An exploration of gender indicative feature data on Twitter

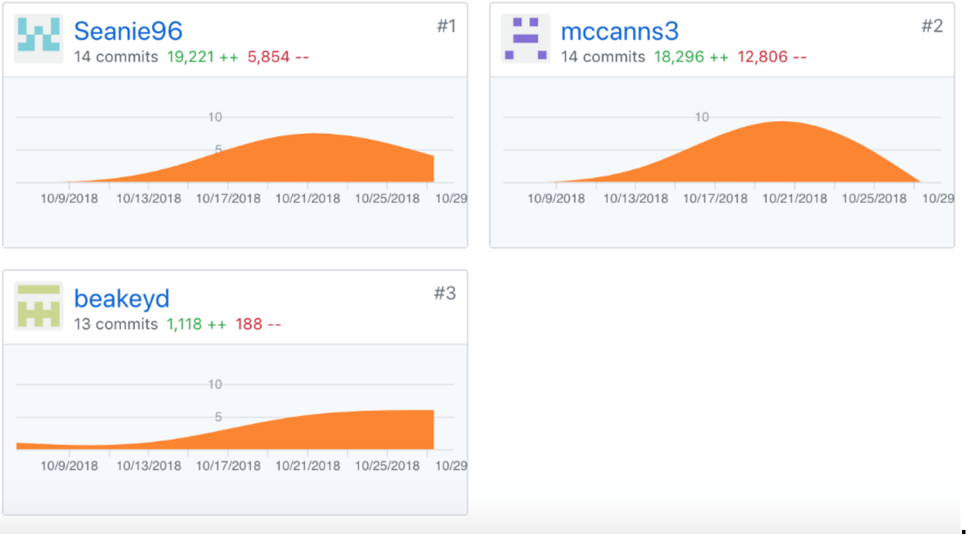
ML1819 Research Assignment 1

Team 41 - Task 107 - How well can the gender of Twitter users be predicted?

Word count - 999

Github Source Code

Github Contributors



Seanie96 #1
14 commits 19,221 ++ 5,854 --

mccanns3 #2
14 commits 18,296 ++ 12,806 --

beakeyd #3
13 commits 1,118 ++ 188 --

Samuel McCann
Trinity College Dublin
Ireland
#15318105
mccanns3@tcd.ie

Sean McDonagh
Trinity College Dublin
Ireland
#15319517
semcdona@tcd.ie

David Beakey
Trinity College Dublin
Ireland
#15335531
dbeakey@tcd.ie

We worked on the project everyday over the course of a week together. Everything that we have submitted, in terms of code, was worked on together.

1 INTRODUCTION

Twitter has become a key platform for companies and public figures to engage with their customer base. Understanding your demographic of users has become a priority of businesses to better interact with them.

Public Profile information is often faked, and twitter does not store much information for each user. This makes it hard for public companies to easily scrape data on their userbase.

Research Question: To explore what data properties on a twitter profile may be indicative of their gender, to allow public bodies to better understand their demographic.

2 RELATED WORK

John D.Burger et al (Burger, Henderson, Kim, & Zarrella, 2011) managed to classify a user's gender with an accuracy rate of 92% when using a user's tweets, their screen name and their description. The user's tweets were concatenated together for a feature. They used a Winnow2 linear classifier to achieve this rate. For ground truth they followed the blogs of users and obtained their gender there.

Clay Fink (Fink, Kopecky, & Morawski, 2012) also examined this problem in detail. They determined gender by searching twitter users Facebook pages. Features were generated for unigrams, hashtags and psychometric properties. Using an 80/20 training to test ratio, they achieved an accuracy rate of 80.60 with a combined feature vector of unigrams, linguistic inquiry word count, and hashtags and using a SVMLight linear classifier.

3 METHODOLOGY

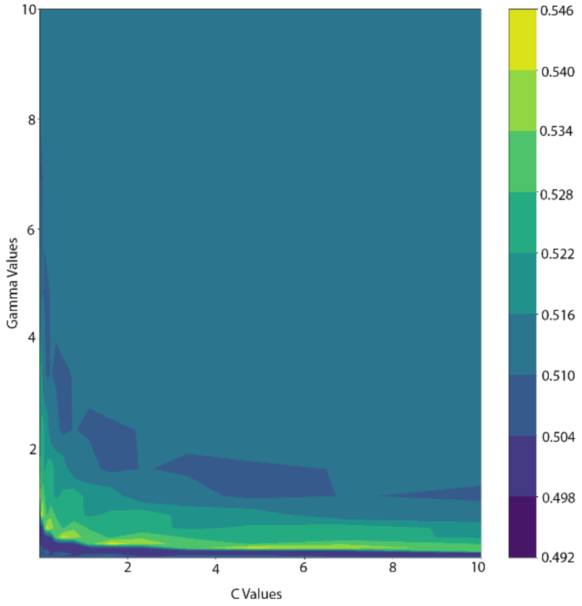
The dataset was obtained through an online research paper on gender classification (Liu & Ruths, 2013). There existed approximately 12,000 user_id's in the dataset. Associated with each user_id is a gender. In order to get relevant data on the users, a script was written to scrape the data from the Twitter API. Data that was thought to have been mediocre to good predictors of gender was then scraped for each user. The dataset was pruned to obtain a balanced ratio of men to females. After the steps stated above, the dataset contained 6000 users.

The problem of differentiating a user into the set of males or females, based off their user account data, is that of a classification problem. There were two algorithms that we could have chosen to use in order to classify users; logistic regression and SVM. SVM was the algorithm that was chosen in the end, due to our belief that for many of the features that we were going to test, there would not exist a linear separation in the data, but rather it was non-linear.

When inputting multiple features into an SVM, we used a series of pipelines to combine them according to an article online (dbaghern, 2017). It was also necessary to convert all the text features to a vector of token counts, using a Count Vectorizer. A tutorial was followed to achieve this (Usman, 2018). The Gender property in our json document was transformed so that 'Male' equals 0, and 'Female' equals 1.

To optimize our C and Gamma parameters for our SVC(Support Vector Classification) SVM's models, we plotted the accuracy of the classifier with a range of the paramaters. The below image shows the resulting plot for the tweet feature based SVM. For NuSVC classifiers, we also classified the best nu value through a grid search process.

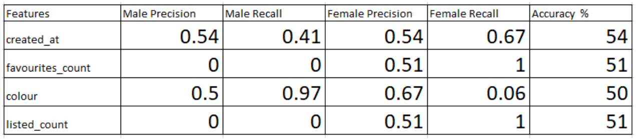
Figure 1 – Gamma C Values



We split the data 90% to training data, 10% to test data. The data proved to not be easily separable, thus a higher training size set resulted in a more accurate model in our experimentation.

4 RESULTS AND DISCUSSION

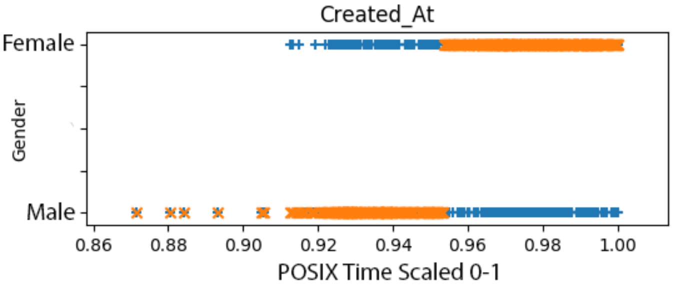
Table 1 - Numerical Features



Features	Male Precision	Male Recall	Female Precision	Female Recall	Accuracy %
created_at	0.54	0.41	0.54	0.67	54
favourites_count	0	0	0.51	1	51
colour	0.5	0.97	0.67	0.06	50
listed_count	0	0	0.51	1	51

You can see from the overall accuracy in Table 1, as well as the Precision and Recall numbers, that the prediction rates for the models which used individual numerical features alone are poor. You can see for example, the model Figure 3 which has favourites_count as its singular feature always predicts Female. This predication is no better than random chance. On their own, these individual features are not good indicators of gender.

Figure 2 - Twitter Account Creation Date



A plot of the Created_At data Figure 2 shows where the classifier is defining the split in the data. Blue crosses on the graph represent the test data used in the model, while orange X's represents the predictions on the test data. X Values are POSIX, time scaled between 0 and 1. Accounts created before 2017 are predicted to be Male, while accounts created after are predicted to be female. This confirms that this model is doing more than just predicting everything to be all Female like in the favourites_count Figure 3 singular feature model, which is below.

Figure 3 - Number of Favourites

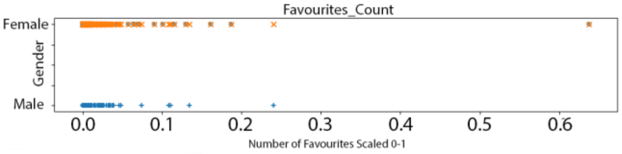
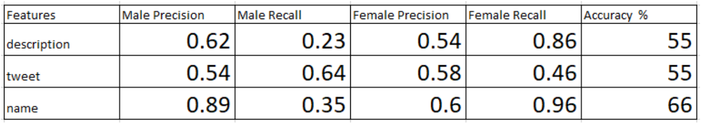


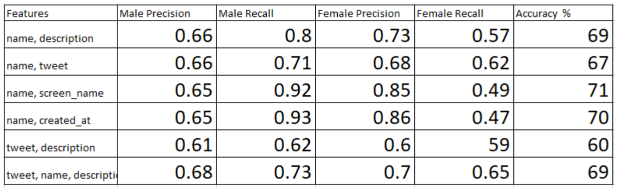
Table 2 - Text Features



Features	Male Precision	Male Recall	Female Precision	Female Recall	Accuracy %
description	0.62	0.23	0.54	0.86	55
tweet	0.54	0.64	0.58	0.46	55
name	0.89	0.35	0.6	0.96	66

The models that uses text-based features Table 2 show slightly better results. Name being the strongest identifier so far, with an accuracy of 66%.

Table 3 - Combined Features



Features	Male Precision	Male Recall	Female Precision	Female Recall	Accuracy %
name, description	0.66	0.8	0.73	0.57	69
name, tweet	0.66	0.71	0.68	0.62	67
name, screen_name	0.65	0.92	0.85	0.49	71
name, created_at	0.65	0.93	0.86	0.47	70
tweet, description	0.61	0.62	0.6	59	60
tweet, name, descripti	0.68	0.73	0.7	0.65	69

However, as we combine the features used in the singular feature based models, the prediction accuracy rises, as high as 70% Table 3. Following John D. Burger's et al (Burger, Henderson, Kim, & Zarrella, 2011) success with using the name, tweet, description and screen_name, we chose to combine this type of data in various combinations to feed as features into our models. Those combinations that returned high accuracy are shown in the above table Table 3

We believe that these results aren't strong enough to used as a classifier in a commercial or professional environment. Our best model, which used name and screen_name data Table 3, was able to reach an accuracy of 71%, which is less than the 92% reported by John D. Burger et al (Burger, Henderson, Kim, & Zarrella, 2011) and the 80% reported by Clay Fink et al (Fink, Kopecky, & Morawski, 2012).

Hence, name is the most indicative of gender, and when combining feature data into a single model, name and screen_name are the most indicative.

5 LIMITATIONS AND OUTLOOK

The main limitation that we encountered, was that a major part of our data was not linearly separable (particularly the numeric data). A polynomial kernel was used to cater for this fact, but still the predication accuracy remained low.

Text classification showed better results over numerical features. An exploration into classifying users based on their tweets could offer more promising results

BIBLIOGRAPHY

Burger, J. D., Henderson, J., Kim, G., & Zarrella, G. (2011). *Discriminating Gender on Twitter.* Massachusetts: The MITRE Corporation.

dbaghern. (2017, October 30). *A Deep Dive Into Sklearn Pipelines*. Retrieved from www.kaggle.com: https://www.kaggle.com/baghern/a-deep-dive-into-sklearn-pipelines

Fink, C., Kopecky, J., & Morawski, M. (2012). *Inferring Gender from the Content of Tweets:.* Maryland: ICWSM.

Liu, W., & Ruths, D. (2013). What's in a Name? Using First Names. *AAAI Spring Symposium - Technical Report*, 10-16.

Usman, M. (2018, August 27). *Text Classification with Python and Scikit-Learn*. Retrieved from stackabuse: https://stackabuse.com/text-classification-with-python-and-scikit-learn/