Featured Dataset

50

Twitter User Gender Classification

Predict user gender based on Twitter profile information

Crowdflower • last updated a year ago

Tags internet twitter gender small featured

Top Contributors

Ed King	1st
RVK	2nd
Kande Bonfim	3rd

Kernels

Classifying User Gender ... run a year ago	13 votes
Gender Identification - A... run 9 months ago	10 votes
Looking for meaningful p... run a year ago	2 votes

Discussion

_golden standard 5 months ago	2 replies
Gender Identification - ... 5 months ago	3 replies
Fix the tweet_id field in ... 8 months ago	0 replies

Description

Help us describe this dataset Edit

This data set was used to train a CrowdFlower AI gender predictor. You can read all about the project here. Contributors were asked to simply view a Twitter profile and judge whether the user was a male, a female, or a brand (non-individual). The dataset contains 20,000 rows, each with a user name, a random tweet, account profile and image, location, and even link and sidebar color.

Inspiration

Here are a few questions you might try to answer with this dataset:

- how well do words in tweets and profiles predict user gender?
- what are the words that strongly predict male or female gender?
- how well do stylistic factors (like link color and sidebar color) predict user gender?

Acknowledgments

Data was provided by the Data For Everyone Library on Crowdflower.

Our Data for Everyone library is a collection of our favorite open data jobs that have come through our platform. They're available free of charge for the community, forever.

The Data

The dataset contains the following fields:

- **_unit_id**: a unique id for user
- **_golden**: whether the user was included in the gold standard for the model; TRUE or FALSE
- **_unit_state**: state of the observation; one of *finalized* (for contributor-judged) or *golden* (for gold standard observations)
- **_trusted_judgments**: number of trusted judgments (int); always 3 for non-golden, and what may be a unique id for gold standard observations
- **_last_judgment_at**: date and time of last contributor judgment; blank for gold standard observations
- **gender**: one of *male*, *female*, or *brand* (for non-human profiles)
- **gender:confidence**: a float representing confidence in the provided gender
- **profile_yn**: "no" here seems to mean that the profile was meant to be part of the dataset but was not available when contributors went to judge it
- **profile_yn:confidence**: confidence in the existence/non-existence of the profile
- **created**: date and time when the profile was created
- **description**: the user's profile description
- **fav_number**: number of tweets the user has favorited
- **gender_gold**: if the profile is golden, what is the gender?
- **link_color**: the link color on the profile, as a hex value
- **name**: the user's name
- **profile_yn_gold**: whether the profile y/n value is golden
- **profileimage**: a link to the profile image
- **retweet_count**: number of times the user has retweeted (or possibly, been retweeted)
- **sidebar_color**: color of the profile sidebar, as a hex value
- **text**: text of a random one of the user's tweets
- **tweet_coord**: if the user has location turned on, the coordinates as a string with the format "[*latitude*, *longitude*]"
- **tweet_count**: number of tweets that the user has posted
- **tweet_created**: when the random tweet (in the **text** column) was created
- **tweet_id**: the tweet id of the random tweet
- **tweet_location**: location of the tweet; seems to not be particularly normalized
- **user_timezone**: the timezone of the user

Did you find this Dataset useful?
Show your appreciation with an upvote

50

Recent Activity

Zach Barnes Ran version 2 of kernel **TWEEETER** 8 days ago

FrankCipollone	Ran version 1 of kernel **Notebook9b526a22d0**	8 days ago
NinaCesare	Commented on dataset discussion **_golden standard**	5 months ago
VijayRaj	Created kernel **Title**	5 months ago
RVK	Commented on kernel **Gender Identification - Analysis**	5 months ago

Similar Datasets

Twitter US Airline Sentiment

Demonetization in India Twitter Data

How ISIS Uses Twitter

#Charlottesville on Twitter

#Inauguration and #WomensMarch Tweets

25,555 views · 3,085 downloads · 5 topics

Our Team Terms Privacy Contact/Support