ML1819 Research Assignment 1

Team 32

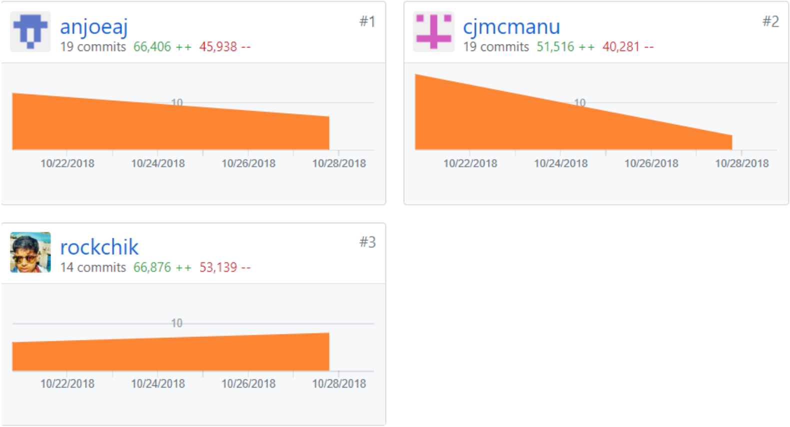
Task:
How well can the gender of Twitter users be predicted? (107)

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

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

Word Count: 999

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



anjoeaj #1
19 commits 66,406 ++ 45,938 --

cjmcmanu #2
19 commits 51,516 ++ 40,281 --

rockchik #3
14 commits 66,876 ++ 53,139 --

Twitter Gender Classification Based on Categorical and Numerical 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 to share is optional. Lastly, the user can share a profile photo, header photo, and choose a theme color. Interpreting this data is problematic due to the limited information available.

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

Vicente et al. [2] 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. Interestingly, an unsupervised learning algorithm, fuzzy c-Means was 96% accurate. However, its accuracy can only be improved by large training datasets.

3 METHODOLOGY

3.1 Acquiring data

Our dataset, created by Figure Eight, was acquired from Kaggle [3, 4]. It contains 20,000 rows of information from Twitter profiles including sample tweets.

3.2 Pre-Processing

The labeled data contains the gender values 'Male', 'Female', 'Brand', and 'Unknown'. Only 'Male' and 'Female' are considered, 65% of the data in total. To normalize numerical features, we implement vector normalization. Where possible, categorical data is converted into Boolean. Some non-uniform values are discarded, while others, like profile image, are set aside as potential future features. For other categorical features, we identify metrics to facilitate comparison. For profile descriptions, we count the number of words, number of hashtags, references to social media, and hyperlinks. Our finalized list is: account creation date, favorites, tweets, description length, hashtags in description, social media references, username length, default color scheme use, tweet length, mentions ('@' in a tweet), tweet hashtags, hyperlinks used.

3.3 Choosing Models

3.3.1 Logistic Regression

Since our prediction is binary, we chose Logistic Regression as the baseline model. Multiple parameters were tweaked to get an optimal prediction. We tested combinations of C - the inverse of regularization strength, penalty, and polynomial features.

3.3.2 Support Vector Machine

SVM was chosen due to its efficiency in classifying high-dimensional data and suitability for binary labeled data. Furthermore, the use of regularization parameters could inhibit overfitting while the kernels could assist with non-linear classification [5].

3.3.3 Extreme Gradient Boost

We achieved good accuracy using the state-of-the-art XGBoost method, a tree boosting method [6]. The model can be defined by metrics like booster type, step size, depth of trees, etc. The model is defined for binary classification and the parameters are tuned for best results. The tuned parameters are shown in Table 1.

Table 1: Optimum XG Boost Feature Values

Parameters	Values
Booster Type	B tree
Evaluation Metric	logloss
Step Size (eta)	0.1
Tree Depth	6
Child Weight	10
Split Value	0.7

Using tuned parameters, the model was 62.46% accurate, slightly better than previous models.

3.4 Optimization

Initial tests were carried out using the default model parameters. Having established a benchmark, optimization was carried out to improve predictions. Optimization was automated using Scikit-Learn's 'Grid Search' function. Parameters, with a range of values, were identified for testing and arranged in a grid. 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. Table 2 shows this process. For SVM, the kernel was taken as a constant with the C and gamma values varying.

Table 2: GridSearch Optimization parameters

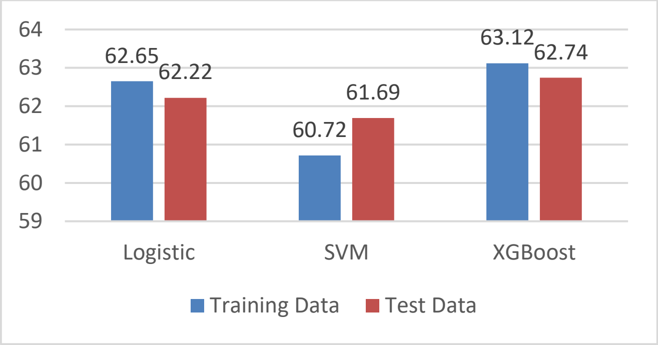
Kernel	Linear	Sigmoid	Gaussian
C	1000	10	10
Gamma	Na	1000	0.001
Training	60.72%	68.84%	57.26%
Test	61.69%	59.46%	56.73%

4 RESULTS AND DISCUSSION

4.1 Results

In this study, we used various machine learning methods. Logistic Regression yields good results with an accuracy of 62.65% on training data and 62.22% on test data. Surprisingly, using the more complex SVM outputs slightly less accuracy, 60.72%, on test data and 61.69% on training data. The highest accuracy is obtained by XGBoost, scoring 63.12% on training data and 62.14% on test data.

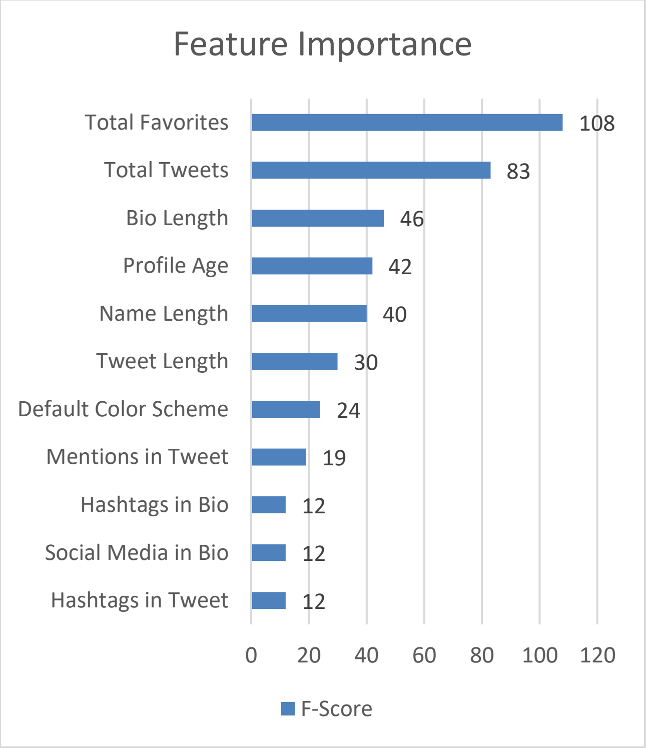
Figure 1: Model Accuracy Comparison



4.2 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.2 shows the Feature Importance of each feature in the prediction. With scores of 12, the number of hashtags in a bio or tweet and the inclusion of a link to other social media, were the least important features. Conversely, the number of favorites and the number of tweets, with scores of 108 and 83 respectively, were effective indicators of gender.

Figure 2: F Score measurement of features



4.3 Interpreting Results

In Section 2, others' performance in answering this question is noted. The models used, and treatment of data yield accuracy levels of 92%[1] and 97%[2]. The underlying data is essentially the same, however, we use a smaller dataset, deviating in how we manipulate data, and in models used. We convert text into numerically quantifiable data to facilitate comparison. Using our models, we assess the potential to predict the gender of a Twitter user. While our models were less accurate, they did illustrate the potential of numerical data in predicting gender, namely, the number of favorites and tweets.

5 LIMITATIONS AND OUTLOOK

Our dataset is limited to 20,000 users and preprocessing of this data reduced it to 12,000. Many features are eliminated based on their relevance to our model. More time is needed to analyze and manipulate the dataset so new insights can be found. In this study, we have taken data attributes like name length and profile length into consideration, but we have excluded other valuable attributes like words in the description and tweet. Future directions include implementations of sentiment analysis using word dictionaries and computer vision technology for gender classification based on profile pictures as proposed by Sayyadiharikandeh et al. [7].

APPENDIX

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

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

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

[5] Q. Team. "Support Vector Machines: A Guide for Beginners," 24/10, 2018; https://www.quantstart.com/articles/Support-Vector-Machines-A-Guide-for-Beginners.

[6] 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.

[7] 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.