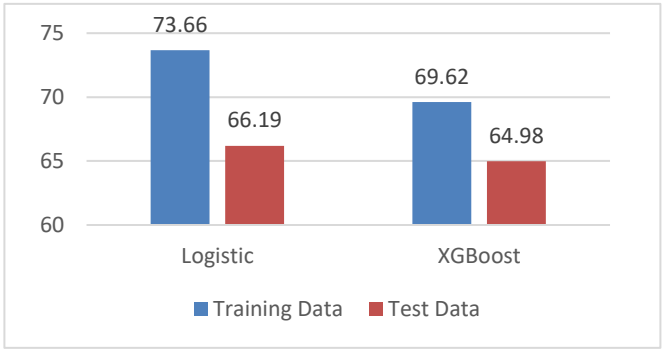
For the tree boosting algorithm, we tuned the hyperparameters using stratified k-fold cross validation coupled with randomized search. For this implementation, we used 3-fold cross-validation with 5 candidates each. The tuned parameters are shown in Table-3.

4 RESULTS AND DISCUSSION

4.1 Results

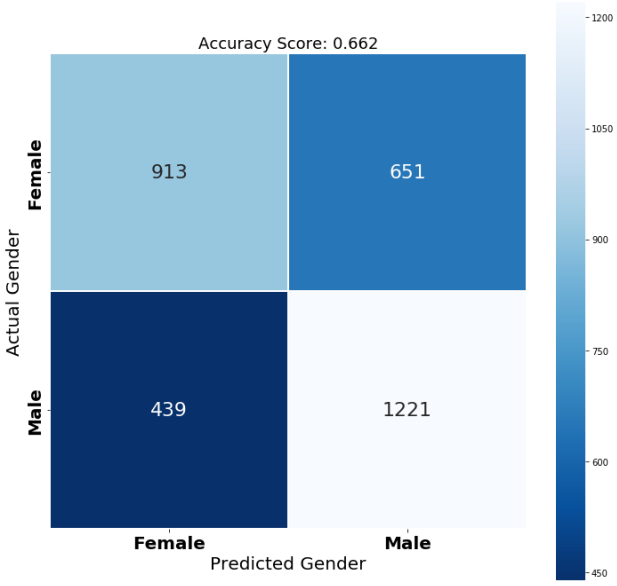
In this study, we used two machine learning methods. Logistic Regression yields good results with an accuracy of 73.66% on training data and 66.19% on test data. Despite being more complex, in this binary classification task, XGBoost output slightly lower accuracy, clocking 69.62% on training data, and 64.98% on the test data.

Figure 2: Model Accuracy Comparison



Logistic Regression yielded a precision of 58.37% in predicting females and 73.55% in males Fig.3. The sensitivity of our model is 67.52% and 65.22% when predicting females and males respectively.

Figure 3: Logistic Regression Confusion Matrix



With area under the curve (AUC) of 0.71 [Fig.4], there is a 71% chance that the model would be able to distinguish between male and female twitter profiles.

Figure 4: Logistic Regression ROC

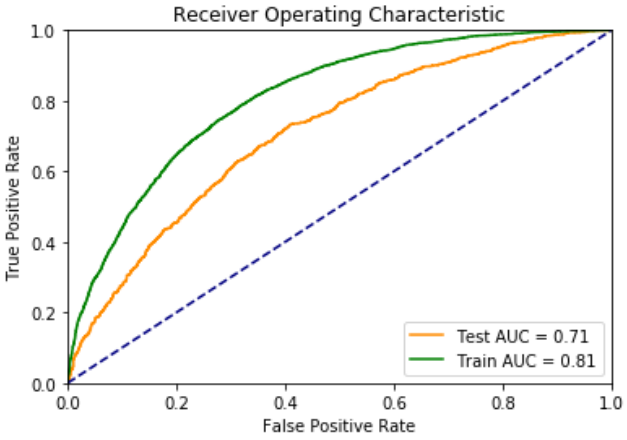
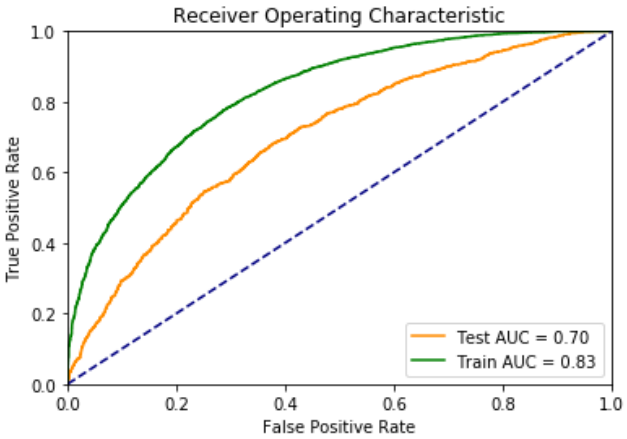


Figure 5: XGBoost ROC



4.2 Interpreting Results

Although the average accuracy achieved using Logistic Regression is 66.2%, through interpretation of the confusion matrix in Fig.3 we see that the difference in precision in predicting females and males is unusual considering the almost equal gender distribution in the test data. More investigation is required to understand why female prediction precision is 58% as opposed to 72% for males. Perhaps, our approach to textual analysis gravitated towards a male prediction. For example, Burger et. al [1] noted that the strongest predictors of a female user were emojis, something we did not account for.

Increasing the depth of the tree in a tree boosting model yields a significant increase in accuracy. Increasing the depth from 6 to 8 yields an accuracy of 69.62% and 64.98%. However, increasing the depth further results in overfitting.

In Section 2, others' performance in answering this question is noted. The models used, and treatment of data yield accuracy levels of 92% [1], 96% [2], 97% [3]. The underlying data is essentially the same, however, we use a much smaller dataset, deviating in how we manipulate data, and in models used. We convert text into numerically quantifiable data to facilitate comparison. We couple these features with a textual analysis of the tweet and user bio content. With analysis of a larger dataset and the analysis of symbols, particularly emojis, a higher overall accuracy could be reached.

5 LIMITATIONS AND OUTLOOK

Our dataset is limited to 20,000 users and preprocessing of this data reduced it to 12,000. Numerical analysis of the information available through a single tweet and a user bio led to accuracy which was not significantly higher than 50% when predicting gender. The accuracy was greatly improved through the addition of textual analysis. With a more sophisticated model for text analysis that supported emoji analysis, even higher accuracy could be achieved. Alternate future directions include computer vision technology for gender classification based on profile pictures as proposed by Sayyadiharikandeh et al. [2].

APPENDIX

Figure 6: Tweet Word cloud

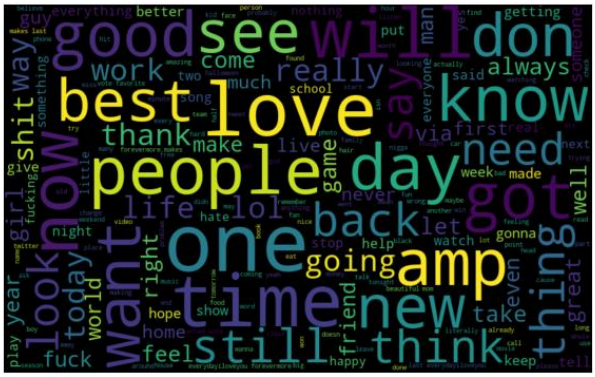


Figure 7: User Bio Word cloud

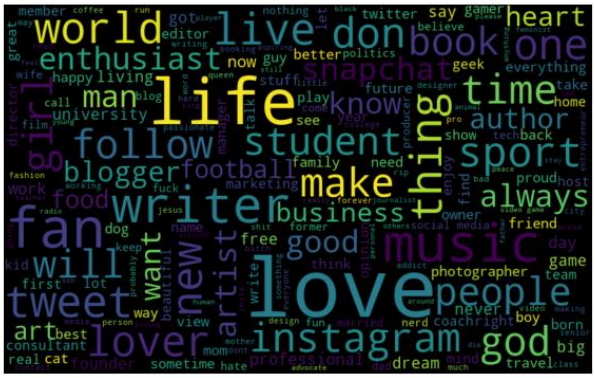


Figure 8: Hashtags Word cloud

REFERENCES

[1] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating gender on Twitter,” in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Edinburgh, United Kingdom, 2011, pp. 1301-1309.

[2] M. Sayyadiharikandeh, G. L. Ciampaglia, and A. Flammini, “Cross-domain Gender Detection in Twitter,” *Proceedings of the Workshop on Computational Approaches to Social Modeling*, vol. ChASM 2016, Nov, 2016.

[3] M. Vicente, F. Batista, and J. P. Carvalho, "Twitter gender classification using user unstructured information." pp. 1-7.

[4] K. inc. "Twitter User Gender Classification," 14/10, 2018; https://www.kaggle.com/crowdflower/twitter-user-gender-classification.

[5] F. E. Inc. "Data For Everyone," 28/10, 2018; https://www.figure-eight.com/data-for-everyone/.

[6] H. Jabeen. "Stemming and Lemmatization in Python," 2018; https://www.datacamp.com/community/tutorials/stemming-lemmatization-python.

[7] P. Joshi. "Comprehensive Hands on Guide to Twitter Sentiment Analysis with dataset and code," 30/11, 2018; https://www.analyticsvidhya.com/blog/2018/07/hands-on-sentiment-analysis-dataset-python/.

[8] T. Chen, and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA, 2016, pp. 785-794.