A Replication of "How Old Do You Think I am?" A Study of Language and Age in Twitter

Anish Gadekar
anishraj@buffalo.edu

Chris Chan
conloonc@buffalo.edu

Shruti Bendale
shrutitu@buffalo.edu

Abstract

In this paper we attempt to replicate the experiment done in the paper "How Old Do You Think I am?" by Dong Nyugen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. Much like the original paper, we will focus on the connection between age and language use among Twitter users. Essentially, we will study how to predict the age of Twitter users based on the content of their tweets. Furthermore, we will analyze the tweets to generate some interesting results.

Introduction

A person's language use reveals a lot about their social identity. It can be based on various groups such as age, gender, political affiliation, etc. The original paper set out to predict the social identity of Twitter users because they believed an individuals language use reveals these traits of the user. They believed social identity can be predicted based on various features such as age, gender, political affiliation, and etc. We found this interesting and decided to replicate their work to see if our results will be similar to theirs.

One of the reasons we chose to replicate this paper, out of any other paper, is because this paper uses Twitter data, as well as, Linear and Logistic Regression. We also had two hypotheses we wanted to test before attempting their research. The first would be that our results would match theirs and in this case our results would support their work. The other would be perhaps we would get an entirely different result, in which case we would like to find out why. We then conduct this experiment using an annotation task which is used to study the relation between a person's language and their age and gender. The prediction task was performed using logistic regression and support vector machines. We also use the collected tweets to identify tweet patterns for users of different age groups.

Corpus Collection

We collected a large set of Tweets from the Twitter API. This set contained over 4000 tweets. We then had around 1000 tweets annotated by 3 annotators. The tweets were collected and annotated for the gender and the age-range by 3 annotators [See Figure 1].

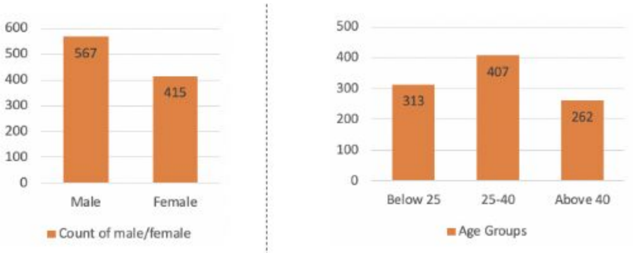

Figure 1: Distribution of tweets across male/female gender (left) and age ranges (right)

Manual Prediction

During our manual prediction phase, we had 3 annotators which were the members of our team. Our team consisted of male and female annotators who are candidates for master's degrees, so the relative age of our team would be considered young. After gathering over 4000 tweets via the Twitter API, each team member would pseudo randomly pick at least 300 tweets each to annotate. In the event where the tweets lead to profiles that could not be viewed, or were not in English the tweets would be discarded and another tweet would be chosen in its place. Some other rules we set in place for annotation tweets include

(1) The account should be publicly accessible
(2) The account should represent a person (ie. the account does not represent a group or an organization) and
(3) The account should have sufficient tweets (at least 10).

We gained access to each tweet via a tweet ID through Twitters website. Then we predicted the age and gender of the Twitter user based on the content of their tweets and their profile picture. Our annotators could take as much time as needed to make a prediction, but in general used several hours to complete the task as a whole.

Inter-Annotator Agreement

The annotations by two annotators were considered for calculating the inter-annotator agreement. A total of 982 tweets were annotated by both. The Inter-Annotator Agreement was evaluated by the third annotator using Cohen's Kappa by evaluating two trials of the same sample. A Kappa value of 1.0 was found for gender and a Kappa value of 0.76 was found for age.

Age Prediction

In this section, we focus on the ways we approached the different features of age prediction. It is also important to note the categories we used are different from the original paper. One such difference is the age category, as we felt many of the Twitter users ages fell under 40. Therefore, we gave 3 categories for users under 40 as we felt it gave our predictions better distinction.

Age Category: 15-, 15-25, 25-40, 40+

We also thought the similarities between the ages in these groups would have more in common with each other, in comparison to the age group categories in the original paper which were 20-, 20-40, 40+.

With the help of the annotated dataset, we trained Logistic Regression and Support Vector Machine models using sklearn libraries. We represent the results in the next section.

Data Preprocessing

We replaced all the user mentions (@user) by a common token. The special characters and stop words are removed. We only keep words that occur at least 10 times in the training documents.The tweets were then vectorized by calculating the tf-idf scores of all the words in the vocabulary.

Methods

We experimented with and compared the predictions of three Machine Learning algorithms:

1. A binary Logistic Regression classifier was used to classify the tokenized tweets with respect to the gender. We use a one versus all method to handle multiclass classification of the age groups.

2. A Support Vector Machine was also experimented with to perform the two classification tasks.

Results

The accuracies for all the 3 methods are given as follows:

Models	**Accuracy**	
	Gender	**Age**
Logistic Regression		
Support Vector Machines		

We think that the accuracy could have been better with a bigger dataset.

Analysis

We performed an analysis of the words occurring most frequently with the help of word clouds. After analysing the tweets, we could see a clear difference between the tweet styles and the words used by people in different age groups [See Figure 3]. We also found that females write slightly more than men (average number of tokens: 2235 versus 2130).

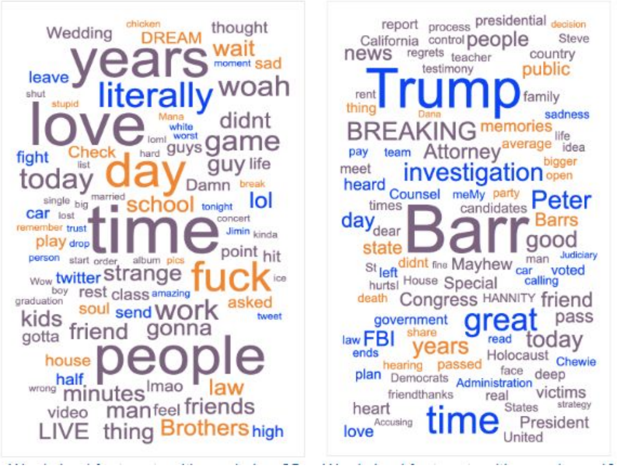

Figure 3: Word clouds for tweets with age below 25 (left) and for tweets with age above 40 (right)

We also noticed that the tweets belonging to people in the age groups 25-40 and 40+ were very similar in nature and hence we represent a common word cloud for both age groups.

Another interesting thing to note is that older people tend to tweet about politics and current events while the word cloud for the younger people was more universal. If a similar analysis is done on a twitter dataset a year later, one will notice a change in the word cloud for the older people. However, we can conclude that the word cloud for the younger people may be the same.

Conclusion

In accordance with the original paper, our models were based on the tweets of users. This has a practical advantages because the data is easy to collect, therefore our models can easily be applied to new Twitter users. One’s social media presence is only a representation of a single facet of one’s identity. A person can choose to represent different facets of their identity at any given moment. One can only say that while younger individuals adopt an informal style of writing on social media platforms such as Twitter, the style of writing becomes more formal and less juvenile for older individuals.

References

Nyugen, D.; Gravel, R; Trieschnigg, D.; and Meder, T. 2013. How Old Do You Think I Am? A Study of Language and Age in Twitter

Argamon, S.; Koppel, M.; Pennebaker, J.; and Schler, J. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression.

Pennancchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to Twitter user classification. In ICWSM 2011.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2012. Gen- der in Twitter: styles, stances, and social networks.

Richard’s deep learning blog

https://richliao.github.io/supervised/classification/2016/11/26/textclassifier-convolutional/

https://www.kaggle.com/tunguz/logistic-regression-with-words-and-char-n-grams