Author profiling

Lovre Mrčela, Marko Ratković, Ante Žužul

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

{lovre.mrcela, marko.ratkovic, ante.zuzul}@fer.hr

June 12, 2016

Content

Introduction

What is author profiling?

Introduction

What is author profiling?

- Author profiling is process of determining author age-group, gender and Big five personality traits.

Introduction

What is author profiling?

- Author profiling is process of determining author age-group, gender and Big five personality traits.
- For this project we are trying to profile author base on their tweeter post.

Introduction

What is author profiling?

- Author profiling is process of determining author age-group, gender and Big five personality traits.
- For this project we are trying to profile author base on their tweeter post.
- Process have two separate task:

Introduction

What is author profiling?

- Author profiling is process of determining author age-group, gender and Big five personality traits.
- For this project we are trying to profile author base on their tweeter post.
- Process have two separate task:
  - classification task for age-group and gender

Introduction

What is author profiling?

- Author profiling is process of determining author age-group, gender and Big five personality traits.
- For this project we are trying to profile author base on their tweeter post.
- Process have two separate task:
  - classification task for age-group and gender
  - regression task for Big five personality traits

Introduction

- Dataset for task was taken for PAN competition.
- Dataset consist 4 language: English, Spanish, Italian and Dutch.
- Official test set isn't available due this year PAN competition.

Approach to solving problem:

- find optimal set of features
- find optimal model

Preprocessing tweets

- substituting *url* with ***URL***
- substituting *usernames* (referenced in replies) with ***REPLY***
- removing stop words from set of user tweets
- converting all tweets to lowercase
- removing repetitions of letter in word (e.g 'coooool' to 'cool')

Set of features

Features used for this problem:

- tf-idf weighting scheme used on trigrams representation of preprocessed user tweets
- number of emoticons
- number of consecutive long repetitions of characters
- number of replies
- number of hashtags
- number of exclamation marks
- average length and standard deviation of posts
- average length and standard deviation of words

Models

Model used for classification:

- Logistic Regression
- Naive Bayes Classifier
- Decision Tree Classifier
- Random Forest Classifier
- SVC (using *rbf*, linear, poly and sigmoid kernels)

Models

Model used for regression:

- Linear Regression
- Decision Tree Regressor
- Random Forest Regressor
- SVR (using *rbf*, linear, poly and sigmoid kernels)

Testing

- Training dataset was divided into subset for training (70%) and for validation (30%)
- Optimal model and hyperparameters were selected by using 10–fold cross–validation
- We have used baseline models because official test set wasn't available at this time

Overview of additional features for each age-group per language

Ne mogu staviti Table 1

Overview of additional features values for gender per language

Table 2

Overview of results of age–group classification per language

table 3

table 4

table 5

Conclusion

- We succeeded to obtain results similar to other published works
- Possible upgrade:
  - Latent Semantic Analysis

The End

Tnx for listening!

Questions??