Latent Attribute Inference of tweeple

Puneet Singh Ludu

Version: 0.1 Date: 19, Feb, 2014

Algorithm

1. Machine Learning models

- Profile features(twitter bio, name, twitter handle, website) [DONE]
- Tweeting behaviour features(e.g. tweets per day) [Work in progress]
- Linguistic features of users [Work in progress]
  - Currently I am using frequency of LIWC2007 classification of words as features
- Network Features [ToDo]
  - I am planning to use verified users by classifying them into genres such as sports, news, entertainment, technology etc. and using them as feature.

Updates

- Downloaded Twitter Bio based data and using bio based features to classify gender
- Downloaded Tweets
- Written code for cleaning the tweets and spell correction
- Written code to load LIWC2007 dictionary to categorize words in tweets in 30 different categories.
- In process of extracting tweets based features
- Next milestone: creating a list of verified users followed by users in the dataset categorizing them manually. for example, Justin Beiber would be categorized as “Teen pop artist” and most active five years “2010-2015”

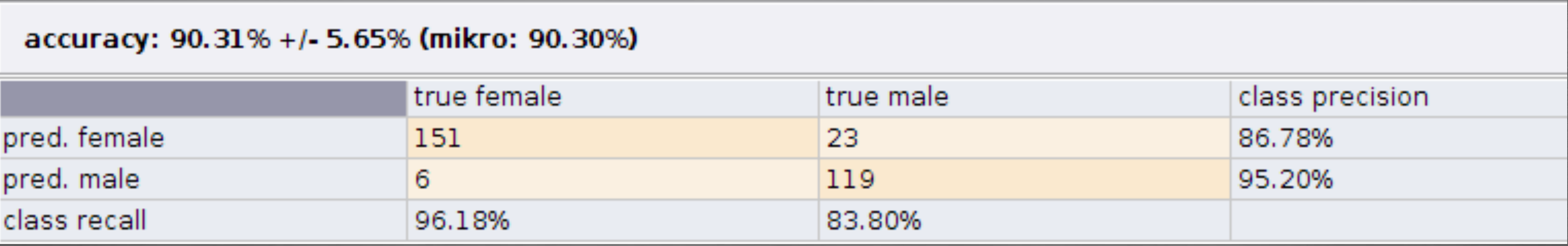
Results

Preliminary results for **Gender classification** based on only profile(bio) based features: Name and screen name

Using: SVM, 10-fold validation on 299 user data

Code: https://github.com/puneetsl/tlassify-gender

Results of various runs: http://goo.gl/QaIKX8



accuracy: 90.31% +/- 5.65% (mikro: 90.30%)

	true female	true male	class precision
pred. female	151	23	86.78%
pred. male	6	119	95.20%
class recall	96.18%	83.80%	

LIWC2007 categories

- pronoun
- article
- verb
- past
- present
- future
- negate
- social
- family
- friend
- affect
- positive emotion
- negative emotion
- anger
- sad
- insight
- cause
- inclusion
- exclusion
- feel
- body
- health
- sexual
- achieve
- money
- death

Tweets based features

- Word category based: using LIWC 2007
- Number of K-top differentiating words for an individual used by group of labeled users
- K-Top bigrams, trigrams
- K-Top hashtags
- Tweeting frequency (tweets per day)

Social Features

- Followers to following ratio
- Verified users following to following ratio
- Verified users categories
- Verified users most active 5 years range

Road blocks

1. Out of 384 users used by Zamal, 86 have closed or deactivated their twitter account, their informations can not be retrieved.
2. Out of 298 remaining users, 35 have locked their timeline this their tweets can not be downloaded.

Thank You