1. Task

- A set of Twitter users and their posts were provided.
- Set was divided into four languages: Italian, English, Dutch, Spanish.
- Each user’s gender, age and personality were given.
- Task was to predict age, gender and personality of unseen users, given a single set of these known users.

2. Data and processing

- Users were equally balanced by author gender. No guarantee of equal balance for other attributes, i.e. age had a marked imbalance.
- As authors may have different numbers of tweets, Up- (and down-) weighting of users was tested to avoid over-fitting to particular authors.
- Investigating effects of Hyperlinks was addressed in two ways:
  - Collecting domain of hyperlinks, replacing hyperlinks with special token.
  - The effect of shares and retweets were not considered in this approach.
- A single document representing each user was formed by aggregating their respective tweets.
- Each document was tokenized with a Twitter-aware tokenizer.

3. Feature extraction

- N-gram language model:
  - Word n-grams with n in the range 1–3 were extracted.
  - Weighted using TF-IDF (term frequency–inverse document frequency).
  - Character level n-grams were not considered.
- Topic model:
  - Topic models identify hidden themes in a document.
  - LDA (Latent Dirichlet Allocation) was employed. This is a generative model in which documents are treated as a finite mixture of topics, such that each word in a document must be generated by one of its topics.
  - A topic model trained on input data was used to label every topic as present or not present within each document.

4. Architecture

- Two feature sets—n-grams and topics—were combined to train Support Vector Machines with a linear kernel from package scikit-learn.
- Resulting model was then presented with previous unseen documents; performing judgements on the author attributes it was trained with.

5. Results

<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>Spanish</th>
<th>Italian</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global Ranking</td>
<td>0.0743</td>
<td>0.6918</td>
<td>0.8061</td>
<td>0.0796</td>
</tr>
<tr>
<td>Average RMSE</td>
<td>0.1725</td>
<td>0.1619</td>
<td>0.1378</td>
<td>0.1409</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>Age</th>
<th>Joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>E</td>
<td>N/A</td>
</tr>
<tr>
<td>0.7394</td>
<td>0.5909</td>
<td>N/A</td>
</tr>
<tr>
<td>0.5211</td>
<td>0.5455</td>
<td>N/A</td>
</tr>
<tr>
<td>RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.1381</td>
<td>0.1669</td>
</tr>
<tr>
<td>N</td>
<td>0.2223</td>
<td>0.2285</td>
</tr>
<tr>
<td>A</td>
<td>0.1918</td>
<td>0.1398</td>
</tr>
<tr>
<td>C</td>
<td>0.1749</td>
<td>0.1412</td>
</tr>
<tr>
<td>O</td>
<td>0.1352</td>
<td>0.1329</td>
</tr>
</tbody>
</table>

Table 1: Results of final software submission including global rankings and individual attribute performance.

- Age, gender and their combination were scored using the accuracy metric for each of the four languages.
- Personality aspects (E=Extraversion, N=Neuroticism, A=Agreeableness, C=Conscientiousness and O=Openness) were scored with RMSE (root mean squared error). An average RMSE for each language is also provided.
- Global ranking is a combination of the joint (age, gender) accuracy and the average personality RMSE.
- These results show that n-grams and topic models are useful in developing multiple language compatible author profiling systems, as consistent results are achieved over the four languages.
- Manipulating hyperlinks was found to have no affect on system performance.
- Up- (and down-) weighting of users to avoid author over-fitting also had no affect on performance.

6. Further Work

- Attempt to generate more robust topic models by training on a large external corpus.
- Assess effect of additional stylometric features such as readability.
- Investigate network and behavioural features for author profiling on social media.

References


Acknowledgments

- This material is based upon work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under Award No. FA9550-14-1-0333.