ENSEMBLE-BASED CLASSIFICATION FOR AUTHOR PROFILING USING VARIOUS FEATURES

M. Czoków1, M. Meina1, K. Brodzińska2, B. Celmer1, M. Patera3, J. Pezacki4 and M. Wilk5
1Faculty of Mathematics and Computer Science, 2Faculty of Physics, Astronomy and Informatics Nicolaus Copernicus University, Toruń, Poland

Introduction

This paper summarizes our approach to author profiling task – a part of evaluation lab PAN’13. We have used ensemble-based classification on a large set of features. Here, all the features are roughly described and evaluation for random forest (ensemble-based classifier obtaining the best accuracy) is presented.

Work methodology

Analytic dataset consists of:
- 8 groups of features.
- Total number of features is 311 for English and 476 for Spanish.

Two approaches:
1. Random forest — eventually applied.
2. Committee for 8 weak classifiers:
   - 8 subsets of features
   - for each subset four classifiers tested (kNN, Linear SVM, SVM with RBF and Naïve Bayes)
   - for each subset of features the best classifier took part in voting

Classification accuracy for Random Forest

<table>
<thead>
<tr>
<th>gender</th>
<th>age</th>
<th>gender + age</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>0.632 ± 0.0019</td>
<td>0.611 ± 0.0019</td>
</tr>
<tr>
<td>Spanish</td>
<td>0.611 ± 0.0027</td>
<td>0.596 ± 0.0089</td>
</tr>
</tbody>
</table>

Baseline 0.3333 0.5 0.4550

Experiments were conducted using k-fold cross-validation with (k = 10).

Topic-specific features

We applied Latent Semantic Analysis:
- With each document we associate 150 coefficients of different topics.
- In order to obtain this we create tf-idf weighted term-document matrix $M$ and approximate its singular value decomposition:
  $$M \approx U_k \Sigma_k V_k$$
- where $U_k$ and $V_k$ can be interpreted as term-topic matrix and topic-document matrix,
- In order to avoid overfitting, topic-specific features enclosed in the analytic dataset are generated with application of 10-folds cross-validation.

Structural features

- Features that describe structure of conversations, e.g. the number of conversations, paragraphs, sentences, special characters and words per sentence.
- Statistics for documents with more than one conversation:
  - minimum, maximum and average conversation length,
  - average edit distance between each pair of conversations.
- Statistics concerning hyperlinks and images.

Cluster analysis

- We created clusters on the base of two groups of features:
  - structural,
  - topic-specific.
- In the set of features we added distances from centroids.

Behaviour profile ↔ author profile

<table>
<thead>
<tr>
<th>Feature</th>
<th>English</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. age</td>
<td>60.66</td>
<td>60.18</td>
</tr>
<tr>
<td>Avg. length</td>
<td>36.74</td>
<td>36.18</td>
</tr>
<tr>
<td>Avg. error</td>
<td>1.76</td>
<td>1.85</td>
</tr>
<tr>
<td>Avg. correction</td>
<td>0.14</td>
<td>0.13</td>
</tr>
<tr>
<td>Avg. number</td>
<td>90.00</td>
<td>89.00</td>
</tr>
<tr>
<td>Avg. idiom</td>
<td>14.47</td>
<td>14.75</td>
</tr>
</tbody>
</table>

Dictionary-based features

In each document we counted number of:
- abbreviations,
- emoticons,
- bad words,
- basic emotion words (e.g. anger, disgust, fear, joy, sadness, surprise),
- connective words (e.g. nevertheless, whatever, secondly)
- words that have little semantical value (e.g. I, the, own, him)
- perseverative words (e.g. you, money, sure, new, roads, health, camp)

Text difficulty & readability

Features based on the following readability formulas: Flesch Reading Ease, Flesch-Kincaid Grade Level, Dale-Chall. These statistics are based on the number of words, sentences, syllables and difficult words (there is Dale-Chall list of 3,000 familiar words and thus, words, which are not on that list, are considered as difficult).

Sequences of parts of speech

- Preprocessing — each sentence tagged into sequence of parts of speech.
- For each document we calculated an average probability, that a tagged sequence from this document belongs to the respective classes (separately for gender and age).
- In order to do this we created 4-gram models (we calculated conditional probabilities that for a given class a given tag occurs in a sequence, when it is preceded by a given sequence of $n - 1$ length).

Errors

Numbers of errors and language mistakes in accordance with the list of 27 standardized ISO 27 errors’ types.

Parts of speech

- Preprocessing — each sentence tagged into sequence of parts of speech.
- Frequencies of particular parts of speech in all conversations of each author.
- Much more parts of speech (features) for Spanish (this number is predefined by tagger).