HOW OLD DO YOU THINK I AM? A Study of Language and Age in Twitter

(A replication project from the research by Dong Nguyen et al)

Shruti Bendale, Anish Gadekar and Chris Chan

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. With the help of Twitter data, we perform a fine grained annotation task which can be used to study the relation between a person's language and their age & gender.

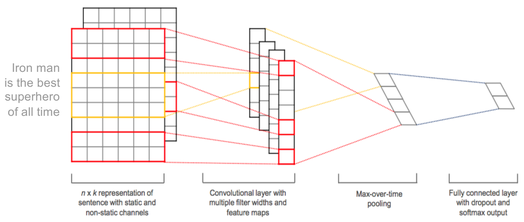
Methods

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.

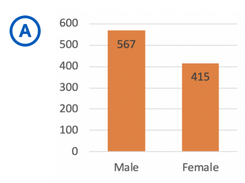
3. Filters were applied to a **Convolutional Neural Network** by using the Keras Merge Layer.

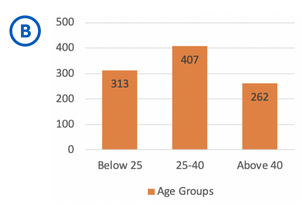


Data Collection

The tweets were collected and annotated for the gender and the age-range by 3 annotators [See Figure A & B]. The ground rules set for the annotation task were as follows:

A. The account should be publicly accessible.
B. The account should represent a person.
C. The account should have sufficient tweets (at least 10).





Inter-Annotator Agreement

Two annotators performed annotations. A total of 982 tweets were annotated by both. The Inter-Annotator Agreement was evaluated by a 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.

Data Preprocessing

A. All user mentions (@user) are replaced by a common token.
B. The special characters and stop words are removed.
C. We only keep words that occur at least 10 times in the training documents.
D. The tweets were then tokenized by calculating the Tf-Idf scores of all the words in the vocabulary.

Results and Analysis:

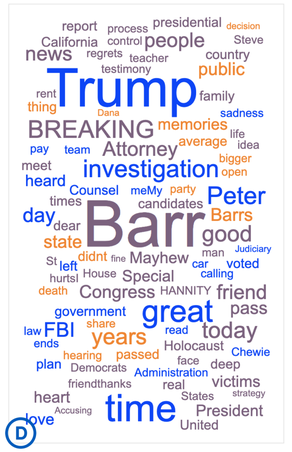
The accuracies for all the 3 methods are given as follows:

	Accuracy	
	Gender	Age Range
Logistic Regression	60%	63%
Support Vector Machine	67%	64%
Convolutional neural network	78%	82%

The convolutional neural seemed to learn the features better than the other two methods. The accuracy could have been better if the dataset was bigger.

After analyzing the tweets, we could see a clear difference between the tweet styles and the words used by people in different age groups [See Figure C & D]. We also found that females write slightly more than men (average number of tokens: 2235 versus 2130)

Word cloud for tweets with age below 25



Word cloud for tweets with age above 40

Conclusion

Our models were based only on the tweets of users. This has as a practical advantage that the data is easy to collect, and thus the 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 also say that while younger people 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 people.

References

1. Nguyen, D.; Gravel, R.; Trieschnigg, D.; and Meder, T. 2013. How Old Do You Think I Am? A Study of Language and Age in Twitter.
2. Argamon, S.; Koppel, M.; Pennebaker, J.; and Schler, J. 2007. Mining the blogosphere: age, gender, and the varieties of self-expression.
3. Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to Twitter user classification. In ICWSM 2011.
4. Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2012. Gen- der in Twitter: styles, stances, and social networks.
5. Richard's deep learning blog; https://richliao.github.io/supervised/classification/2016/11/26/textclassifier-convolutional/
1. https://www.kaggle.com/tunguz/logistic-regression-with-words-and-char-n-grams

University at Buffalo The State University of New York

Computer Science and Engineering
cse.buffalo.edu