ML1819 Research Assignment 1

Team 9

107 - Twitter Users Gender Prediction

Team members:

Nicholas Bonello 18307199

Siddharth Tiwari 18300621

Zihan Huang 18300321

Work Contribution:

Nicholas Bonello: Plot and analyze the tweets count and favorite counts, side bar color and link color, color in RGB separately. Process the tweet content and description and apply logistic regression. Write the report.

Siddharth Tiwari: Process the tweet content and description and apply Naïve-Bayes on it. Plot and analyze hashtag count. Plot and analyze emoji count. Write the report.

Zihan Huang: Looking for related works; Plot and analyze the length of tweets and description; Plot and analyze the bar color hex into decimal and plot; Write the report.

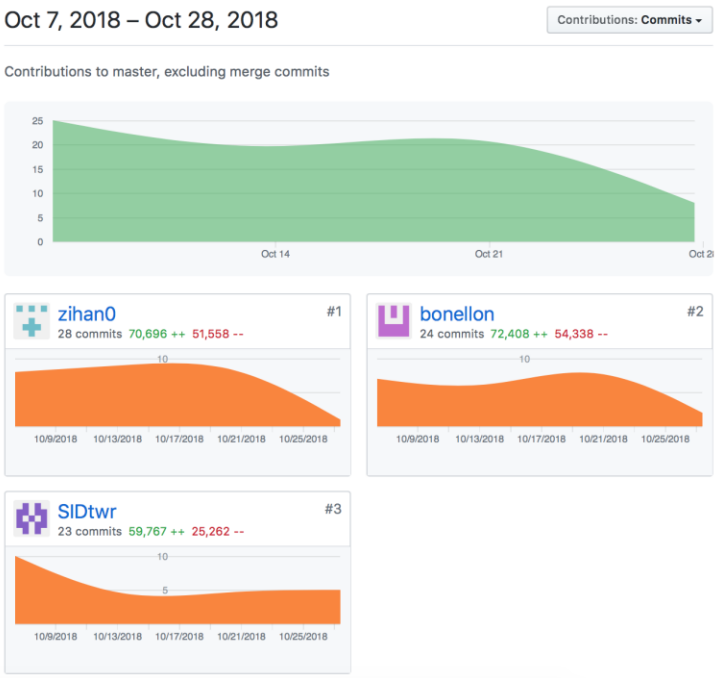
Word Count: 998 words

Source Code Repository:

https://github.com/zihan0/ML1819-task-107-team-09.git

Source Code Repository Activity:

https://github.com/zihan0/ML1819-task-107-team-09/graphs/contributors



ML1819 Research Assignment 1

Twitter Users Gender Prediction

Nicholas Bonello
School of Computer Science and Statics
Trinity College Dublin, University of Dublin
Dublin Ireland
Bonellon@tcd.ie

Siddharth Tiwari
School of Computer Science and Statics
Trinity College Dublin, University of Dublin
Dublin Ireland
stiwari@tcd.ie

Zihan Huang
School of Computer Science and Statics
Trinity College Dublin, University of Dublin
Dublin Ireland
huangzi@tcd.ie

1 INTRODUCTION

Gender prediction is an important tool that can be used to improve existing predictive models. Existing works focusing on gender prediction through blogs or microblogs such as twitter generally make use of analysing the language used in text – in this case tweets and user biography.

In this paper we investigate the possibility of predicting twitter users' gender based on public information. We will evaluate the potential of simple statistical measures such as tweet counts, favourite counts per tweet and profile background colours. We will also apply natural language processing and machine learning algorithms to the text in tweets to understand the differences between male and female twitter users.

2 RELATED WORK

Predicting gender through social media data is generally considered to be a text classification problem. According to Chen *et al.* [1], K-Nearest Neighbour (KNN) is an effective and easily implemented machine learning algorithm, but not perfect for text classification purposes. They proposed an algorithm that combines Latent Semantic Indexing (LSI) methods with KNN to compromise the shortages KNN has. From their results, the effectiveness in processing large scale data improved with the ne w algorithm.

Naïve-Bayes [2] and Support Vector Machine (SVM) [3] also are popular techniques for text classification.

3 METHODOLOGY

A. Data collection

A readily available csv dataset containing a list of tweets and related twitter profile information such as tweet-counts, favourite counts, user biography, etc. was taken from Kaggle [5]. The dataset also provides labelled data on the user gender; male, female or brand.

B. Data Processing

First, we manually removed all the extra columns such as user location that clearly don't have any effect on gender, as well as all the rows that were predicted to be a brand. In order to fine tune our labelled data, we removed all rows where the gender prediction accuracy was less than 80%. Then, we plotted different graphs using two features at a time to determine whether there are any obvious factors that clearly correlate to gender.

C. Machine Learning Algorithm

We attempted to create a classifier that could accurately solve the logistic regression problem of predicting a twitter users gender based on two different features at a time. We used 80% of our dataset to train the model and the remaining 20% as future data to test our model.

The first step was to create graphs that would help us visualise the data so that we can determine which of the features can be used to predict the user's gender.

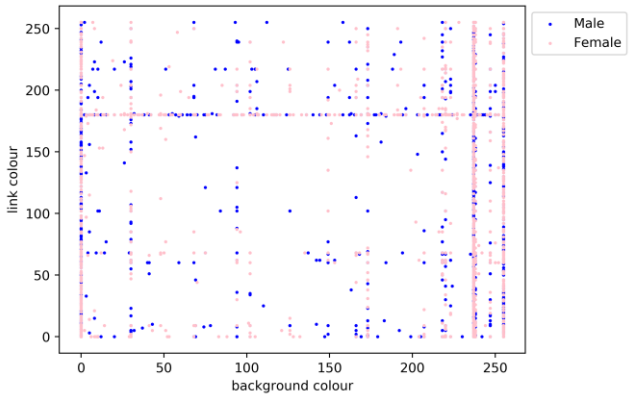


Figure 1 - Link Colour vs Background colour in Red spectrum

The graph above shows the link and background colour frequency distribution in the red spectrum. Values 0 and 255 are the default values which explain why there are so many users picking them.

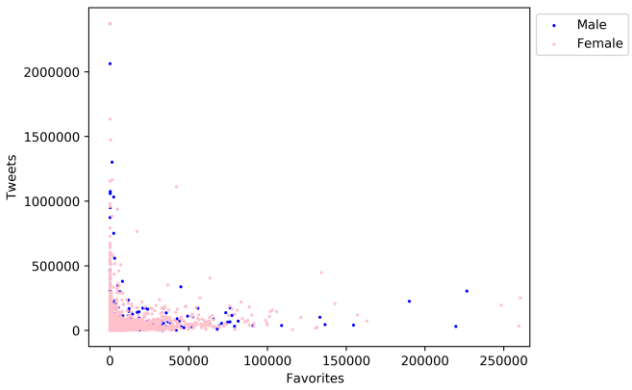


Figure 2 - Number of tweets vs Number of favourites per user

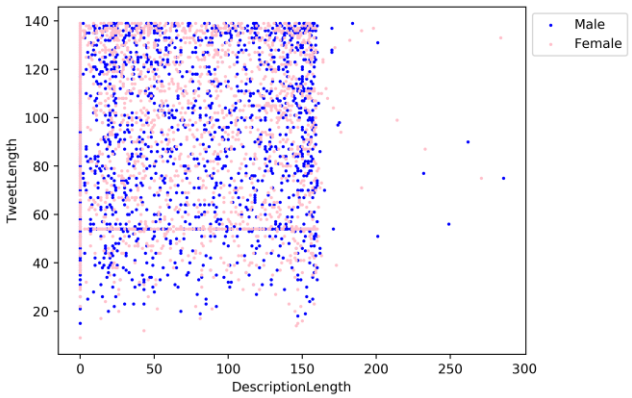


Figure 3 - Tweet Length vs User Description Length

It became instantly clear that none of the provided features could be used to distinguish between a twitter user's gender. Our next approach was to follow the related works methods and analyse both the tweet and the biography text and apply different machine learning algorithms to try and predict gender based on text data, and to what accuracy.

5 RESULTS & DISCUSSION

Our first attempt was to look at the different features such as favourites counts, tweet counts, background colour, link colour and even the number of hashtags used per tweet – labels that were easily obtained from the dataset. These labels were not a good discriminant for predicting gender.

We then created a bag-of-words algorithm that calculates the frequency of word usages per gender. A logistic regression classifier was used to train our dataset on the top words.

When run against the test data, the logistic regression model had an accuracy of 53.34%. To improve on this result, we attempted to use a Naïve-Bayes algorithm instead of the logistic regression but still had relatively weak results – 57.27%.

We explored various options to get a better prediction rate; considering different natural language processing techniques including stop-word removal, punctuation removal and stemming. This improved the scores of both above models as seen in the table below.

Classification	First Model	NLP Techniques
Logistic Regression	53.34%	64.7%
Multinomial Naïve Bayes	57.27%	65.35%

Table 1 – Accuracy Results

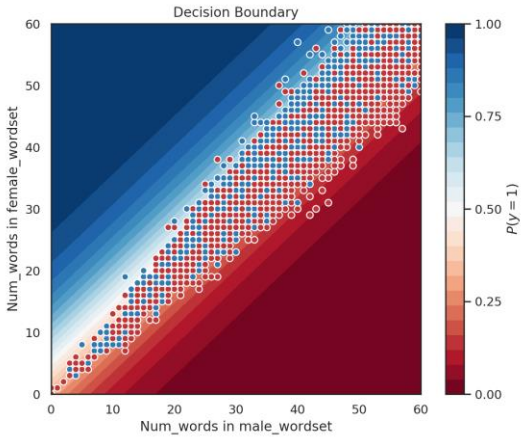


Figure 3 – Decision Boundary depicting our trained logistic regression classifier

All papers that have previously attempted to predict twitter users' gender based on their profile data have all done so through semantic analysis of tweet text and user biographies. Our results agree with this statement, showing that none of the other provided features that were provided in the dataset have any relevance to the gender of that specific user.

Previous works could successfully predict gender with an accuracy of 67.2% when considering randomly obtained tweets using an n-gram model [6].

Burger *et al.* demonstrated that the accuracy of a model improves significantly when more features are considered [5]. As can be seen below, our results match theirs when we only consider the user description and a single tweet.

Baseline (F)	54.9%
One tweet text	67.8
Description	71.2
All tweet texts	75.5
Screen name (e.g. *jsmith92*)	77.1
Full name (e.g. *John Smith*)	89.1
Tweet texts + screen name	81.4
Tweet texts + screen name + description	84.3
All four fields	92.0

Figure 4 - Burger et al. Results

5 LIMITATIONS & OUTLOOK

Given the limited amount of time available for this project, many proposals have been suggested that we simply were not able to do.

For starters, the provided dataset contained extra information while also lacking important information such as additional tweets per user. Creating our own larger dataset with a more users and tweets per user would have provided us with much more training data and better results.

In terms of implementation techniques, we plan to develop an SVM implementation to compare this result with the already obtained results.

In a dataset as small as the one we used, dedicating 20% of our data to test data is a significant amount that would potentially cause a loss of accuracy in our model. In the second phase we plan to implement cross-validation techniques and experiment with the percentages of training and test data to find the most optimal results.

REFERENCES

[1] Jianle Chen, Tianqi Xiao, Jie Sheng and A. Teredesai, "Gender prediction on a real life blog data set using LSI and KNN," 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, 2017, pp. 1-6.

[2] I. Rish, "An Empirical Study of the Naïve Bayes Classifier", In Proceedings of the IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence, Vol. 3, Issue 22, pp. 41-46, 2001.

[3] B.E. Boser, I.M. Guyon, and V.N. Vapnik, "A training algorithm for optimal margin classifiers", COLT '92 Proceedings of the fifth annual workshop on Computational learning theory, pp. 144-152, Pittsburgh, Pennsylvania, USA, July 27 - 29, 1992.

[4] Kaggle.com. (2018). Twitter User Gender Classification. [online] Available at: https://www.kaggle.com/crowdflower/twitter-user-gender-classification [Accessed 9 Oct. 2018].

[5] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In 2nd International Workshop on Search and Mining UserGenerated Content. ACM.

[6] Burger, J.D., Henderson, J., Kim, G. and Zarrella, G., 2011, July. Discriminating gender on Twitter. In Proceedings of the conference on empirical methods in natural language processing (pp. 1301-1309). Association for Computational Linguistics.