Twitter & Gender

The Tweeps (Group 19)
D.J. Anderson, Nishant Jha, Edlene Miguel, Kirsten White

#problemStatement

There is a lot of metadata associated with each tweet that gets posted.

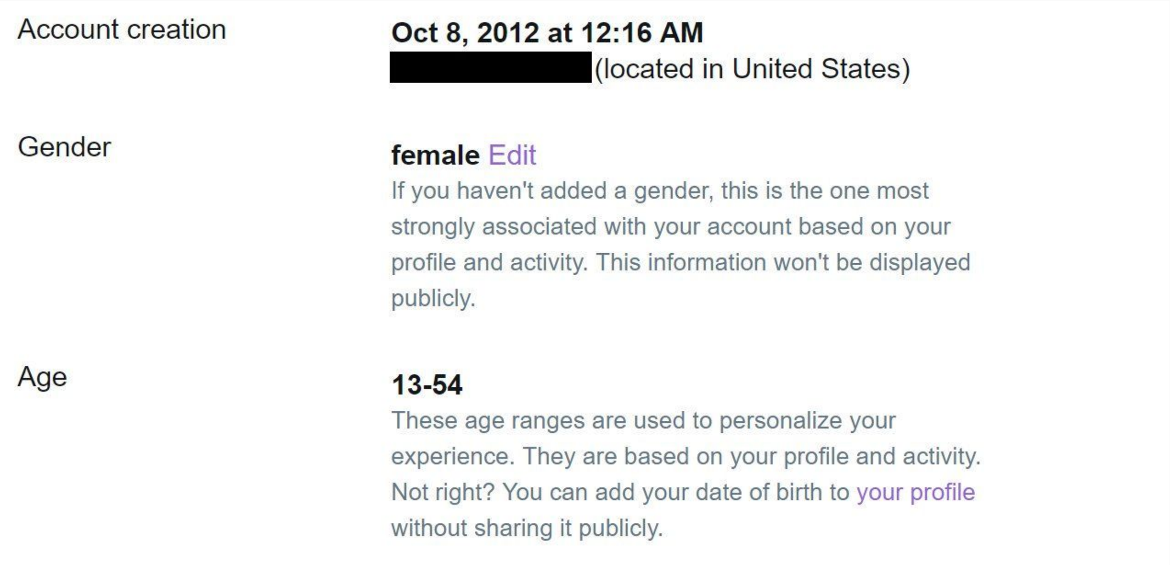
Can we identify a Twitter user’s gender based on their tweet metadata?

#motivation

- User Verification: Are you who you say you are?
  - Safety - dating apps, Craigslist, etc.
- Business Analysis
  - Targeting markets: showing ads to specific audience
  - $$$

#background

In May 2017, users began noticing that if they had not added a gender to their profile already, Twitter speculatively did it for them.



Account creation

Oct 8, 2012 at 12:16 AM
[redacted] (located in United States)

Gender

female Edit
If you haven't added a gender, this is the one most strongly associated with your account based on your profile and activity. This information won't be displayed publicly.

Age

13-54
These age ranges are used to personalize your experience. They are based on your profile and activity. Not right? You can add your date of birth to your profile without sharing it publicly.

https://twitter.com/Vronos/status/865238426429386752/photo/1

#background

- https://forge.fiware.org/docman/view.php/47/5275/GenderRecognitionAlgorithmGRA.pdf
  - Decision trees and waterfall evaluation to obtain ~87% accuracy
  - Used screen name, description, and profile color attributes
- http://www2.cs.uregina.ca/~hilder/my_students_theses_and_project_reports/ugheokeMScProjectReport.pdf
  - Support vector machine to obtain ~87% accuracy
  - Used ranker-filter algorithms from WEKA toolkit
    - Feature-selection algorithms rank features based on the score computed for each feature
      - Used chi-square, information gain ratio, relief, and symmetrical uncertainty algorithms

#challengesAndGoals

- Emojis
- Bigrams
- Data includes non-numeric and numeric values
- Irregularities in informal written language
- Gender Types/Categories?

#data

- Data includes:
    - User data (name, user description, date/time account created, number of tweets favorited, profile image, etc.)
    - Tweet data (date/time posted, text, location, timezone, etc.)
    - Gender classification data
        - Male/Female/Brand (gender)
        - Confidence value (gender:confidence)
        - Golden/Finalized - if classification has been confirmed or not (_unit_state)

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

#preprocessing

- Parsed tweets and descriptions into individual words
- Removed stopwords
- Removed non-alphabetic characters and accented characters
- Translated color hex codes into color names, then into numeric labels
- Identified most common words for each gender

#featureSelection

- Computed “maleness”, “femaleness”, and “brandness” of tweets and descriptions
- Omitted date/time of account creation, date/time of tweet creation, user location, and tweet location
- Kept description content, tweet content, link color, sidebar color, number of tweets, number of retweets, and number of favorites

#process

- Preprocess data
- Perform feature extraction and selection
- Execute algorithms (decision tree and kNN)
- Compute and compare metrics

#algorithms

- Decision Trees
- K - Nearest Neighbors

#modelSelection

- k-fold cross validation
    - Decision tree: k = 4
    - kNN: k = 10

#results

- kNN implementation
  - Optimal number of neighbors: 33
  - Accuracy: 48%
- Decision Tree implementation
  - 4 folds: 51.7%, 51.3%, 50.6%, 51.4%
  - Max: 51.7%
  - Average: 51.3%

#softwareAndHardware

Software:

- Anaconda Notebook (spyder, jupyter)
- sklearn
- pandas
- matplotlib

Hardware:

- Personal Laptop Computers

#conclusions

- Overall, we experienced mild success with our implementations of machine learning algorithms, but could have seen marked improvement upon that success had more features been included. Some examples of these features are discussed in the “Future Work” section later in this paper.

#references

Crowdflower. Twitter User Gender Classification | Kaggle, 21 Nov. 2016, www.kaggle.com/crowdflower/twitter-user-gender-classification.

Notopoulos, Katie. “Twitter Has Been Guessing Your Gender And People Are Pissed.”BuzzFeed, www.buzzfeed.com/katienotopoulos/twitter-has-been-guessing-your-gender-and-people-are-pissed?utm_term=.sfe90R3MnL#.ovGDpX9j7o.

“Classification And Regression Trees for Machine Learning.” Machine Learning Mastery, 20 Sept. 2017, machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

http://www2.cs.uregina.ca/~hilder/my_students_theses_and_project_reports/ugheokeMScProjectReport.pdf

https://forge.fiware.org/docman/view.php/47/5275/GenderRecognitionAlgorithmGRA.pdf

#questions