ML1819 Research Assignment 1

• Team 28 • Task 107: How well can the gender of Twitter users be predicted? • Words •

Source Code Repository: <https://github.com/dealrachaan/ML-1819-task-107-team-28>

Source Code Repository Activity: <https://github.com/dealrachaan/ML-1819-task-107-team-28/graphs/contributors>

Dataset: <https://www.kaggle.com/crowdflower/twitter-user-gender-classification>

Ethan Hamman

15324710

Implemented Methodology

Pre-processed Dataset

Wrote Report

Jane D’Altuin

15315850

Implemented Methodology

Sourced Dataset

Pre-processed Dataset

Wrote Report

Machine Learning: Using Keywords to Identify Twitter Users by Gender

Jane D’Altuin†  
 SCSS  
 Trinity College Dublin  
 Dublin, Ireland  
 daltuinj@tcd.ie

Ethan Hamman  
 SCSS  
 Trinity College Dublin  
 Dublin, Ireland  
 [hammane@tcd.ie](mailto:hammane@tcd.ie)

INTRODUCTION

Machine Learning is often used to classify users of an online service into demographic traits in order to better tailor the service to users’ interests and to present the platform to advertisers who are interested in targeting certain demographics. The research problem posed to this project is whether the gender of a Twitter user can be identified given details of their profile. This project uses the metrics of the text of a tweet from the user, the user’s profile description (known as a ‘bio’), the user’s chosen colour to accent their profile, and the user’s screen name. The WEKA library of machine learning algorithms was trained on a large dataset to identify the gender of users from this data. Being able to identify the gender of Twitter users would be useful when deciding what tweets and profiles to promote to users as well as deciding what ads to show to these users.

RELATED WORK

Deitrick et al (2012) [1] used a neural network called the Modified Balanced Winnow to classify Twitter users’ gender based on Tweet content as well as other traits selected by WEKA’s attribute selection algorithms. The writers understood men and women to write differently. While the character limit of Twitter (140 at the time of publication) limits the amount of data per tweet that can be used for classification, this also pushes users to engage in abbreviations and initialisms. As the type of language used in these contexts tends to be heavily stylized and influenced by social groups, this makes identifying users’ gender easier.

Filho, Castro and Pasti (2016) [1a] analysed Twitter text and categorized the writers into male and female categories based on meta-text attributes such as syntax, lexicon, structure and morphology. Analysing tweets made in Portuguese by Brazilian journalists, they found that characters and syntax and textual structure were some of the strongest indicators of the gender of the author.

METHODOLOGY

The dataset ‘**Twitter User Gender Classification’** from Kaggle was used. This data had been sorted by humans into being from a male, female or branded Twitter account. Some of the data was classified as being unknown, and each data item was given a confidence interval associated with its assignment if it was classified as male, female or brand.

This data was presented as an Excel spreadsheet containing the text of a tweet, the Tweet author’s perceived gender, a confidence interval for that gender assessment, the author’s Twitter name, handle, profile description, and details about the Tweet and author such as date Tweet was posted, date the profile was created, number of favourited/liked Tweets, etc.

This data was first trimmed to contain only the details of users who were male or female with an assessment confidence of 1. This yielded 12,894 items of the original 20,000-item data set. In this set were 6,194 tweets from men (48.04%) and the rest (6700 tweets, 51.96% of the set) were from women. This data was then reduced to show only the text of a tweet from the user, the user’s profile description (known as a ‘bio’), the user’s chosen colour to accent their profile, and the user’s unique handle. The colour was encoded using hex coding. This data was saved to an .ARFF file.

The intention for this program was to identify the words, self-descriptions and profile colours associated with different genders in this data set.

The WEKA Machine Learning Library was used to assess the data. The WEKA Zero Result algorithm was used to establish a baseline to compare other algorithms against. The algorithms used were the User Classifier algorithm and Naïve Bayes Multinomial Result algorithm. All of these algorithms were tested using 10-fold cross-validation.

RESULTS AND DISCUSSION

This program correctly assesses the gender of Twitter users based on Tweet content 50.15% of the time. This figure comes from running the program over a test data set which comprised 10% of the processed data for a total of ~1000 Tweets, of which half were from female users and half from male users.

Due to difficulties in implementing the program, the point at which Twitter data transitioned to Tweets from female users to Tweets from male users was hard-coded into the program. For future versions of this project, it would be desired to enable the program to read the gender of the user from training data without this process.

This project takes a blunt approach to assessing Twitter users’ gender and ranks all keywords equally. A more refined implementation of this program would involve weighting words by gender and using this weighted scoring system to assess Twitter users’ gender.

LIMITATIONS AND OUTLOOK

The primary limitation on this project was that there was insufficient work done to complete the course code. The completion of this work will have to be done to develop on this project. Variations on the methodology to consider going forward include varying the percentage of most popular words by gender considered, weighting words by their popularity by gender, and analyzing the impact of other factors such as profile description and Twitter username on gender analysis.

REFERENCES

[1]

[2]

[3] Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. Twitter User Gender Inference Using Combined Analysis of Text and Image Processing. (August 2014). Accessible at http://www.aclweb.org/anthology/W14-5408

[4] H. Andrew Schwartz, Johannes C. Eichenstadt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. (September, 2013). Retrieved October 16, 2018 from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0073791

[5] Ana Brassard, Tin Kuculo. 2017. Stereotypes Can Be in Writing Too – Age and Gender Identification of Twitter Users. (July 2017). Accessible at <https://www.fer.unizg.hr/_download/repository/TAR-2017-ProjectReports.pdf#page=11>

[6] Statistia (2018). *Global Twitter user distribution by gender 2018 | Statistic*. [online] Statista. Available at: https://www.statista.com/statistics/828092/distribution-of-users-on-twitter-worldwide-gender/