ML1819 Research Assignment 2

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.

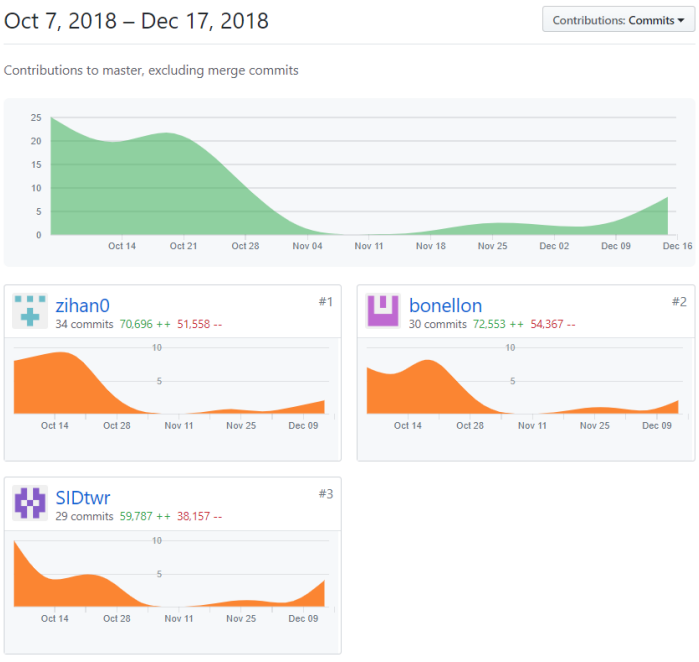
Word Count: 1439 words

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 2

Twitter Users Gender Prediction

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

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

Zihan Huang
School of Computer Science and Statistics
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 microblogs such as twitter generally make use of analysing the language used in text.

In this paper we investigate the possibility of predicting twitter users' gender based on public information. We evaluate the potential of simple statistical measures such as tweet counts, favourite counts per tweet and profile background colours and apply natural language processing 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 new algorithm.

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

3 METHODOLOGY

A. Data collection

A readily available csv dataset containing a list of 20,000 tweets and related twitter profile information including tweet-counts, favourite counts, user biography, etc. was taken from Kaggle [5]. A detailed explanation table is shown below.

Table 1 - Breakdown of Raw Dataset

Column	Description
Gender	Gender of the user. Male, female, brand and unknown.
Created time	The time when this tweet posted
Description	Users' profile
Fav_number	Count of favourite of tweets
Link_color	The colour of link of tweets
Name	Users' name
Sidebar_color	Users' preference in side bar colour
Text	Content of tweets
Tweet_count	Count of tweets
gender: confidence	Gender prediction accuracy
profile_yn: confidence	Prediction accuracy on users' existence

B. Data Pre-processing

The first step was to manually remove all the columns that we deemed unnecessary when attempting to predict user gender; such as a user's location. All users identified as brands or unknown genders were also removed from our model – leaving 65% of the total data. All rows where the gender prediction accuracy was less than 80% was also removed.

The remaining eight columns listed below in table II were used for our classification model, including the gender and confidence which is used as a benchmark to help our model when making decisions.

Finally, graphs of the chosen features were plotted, two at a time for easy representation to determine whether there are any obvious factors that clearly correlate to gender.

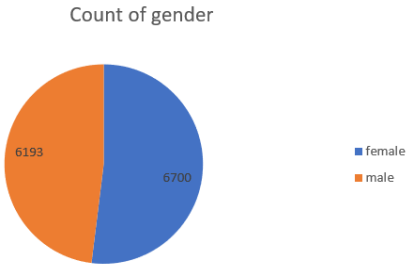


Figure 1- Male and Female counts

Table 2 - Cleaned Dataset Description

Column	Description
Gender	Gender of the user, male or female.
Description	Users' profile
Fav_number	Count of favourite of tweets
Link_color	The colour of link of this tweet
Sidebar_color	Users' preference in side bar colour
Text	Content of tweets
Tweet_count	Count of tweets
gender: confidence	Gender prediction accuracy

C. Machine Learning Algorithms

In the previous phase, we split our data into an 80% training set and the remaining 20% as future data to test our model. This time, we experimented with different percentage values to find how significant minor fluctuations in this percentage are to our final model – described in further detail in the Results section.

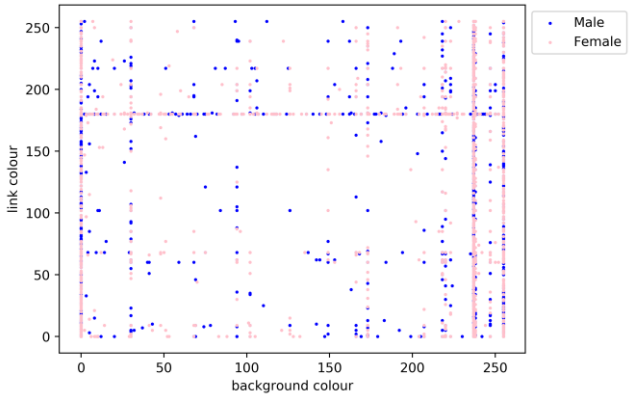


Figure 2 - Link Colour vs Background colour in Red spectrum

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.

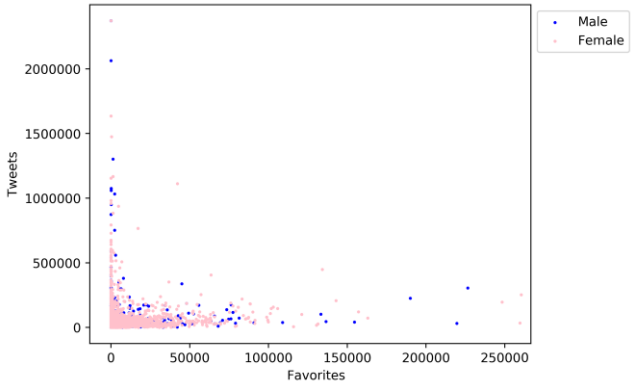


Figure 3 - Number of tweets vs Number of favourites per user

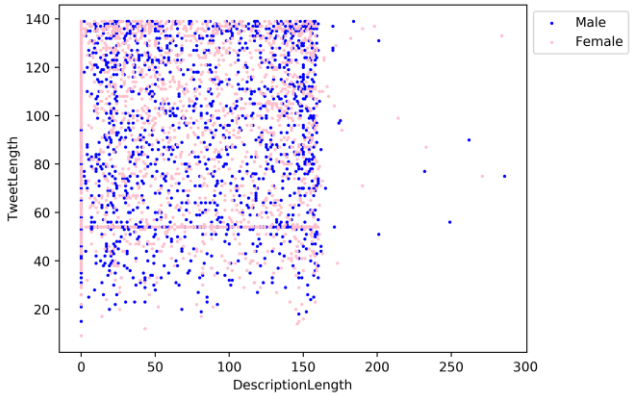


Figure 4 - Tweet Length vs User Description Length

It was clear that none of the provided features could be used to predict a twitter user's gender. Our next approach was to follow the methods seen in the recommended works and analyse both the tweet and the biography text; applying different machine learning algorithms to try and predict gender based on text data.

We created a bag of words from the tweet & description fields of the twitter dataset. The fields were cleaned by converting all words into lower case and removing punctuation. The remaining words are then added to the bag and the frequency distribution of words are calculated. The top 4000 words with the highest frequency are used as features for defining the training data set.

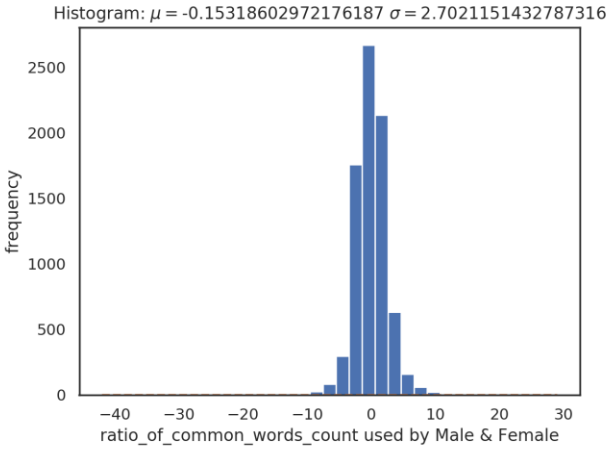


Figure 4 – Distribution of words in the dataset

If a word is a good descriptor for identifying someone as male or female it will predominately be used by either one of the genders.

The above graph plots the ratio of the number of times a common word is used by both genders' vs the number of times the same ratio can be observed for different words. The graph for values of x around x = 0 indicate the number of common words used equally by both genders. As the ratio become more negative the frequency value at that ratio indicates the number of words predominantly used by women. Similarly, for x > 0 the frequency at higher ratios indicate the number of words predominantly used by men.

The plot also gives an estimate of the quality of dataset for training the model. A bad dataset set will show higher levels of frequency or data distribution around zero. A good data set on the other hand will have a balance between the number of common and unique words primely used by male and female. Here the mean is centred around **-0.15** indicting the dataset is slightly biased towards females but the curve resembles a normal distribution.

We are looking for words where the value of **r > 1.4** , **r < -14** for training the model as they are good discriminant of gender.

Table 3 – Example of some unique & common words used by male & female

Male	Female	Common
Battle, victory, playing, economy, tax, government, Ebola etc.	relationships, shopping, besties, cute, fashion, beautiful, love etc.	Angry, regrets, parties, laughing, texting etc.

5 RESULTS & DISCUSSION

Our first attempt involved looking 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 independently to find any indicator that these features correlated with gender. We found that 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. When run against the test data, the logistic regression model had an accuracy of 68.97%.

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 the above models as seen in the table below.

Table 4 – Optimal Accuracy Results

Classification	First Model	NLP Techniques
Logistic Regression	53.34%	68.97%
Multinomial Naïve Bayes	57.27%	65.76%

We compared some different training and test set sizes to evaluate the differences in accuracies. We found that even though the 80/20 split provided the best results over the 10 iterations for all 3 of the different algorithms used.

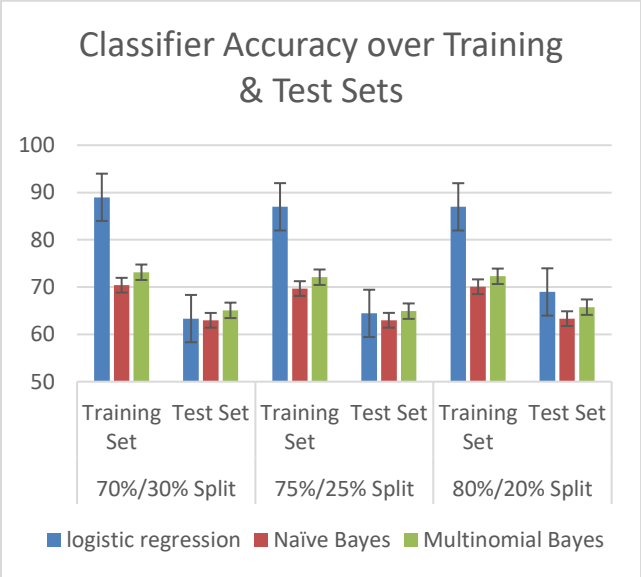


Figure 5 - Algorithm accuracies across Training/Test sets

Table 5 – Complete Results

	70/30 Split	**75/25 Split**	**80/20 Split**
Logistic Regression	63.34	64.44	68.97
Naïve Bayes	62.98	62.99	63.33
Multinomial Bayes	65.08	64.91	65.76

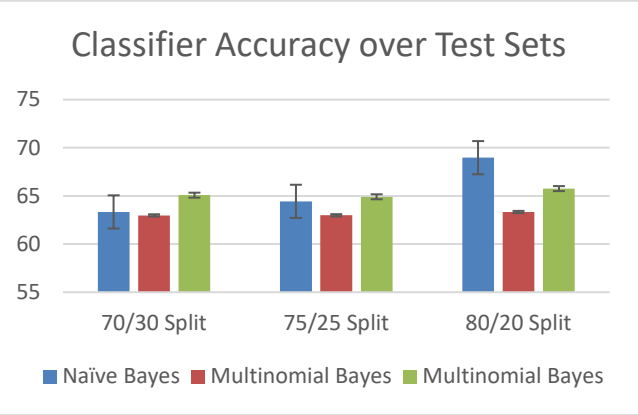


Figure 6 - Algorithm Accuracies over Test Sets

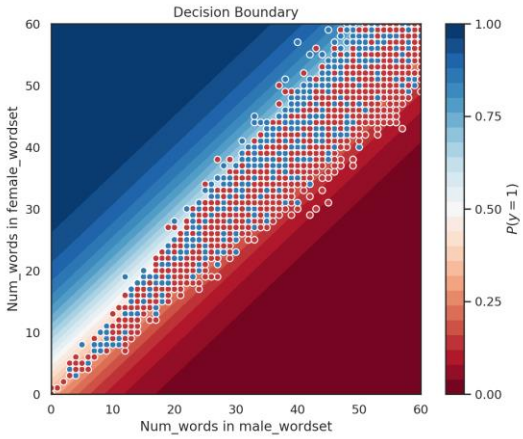


Figure 5 – Decision Boundary depicting our trained logistic regression classifier

Our training set contained N features, so it wasn't possible to visualize the decision boundary without reducing the dimensionality.

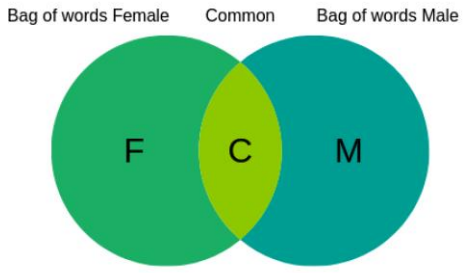


Figure 6 – Venn's diagram depicting the overlap of Male/Female word set

Bag of words male = Unique words M + common word(M)
Bag of words female = Unique words F + common word(F)

In order to solve this problem, we created two separate word lists; the bag of words male & female which contain unique words either used by males or females. To select common words with a high discriminant value, a list of words commonly used by males and females is created.

For every word W_i in the common word list L a ratio r_i (r_i = count_Male(W_i) / count_Female(W_i)) between the number of times and the word is used by male and female is calculated. If the ratio is between 1 – 1.4 the word is discarded as it is not a good discriminant. For values greater than 1.4, the words are added to common words(M). For values of ***r*** less than one, their reciprocal is calculated, and the word is added to common word(F) if the ratio is greater than 1.4.

if(r>=1) and (r>=1.4)
    add word to Common Word(M)
if(r==1) or (r<=1.4)
    discard word (poor discriminant)
if(r<=1) and (1/r>=1.4)
    add word to Common Word(F)

These bags of words are then used to create a training set that takes the number of times the words appear in the male and female word sets. A maximum accuracy of 68.97% was achieved when the cut off value of **r** was set to 1.4 using logistic regression.

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 in the dataset show any relevance to gender.

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 7 - Burger *et al.* Results

6 LIMITATIONS & OUTLOOK

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.

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.