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, SVM, Report

Avi: Data Preprocessing, XGBoost, Report

Anjoe: Data Preprocessing, Logistic Regression, Report

Word Count: \_\_\_

Repository:

<https://github.com/anjoeaj/ML-1819--task-107--team-32>

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

*picture of contributions*

**Twitter Gender Classification Based on Categorical and Numerical Data** \_\_\_\_\_\_\_words/sentiment analysis?

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
|  |

# 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. To gain further insight about 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.

# Related Work

Burger et al. [[1](#_ENREF_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](#_ENREF_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.

# mETHODOLOGY

## Acquiring data

Our dataset, created by Figure Eight, was acquired from Kaggle [[3](#_ENREF_3), [4](#_ENREF_4)]. It contains 20,000 rows of information from Twitter profiles including sample tweets.

## 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. For other categorical features, we identify metrics to facilitate comparison. For profile descriptions, we count the number of words, number of hashtags, and links to other social media. An outline of our finalized feature list can be seen in [Table. 1](#feature_list). Initial tests were carried out based on the statistical features only. After these, the textual content of the user’s profile and a sample tweet was analysed. The two text groups were merged and then cleaned. Non-alpha characters and words with less than three characters were removed. The hashtag symbol was removed but the hashtag word was kept. Meanwhile, mentions of other profiles using the ‘@’ symbol were removed. Lemmetizer 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. Next, Bag of Words was used to count the frequency of words used in the text. Words which appeared with frequency of higher than ninety percent were ignored. Similarly, any word which only occurred once was ignored [[5](#_ENREF_5)]. In this case, we analysed the three-thousand most frequent words.

**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 age of the Twitter profile |
| 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 |

## Choosing Models

### **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, and ‘solver’, the algorithm used for the optimization problem. Combinations of these were tested to identify the optimal values.

### **Extreme Gradient Boost**

We achieved good accuracy using the state-of-the-art XGBoost method,  a tree boosting method [[6](#_ENREF_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 2](#Table1).

**Table 2:** 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.

## Optimization

Initial tests were carried out on the stastical features listed in [Table. 1.](#feature_list) The purpose of these tests were to identify the most useful statistical features for gender classification. To identify the most useful features, their F-Score was calculated. 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](#Figure2) shows the Feature Importance of each feature in the prediction. The features which contributed the least to the model’s prediction were: the inclusion of hashtags in a tweet or user bio, and the linking of other social media accounts in the user bio. Conversely, the number of favorites and the number of tweets, with scores of 108 and 83 respectively, played a significant role in the predictions made by the model. In order to streamline the model, feature selection was carried out based on these F-Scores. A threshold of forty was set for feature selection. Therefore, only 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.

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](#Table2) shows this process. For SVM, the kernel was taken as a constant with the C and gamma values varying.

**Figure 2: F Score measurement of features**

**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% |

# RESULTS AND DISCUSSION

## 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

## Interpreting Results

In [Section 2](#_Related_Work), others’ performance in answering this question is noted. The models used, and treatment of data yield accuracy levels of 92%[[1](#_ENREF_1)] and 97%[[2](#_ENREF_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.

# 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](#_ENREF_7)].

**APPENDIX**

1 Introduction

2 Related Work

3 Methodology

3.1 Acquiring Data

3.2 Processing Data

3.3 Choosing models

3.3.1 Logistic Regression

3.3.2 Support Vector Machines

3.3.3 Extreme Gradient Boosting

3.4 Optimization

4 Results and Discussion

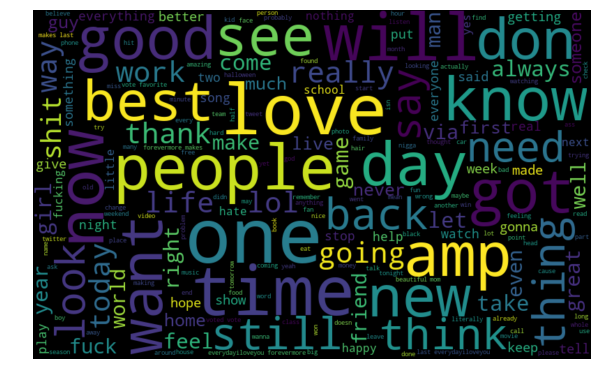
4.1 Results

4.1 F-score values

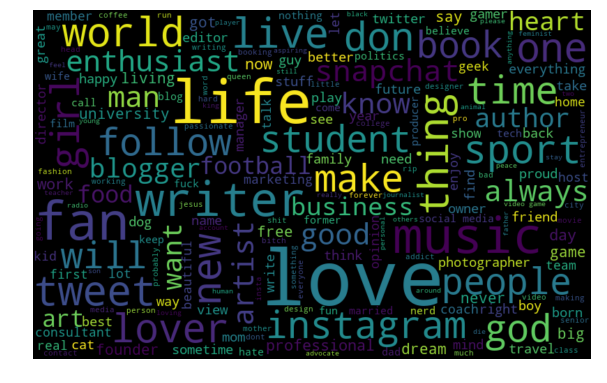
4.1 Interpreting results (Conclusion)

5 Limitations and Outlook

**Figure 3: Tweet Word cloud**



**Figure 4: User Bio Word cloud**



**Figure 5: 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. 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] 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/>.

[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.