ML1819 Research Assignment 1

Team 9

Twitter Users Gender Prediction

Team members:

Nicholas Bonello 18307199

Siddharth Tiwari 18300621

Zihan Huang 18300321

Work Contribution:

Nicholas Bonello:

Siddharth Tiwari:

Zihan Huang:

Word Count:

(Exceptional Circumstances )

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

Commit activity:

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 (10)

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

In this paper we investigate the possibility of predicting twitter users’ gender based on public information. We will be evaluating the potential of using simple statistical measures such as tweet counts, favourite counts per tweet, profile background colours and link colours.

We will also be applying natural language processing and machine learning algorithms to the text in tweets to try and understand the differences between male and female twitter users.

2 RELATED WORK (10)

There are numbers of papers on predicting personal attributes based on social data [1]. Kosinski *et al.* demonstrated that even simple algorithms can predict personal attributes based on the patterns of Facebook likes an indicator of peoples’ preferences [2].

A service called Personality Insights, designed and developed by IBM provides personality traits including factors such as the users personality, needs, and even personal values [3].

3 METHODOLOGY (30)

1. Data collection

A readily available dataset containing a list of tweets and related twitter profile information was found. This dataset has approximately 20,000 tweets, along with its creator’s gender, the time when it was published, the side bar colour of the creator’s profile, the total tweet count from the creator, total count of favourites on this tweet, content of this creator’s description and the tweet’s content.

1. Data Processing

Data processing was performed in several steps. First, we decided to plot different graphs using just two features at a time to determine whether there are any obvious factors that clearly determine gender.

First, we attempted to use simple statistical data just as analysing the tweet count, favourites count, even the length of a tweet or the number of hashtags used. Then we converted the different background colours and link colours into RGB values – and attempted to extract just the one of the three primary colours to visualise any gender indicators based on colour.

1. 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.

Afterwards, we also attempted to implement NLP algorithms to process the tweet text. To do this we removed all English stop-words and calculated the word frequency per-gender for the rest of the words. Finally, we applied a Naïve-Bayes classifier on our training set and calculated the accuracy of our newly trained model on our test-set.

These two different approaches were then compared to each other to determine which model is better at predicting gender.

4 RESULTS & DISCUSSION (30)

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

We plotted the dataset using tweet-counts vs favourite counts, as shown in the figure 1 below. Females are represented as pink dots and males as blue dots.

Figure 1

As it can be seen in figure 1, there’s no clear separating line. Females and males are seemingly randomly spread all across the graph. This was expected for this graph since it would not make sense for the number of tweets and favourites to be distinctly different based on gender.

Afterwards, we tried plotting the tweet length compared to the user description length as seen in figure 2 below, but again there was no distinction between genders.

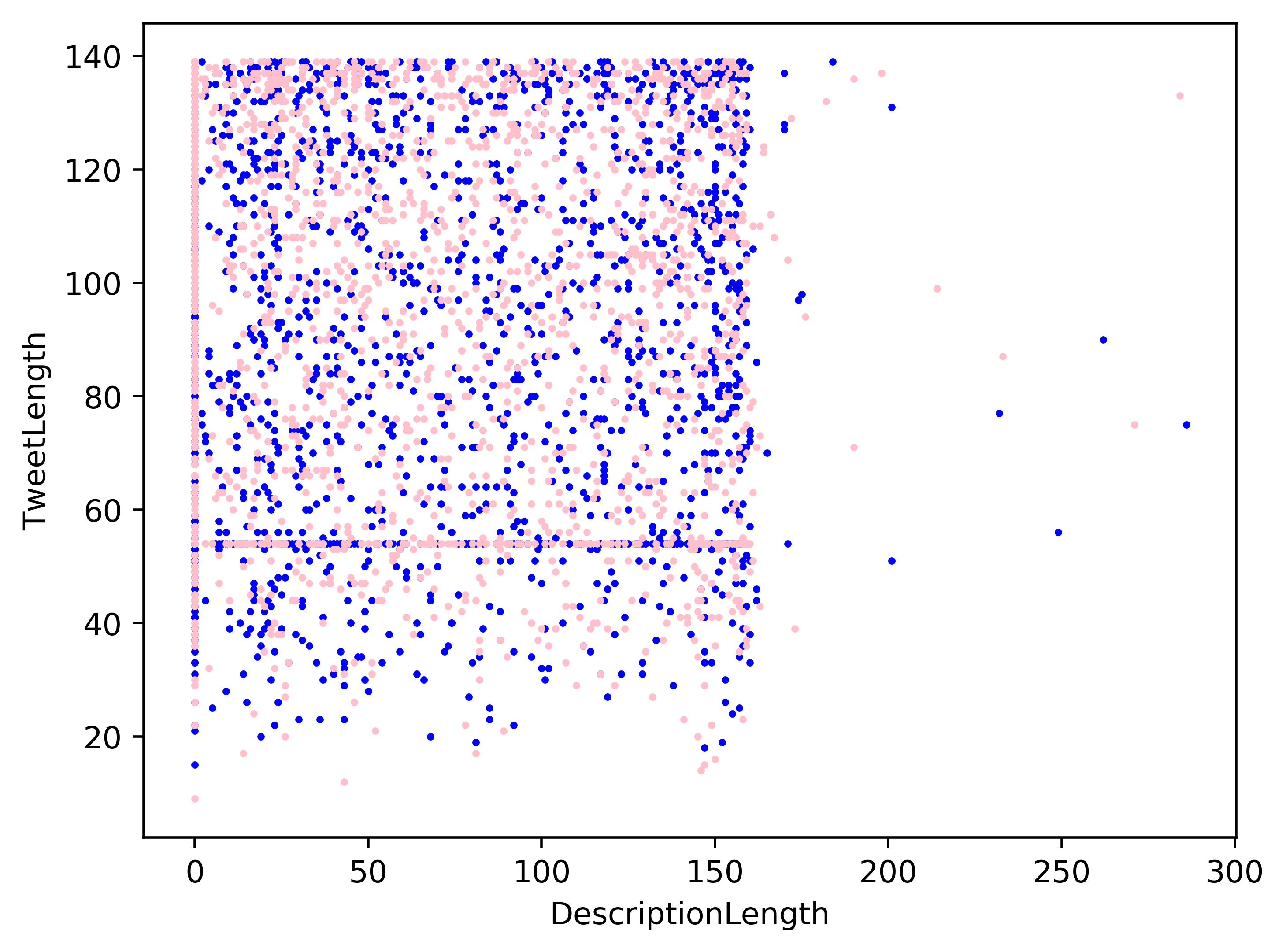


Figure 2

Furthermore, we also plotted the dataset using link and background colours as seen below in figure 3. The graph represents the red-spectrum link and background colours when converted from hex-triplets to RGB.

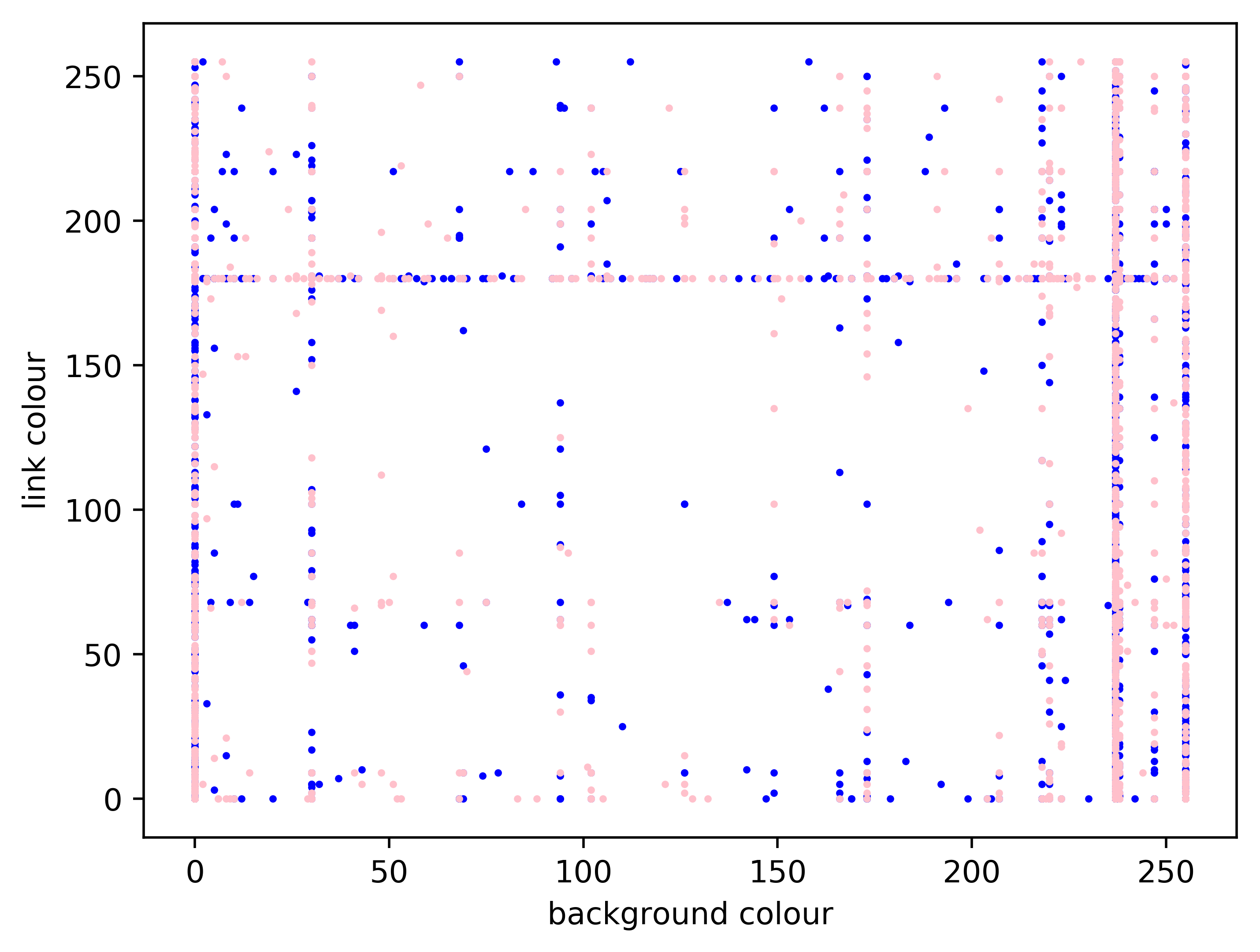


Figure 3

The graph represents the red colour spectrum however all other spectrums produce similar results. We found out that while there is quite a difference in colour customization by users, there is no clear difference between male and female users.

From there, we decided to seek possibility in content. Based on different writing habits between two genders, by processing the text, there may appeals a better classification solution.

5 LIMITATIONS & OUTLOOK (5)

ACKNOWLEDGMENTS

REFERENCES

[1] Yo Take,Sasahara Kazutoshi. *2017 IEEE International Conference on Big Data (Big Data) Big Data (Big Data), 2017 IEEE International Conference on*. :3168-3174 Dec, 2017.

[2]  M. Kosinski, D. Stillwell, and T. Graepel, “Private traits and attributes are predictable from digital records of human behavior,” Proceedings of the National Academy of Sciences, vol. 110, no. 15, pp. 5802–5805, 2013.

[3] “IBM Watson Personality Insights,” https://www.ibm. com/watson/services/p ersonality- insights/.

[4] Jain Amita, Singh Amishapriya, Tayal Devendra Kumar, Vij Sonakshi, “Amalgamating Data Analytics and Machine Learning for Predicting Sex Ratio and Infant Mortality Rate to Improve Gender Composition,” Global Journal of Enterprise Information System. Jul-Sep2016, Vol. 8 Issue 3, p1-5. 5p.

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00