Author Profiling

Borna Sirovica,Filip Zelić,Ivan-Dominik Ljubičić

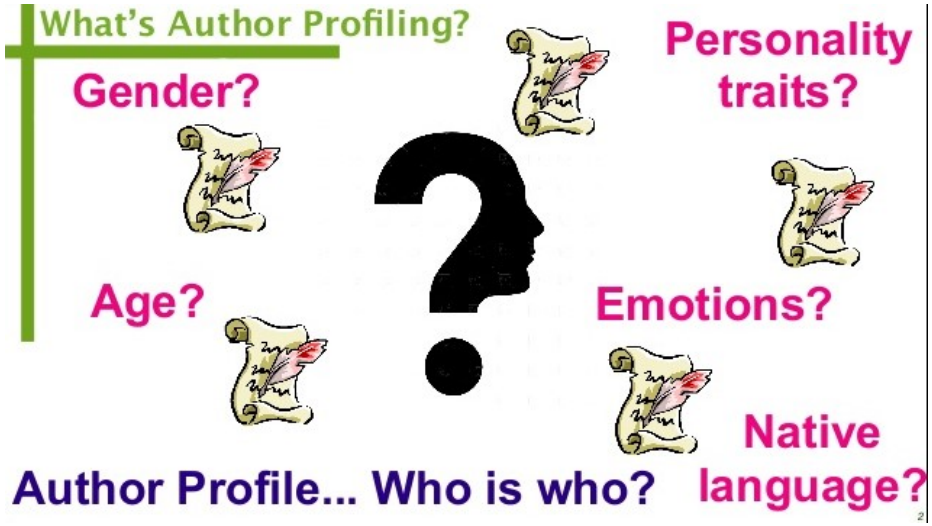
Fakultet elektrotehnike i računarstva

June 13, 2016

Content

Author Profiling

- Identifying information about an author by analyzing their writing style
- Distinguishes between classes of authors studying their sociolect aspect
- Growing importance in forensics, security, marketing, and social networks

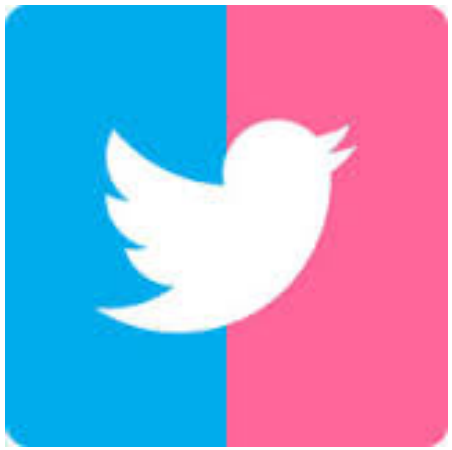


Our task

- Predicting author profiling aspects
  - Demographics (Classification)
    - age
    - gender
  - Big Five personality traits (Regression)
    - extroversion
    - stability
    - agreeableness
    - conscientiousness
    - openness

Dataset

- English tweets dataset from PAN 2015 competition
  - XML format, 152 users
  - Age and gender labels
    - 18-24, 25-34 35-49, 50+
    - M, F
  - Personality traits
    - normalized numeric rating in [-0.5,0.5] range



Preprocessing

- Creating a document of tweets for each user
- Cleaning tweets from
  - XML tags
  - hashtags
  - @mentions
  - URLs
  - stop words
  - duplicates

Feature extraction

- Custom designed features
- Extracted from publicly available resources

Features

- Lexicon Features
  - NRC Word-Emotion Association Lexicon
  - Internet slang word lexicon
  - Frequent male and female words lexicon
  - Frequent male and female function words lexicon
- Twitter-specific Features
  - hashtags
  - URLs
  - user mentions
  - emoticons

Features (cont'd)

- Ortographic Features
  - letter case
  - character flooding
  - word length
  - tweet length
  - specific characters (exclamation points, question marks, apostrophes...)
- Term-level Features
  - n-grams (TF-IDF matrices of unigrams and trigrams)
  - F-score

Machine learning models

- Classification models
  - Support Vector Machine Classification(SVC) with linear kernel and RBF kernel
  - Logistic Regression
  - Random Forest
- Regression models
  - Support Vector Machine Regression (SVR) with linear kernel and RBF kernel
  - Linear Regression

Model selection techniques

- Splitting the dataset into a training set and a test set (70–30%)
- Hyperparameter optimization
  - Grid search
  - k-fold cross validation
- Scaling feature values
- Feature selection tools
  - Tree-based feature selection
  - Recursive feature elimination

Results

- Age evaluation
  - Best feature combination
    - Twitter-specific features
    - Orthographic features
    - TF-IDF scores of frequent trigrams
- Gender evaluation
  - Best feature combination
    - Male and female based lexicons
    - Orthographic features
    - F-score measure

Subtask	Classifier	Accuracy score
Gender	SVM (RBF kernel)	77.2 %
Age	SVM (RBF kernel)	76.1%

Results (cont'd)

- Personality traits evaluation
  - Best feature combination
    - Word-emotion based lexicons
    - F-score measure
    - POS tags frequencies
    - Orthographic features
  - Root mean squared error(RMSE) measure between predicted and actual score

Personality trait	Regression model	RMSE
Extroverted	SVR (RBF kernel)	0.142
Stable	SVR (RBF kernel)	0.170
Agreeable	SVR (RBF kernel)	0.145
Conscientious	SVR (RBF kernel)	0.153
Open	SVR (RBF kernel)	0.156

Implementation

- All coding done in Python
- Used libraries
  - scikit-learn
    - ML models
    - Grid search
    - Feature selection tools
    - Cross-validation
    - Scalers (Standard, MinMax)
  - nltk
    - POS tagger
    - TF-IDF vectorizer
    - Porter stemmer
    - Tweet tokenizer

Conclusion

- Challenging task
- Various combinations of features used for training our machine learning models
- Good results
- Future work:
  - Evaluate on real test dataset
  - Use better feature selection tools
  - Better usage of lexicons

End

Thank you for attention.

Questions, comments ?