An exploration of gender indicative feature data on Twitter

ML1819 Research Assignment 2

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

*Word count - TODO*

[Github Source Code](https://github.com/beakeyd/ML1819--task-107--team-41)

[Github Contributors](https://github.com/beakeyd/ML1819--task-107--team-41/graphs/contributors)

.

TODO Update Contribution Screenshots

|  |  |  |  |
| --- | --- | --- | --- |
| 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 d[beakey@tcd.ie](mailto:beakeyd@tcd.ie) |
|  |

TODO – UPDATE CREDITS - 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 consumer base. Understanding your demographic of users has become a priority of businesses to better interact with them.

Privacy has become more of a concern nowadays on the internet. It is harder for companies to scrape relevant data on their demographics from publicly available information on twitter profiles, such as gender.

We wish to discover how well we can infer or predict the gender of twitter users based off their publicly available profile information.

# 2 Related Work

John D.Burger et al [1] managed to classify a user's gender with an accuracy rate of 92%. They used a Winnow2 linear classifier on text features such as Screen Name, Profile Description and Tweets. This showed success, so we shall try build off those features. To build their ground truth they followed the blogs of users and obtained their gender there. TODO Possibly delete this line, and just mention how the ground truth for our data was obtained.

Clay Fink [2] also examined this problem in detail. They determined gender by searching twitter users Facebook pages todo same as above. They achieved an accuracy rate of 80.6% with a combined feature vector of unigrams, linguistic inquiry word count, and hashtags and using a SVMLight linear classifier.

# 3 METHODOLOGIES

**TODO – Split methodology into multiple parts**

## 3.1 Dataset

## 3.2 Pre-Processing

## 3.3 Algorithms

## 3.4 Evaluation

The dataset was obtained through an online research paper on gender classification [3]. 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.

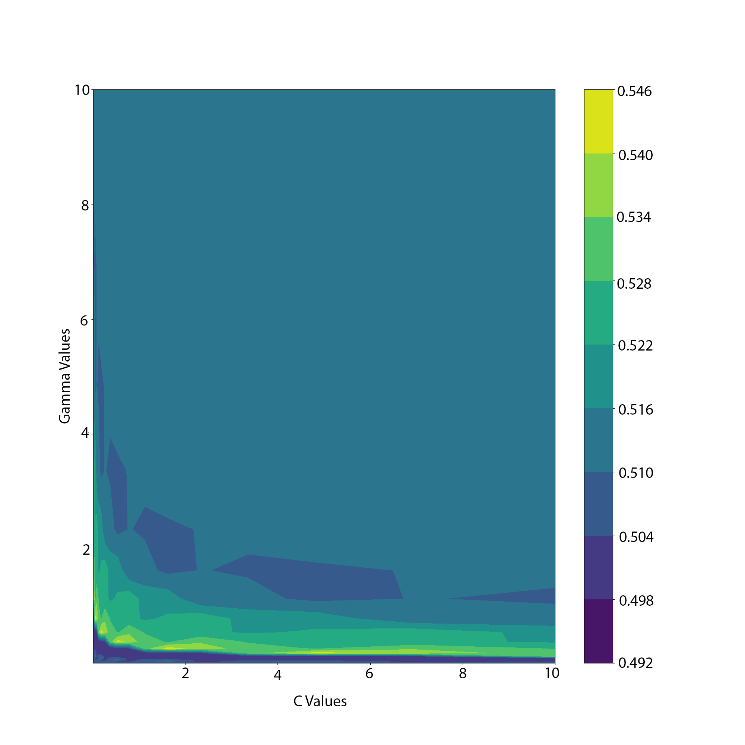
The problem of differentiating a user into the set of males or females, based off their user account data, is that of a classification problem. There were two algorithms that we could have chosen to use in order to classify users; logistic regression and SVM. SVM was the algorithm that was chosen in 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 [4]. 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 [5]. 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.

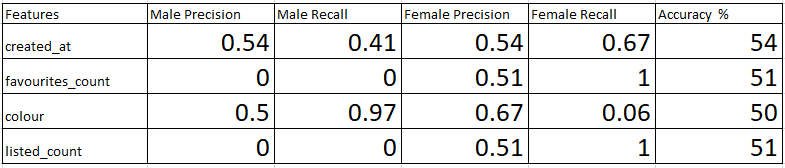
For NuSVC classifiers, we also classified the best nu value through a grid search process.

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 Discussions

Table 1 - Numerical Features



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

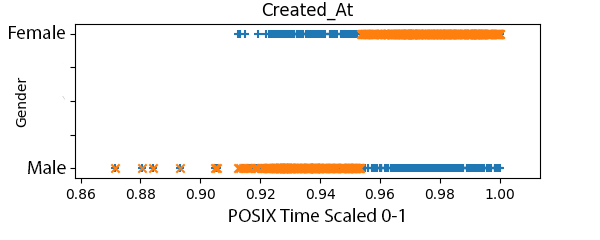
A plot of the Created\_At data Figure 2 shows where the classifier is defining the split in the data. Blue crosses on the graph represent the test data used in the model, while orange X’s represents the predictions on the test data. X Values are POSIX, time scaled between 0 and 1. Accounts created before 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.

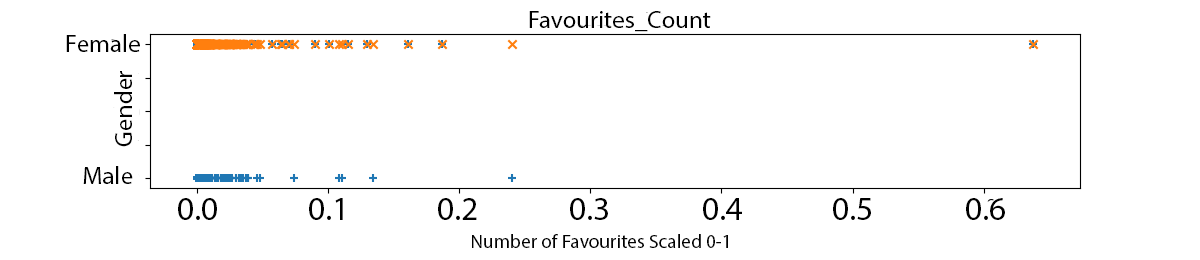
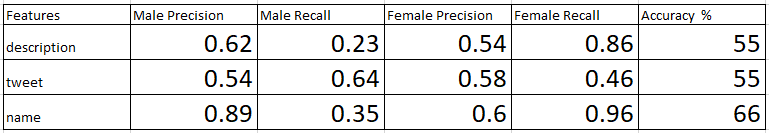
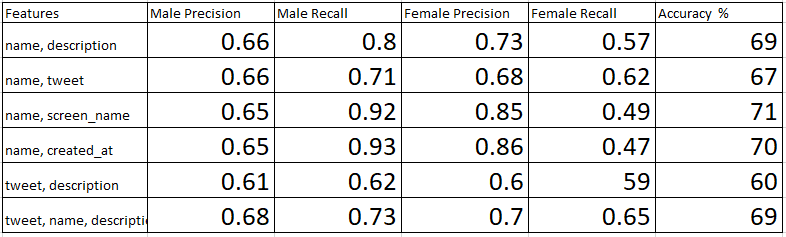
Figure 3 - Number of Favourites

Table 2 - Text Features



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



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 [1] 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 [1] and the 80% reported by Clay Fink et al [2].

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

TODO – add reference comparing our results to previous results

# References

|  |  |
| --- | --- |
| [1] | J. D. Burger, J. Henderson, G. Kim and G. Zarrella, "Discriminating Gender on Twitter," The MITRE Corporation, Massachusetts, 2011. |
| [2] | C. Fink, J. Kopecky and M. Morawski, Inferring Gender from the Content of Tweets:, Maryland: ICWSM, 2012. |
| [3] | W. Liu and D. Ruths, "What’s in a Name? Using First Names," *AAAI Spring Symposium - Technical Report,* pp. 10-16, 2013. |
| [4] | dbaghern, "A Deep Dive Into Sklearn Pipelines," 30 October 2017. [Online]. Available: https://www.kaggle.com/baghern/a-deep-dive-into-sklearn-pipelines. |
| [5] | M. Usman, "Text Classification with Python and Scikit-Learn," 27 August 2018. [Online]. Available: https://stackabuse.com/text-classification-with-python-and-scikit-learn/. |