Text-based Gender Prediction via Deep Learning and Random Forests

Miroslav Vitkov, Prashant Dangwal

January 9, 2019

1 Task

Gender prediction over a short direct speech in written form can be valuable in various fileds. One possible application is estimating the effectiveness of prose authors to describe a character of a different gender than themselves analogously to the Turing test. An abundant source of real-world training and test data is twitter. A corpus will be created and then annotated with profile details of the posters. After collection, the corpus will be trimmed to posts from users with sufficient profile details and number of tweets. A naive DL and RF models will be trained on minimal number of features. Performance metrics will be established and reported. Time permitting, the classifier part will be expanded both by number of algorithms and by algorithm sophistication.

2 Baseline model

Our baseline model will be the observation that 55

3 My Subtask

I will prepare the boilerplate code: corpus collection, data filtering, automated data labling, naive classifier, estimating performance. Prashant will work on advanced classifiers.

4 Challenges

Obtain twitter developer id. Allocate time to collect sufficient data. Avoid biase sampling. Prune obtained corpus. Select features. Operate new software (DL and RF models). Ensure statistical correctness of reported results. Present all those in an academic report.

5 Objectives

Overcome more than half of the challanges.

6 Very very very awklard bibliography

Bill Heil and Mikolaj Jan Piskorski. 2009. New Twitter re- search: Men follow men and nobody tweets. Harvard Busi- ness Review , June 1