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 analyzing 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 without analyzing the any language usages. Instead, we will be evaluating the potential of using simple statistical measures such as tweet counts, favorite counts per tweet, profile background colors and link colors. We will also be considering some simple tweet metrics such as the hashtag frequency, punctuation and even smileys to fully understand the differences between male and female twitter users.

**2 RELATED WORK (10)**

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

Also, this is not limited to academic fields, a service called Personality Insights was designed and developed by IBM, personality traits including the Big Five factors, needs, and values can be predicted by it [3].

**3 METHODOLOGY (30)**

1. *Data collection*

We looked up online and found a dataset of tweets and related information. This dataset has approximate 12,000 tweets, along with its creator’s gender, the time when it’s published, the side bar colour from the creator’s setting, the total tweet count from the creator, total count of favourite on this tweet, content of this creator’s description and tweet’s content.

1. *Data Processing*

Data processing was performed in several steps. We decided to plot the dataset using only two features at a time in order to see if there exists a possible separating line. In order to do this, we proc-

ess the text content into data for example length of it, that can be used to draw graph.

We calculated the length of each tweets, convert the hexadecimal color data to decimal.

We plotted the dataset using favorite count and total tweets count, and there’s no correlation shown to separate gender.

1. *Machine Learning Algorithm*

Since we are try to predict the gender based on two features of a certain tweet, and this is a simple classification problem. So we decided to use the logistic regression algorithm.

We use 80 percent of our dataset to train the model and the rest 20 percent as future data to test out our model.

**4 RESULTS & DISCUSSION (30)**

We plotted the dataset using tweets count and favourite count, as shown in the figure 1 below. Females are shown with pink dots and males are shown as blue dots.

Figure 1

As it can be seen in figure 1, there’s no clear separating line. Females and males lie randomly among the graph.

And we plotted the dataset using tweets length and user description length, graph shown below as figure 2.

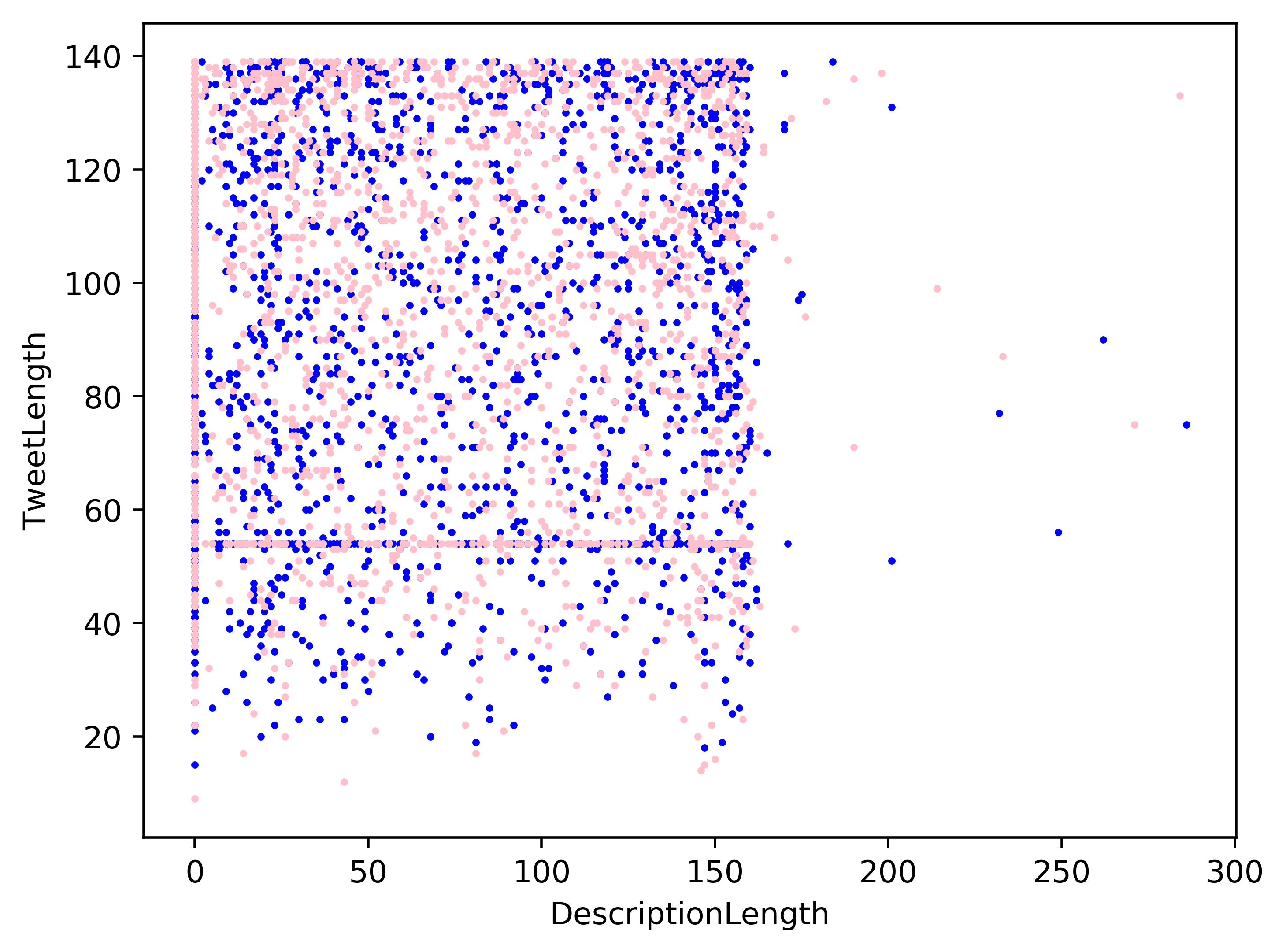


Figure 2

As can be seen in figure 2, there is no possible way to do a good classification on genders by these two features.

Furthermore, we plotted dataset using tweets count and side bar colour and as shown below in figure 3.

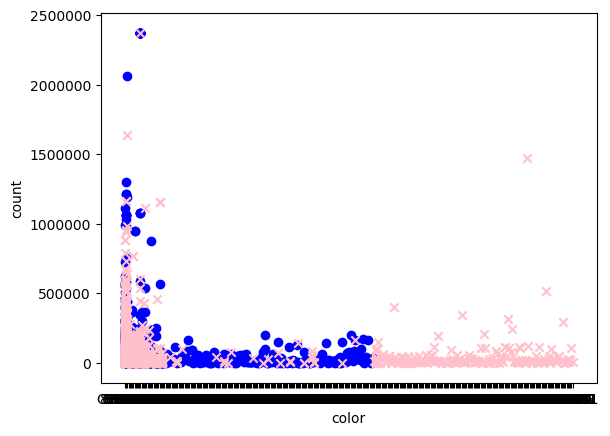


Figure 3

We found out that there is a quite difference in colour that customize by users, we decided to look into that and do more processing on the colour data.

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