ML1819 Research Assignment 2

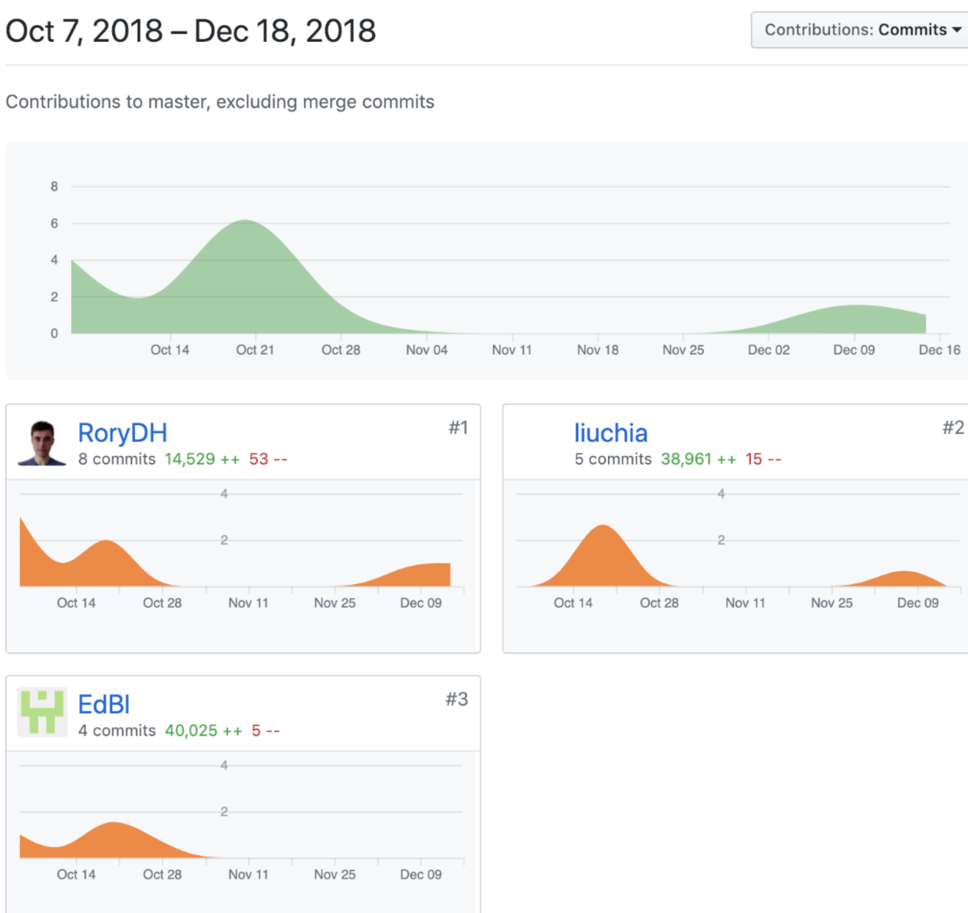
Team 47

107. How well can gender be predicted, and which features are particularly predictive?

15326006 Chia (Jason) Lau	Research, classifier design, writing
15325303 Rory Hughes	Research, data download and enrichment, writing
15323612 Edvardas Blazevicius	Research, data collection, writing, report editing

Word count: 1296

Source code repository:
https://github.com/RoryDH/ML1819--task-107--team-47



Predicting a Twitter User's Gender Using Machine Learning

Chia J. Lau
lauch@tcd.ie

Rory Hughes
hughesr8@tcd.ie

Edvardas Blazevicius
blazevie@tcd.ie

1 INTRODUCTION

The vast amounts of user generated data available on social networks is multiplying year on year at an ever increasing rate. With many social networks being over a decade old, this makes them bottomless wells of user generated data.

Twitter does not ask its users for their gender. Armed with megabytes of other data on each member, it would be interesting to see how well their gender could be predicted. On top of that it would be interesting to investigate what subsets of that data are the best predictors of their gender.

1 INTRODUCTION

The vast amounts of user generated data available on social networks is multiplying year on year at an ever increasing rate. With many social networks being over a decade old, this makes them bottomless wells of user generated data.

Twitter does not ask its users for their gender. Armed with megabytes of other data on each member, it would be interesting to see how well their gender could be predicted. On top of that it would be interesting to investigate what subsets of that data are the best predictors of their gender.

2 RELATED WORK

A lot of work has already gone into trying to classify a user's gender based on their Twitter data.
You et al. [1] attempted predicting a social media user's gender using the images they post.
Zamal et al. [2] and Burger et al. [3] tried predicting a Twitter user's gender by applying an n-gram feature model to the user's data such as their real name, username, bio and tweets.
Tenuto of Figure Eight [4] tried predicting whether a Twitter user is male, female or a brand by analyzing key words and emoji's in their tweets.
Liu and Ruths [5] attempted predicting a Twitter user's gender using their first name only.
Rao et al. [6] tried predicting a Twitter user's gender by analyzing the sentiment of their tweets.
Alowidbu et al. [7] attempted predicting a Twitter user's gender by analyzing their use of colors as well as the phonemes found in their names.

3 METHODOLOGY

Dataset

The data we used originated from two datasets. The first was a CSV[9] of 14,000 Twitter users with their genders guessed by humans along with their confidence. The second dataset was a list of 40,000 Twitter usernames[10]. There was no gender data associated with the names so we used a gender guesser module[8] which given a first name outputs a gender and its confidence in its prediction.

Using the Twitter API, we collected more data for each user. Some users were removed from our dataset as their profile was not public or their account had been deleted.

Only users who genders were certain according to their source dataset was used. For the first dataset, this means the community vote agreed unanimously. For the second, the gender guesser module returned a high confidence. Users who did not have first names and last names were also removed.

We split the dataset into two bins, one for female and one for male and then drew an equal amount from each. This resulted in a final dataset of 6118 males and 6118 females.

Features

The Twitter API gave us access to the following which we made use of as features:

- First Name
- Last Name
- Username
- Description
- Follower Count
- Friend Count
- Favourites Count
- Custom Profile Colours (Text, Background, Link, Sidebar Border, Sidebar Fill)

The colours were given as an RGB hex string which we converted to HSV format. The motivation for this was that Hue, Saturation and Value captures the way humans perceive colour much better than RGB.

The first three and last three letters of the names, the username length and description length were the features based on the string data given. Overall, this created 38 features.

Feature scaling was performed via standardization.

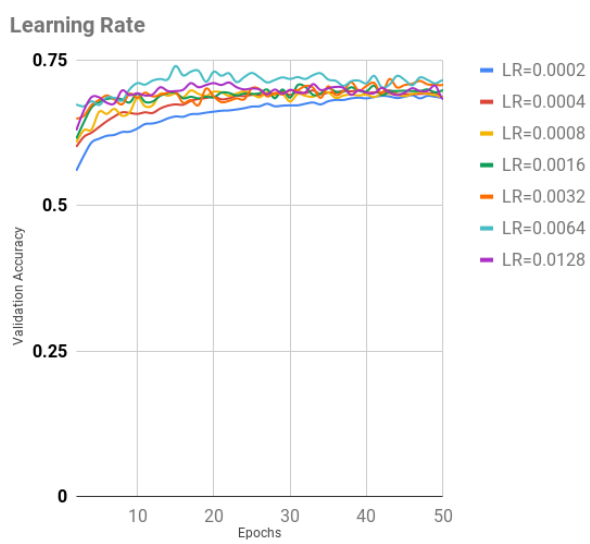
Algorithm

The machine-learning algorithm that we used was neural networks. The library that was used was Tensorflow. Neural networks were chosen over linear regression or Support Vector Machines as it would handle non linearly separable data much better.

The following hyperparameters were used

Number of Hidden Layers	1
Activation Function	Relu
Optimization Algorithm	AdamOptimizer
Learning Rate	0.0064
Epochs	20

The choice of learning rate and epochs were determined experimentally by parameter sweeping :



Evaluation

Holdout Evaluation is used with 80% Training Set and 20% Test Set. The Training Set is split 80/20 again for the Validation Set.

The dataset contains an equal number of male users and female users which allows us to use accuracy as a reasonable metric. A baseline accuracy is 0.5 (with random guessing). Other attempts at Twitter gender classification have achieved accuracies hovering between 0.72 (Rao et al. [6]) and 0.92 (Burger et al. [3]).

Feature Importance will be measured by training the model on the dataset without the feature to be measured and comparing the accuracy of this new model (Anew) with that of the model trained with all features (Abase). More precisely,

$$\frac{A_{new} - A_{base}}{A_{base}}$$

The more negative this value, the more important the feature is as its absence has worsened the accuracy of the model.

4 RESULTS AND DISCUSSION

The accuracy of the model using all features was 70.65%.

The multiplicative percentage change on absence for the most and least important features were as follows:

Rank	Feature	% Change
1	Last Letter of First Name	-10.52%
2	Third Last Letter of Username	-4.51%
3	First Letter of Username	-3.93%
4	Second Last Letter of First Name	-3.87%
5	Third Last Letter of First Name	-3.64%
6	Follower Count	-3.12%
7	Second Last Letter of Username	-3.06%
8	Third Last Letter of Last Name	-2.94%
9	Profile Link Color Brightness	-2.89%
10	Profile Text Color Saturation	-2.77%

Rank	Feature	% Change
33	Profile Link Color Saturation	0.05%
34	Profile Text Color Hue	0.23%
35	Second Letter of Last Name	0.23%
36	Second Last Letter of Last Name	0.28%
37	Description Length	2.37%

By removing the worst features, we can predict the gender of a Twitter user with 74.16% accuracy.

5 LIMITATIONS AND OUTLOOK

For the first part of the project we were using a small data set of 40,000 users - the next smallest dataset that we saw in the referenced research papers had 150,000 users. We would like to increase our dataset by at least a factor of three.

We experimented with using phonemes (the sounds of the syllables of the word) as a representation of names instead of letters but our results were unfortunately poorer however we are confident that further experimentation with phonemes may succeed.

We also looked into text analysis of each users' tweet content however were blocked early on by the restrictive twitter API limitations. Although we found datasets online with tweet text, we instead wanted tweets for each of the the users whom had gender labels in our original dataset.

Furthermore, we also considered analyzing image content. Both user profile pictures and images attached to to tweets could be analysed for color, objects and possibly facial features but this was outside of scope for this project.

REFERENCES

[1] Q. You, S. Bhatia, T. Sun and J. Luo, "The Eyes of the Beholder: Gender Prediction Using Images Posted in Online Social Networks," 2014 IEEE International Conference on Data Mining Workshop, Shenzhen, 2014, pp. 1026-1030.
DOI: 10.1109/ICDMW.2014.93

[2] Al Zamal, Faiyaz & Liu, Wendy & Ruths, Derek. (2012). Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors.

[3] L. Burger, J. Henderson, G. Kim, and G. Zarrella, “Discriminating
gender on Twitter”. In Proceedings of EMNLP’11
1301–1309, 2011.

[4] Justin Tenuto. Figure Eight. *Using machine learning to predict gender.* https://www.figure-eight.com/using-machine-learning-to-predict-gender [Accessed October 2018]

[5] Liu, Wendy and Derek Ruths. “What's in a Name? Using First Names as Features for Gender Inference in Twitter.” AAAI Spring Symposium: Analyzing Microtext (2013).

[6] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In Proceedings of the 2nd international workshop on Search and mining user-generated contents (SMUC '10). ACM, New York, NY, USA, 37-44. DOI: https://doi.org/10.1145/1871985.1871993

[7] Alowibdi, Jalal & Buy, Ugo & Yu, Philip. (2013). Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter. 1. 365-369. 10.1109/ICMLA.2013.74.

[8] gender-guesser, pip module, https://pypi.org/project/gender-guesser/

[9] Dataset with users with pre-classified genders, https://www.kaggle.com/crowdflower/twitter-user-gender-classification

[10] Large dataset containing Twitter usernames, https://www.kaggle.com/hwassner/TwitterFriends