Author Profiling Using Semantic and Syntactic Features

György Kovács§, Vanda Balogh†, Purvanshi Mehta‡, Kumar Shridhar¤, Pedro Alonso§, and Marcus Liwicki‡*

§Embedded Internet Systems Lab, Luleå University of Technology, Luleå, Sweden (gyorgy.kovacs@ltu.se, pedro.alonso@ltu.se, marcus.liwicki@ltu.se)
†Institute of Informatics, University of Szeged, Szeged, Hungary (bvanda@inf.u-szeged.hu)
¤MindGarage, Kaiserslautern, Germany, (purvanshi.mehta11@gmail.com, shridhra.stark@gmail.com)

Key contributions

- Examined a wide set of features for both tasks
- Examined a variety of machine learning algorithms for the bot detection class
- Attained 99% accuracy on validation and ~89% on the test set for bot detection

Task

- Bot detection and gender classification
- Classify twitter profiles based on tweets
- English language

Features

- **URL features**: average no. of URLs
- **Emoticons**: average emoticon count
- **Stylistic features**:
  - Character floodings
  - Average no. of capital letters per word
  - No. of sentences
  - No. of tokens
  - Flesch reading-ease score
  - Tokens repeated more than twice
  - Maximum repetition count of token
- **POS tags**: average no. of spaCy [1] POS tags per profile
- **Topic features**: Prevalence of words in the profile belonging to SEMCAT [2] and SemCor [3] categories (133 altogether)

Results

<table>
<thead>
<tr>
<th>Classifier</th>
<th>F1-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdaBoost</td>
<td>99</td>
</tr>
<tr>
<td>RandomForest</td>
<td>97</td>
</tr>
<tr>
<td>Bagging Classifier</td>
<td>97</td>
</tr>
<tr>
<td>Gradient Boost</td>
<td>98</td>
</tr>
<tr>
<td>SVMs</td>
<td>94</td>
</tr>
<tr>
<td>BiDirectional LSTM</td>
<td>83</td>
</tr>
</tbody>
</table>

PAN evaluation

<table>
<thead>
<tr>
<th>Task</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bot/Human</td>
<td>99%</td>
<td>89%</td>
</tr>
<tr>
<td>Female/Male</td>
<td>94%</td>
<td>36%</td>
</tr>
</tbody>
</table>

Conclusions and future work

- An efficient syntactic and semantic feature extractor is introduced
- Several types of features included in the examination
  - URL, emoticons, tokens, capital letters
  - Syntactic features extracted using POS tags
  - Semantic features extracted using the SEMCAT and SemCor dataset
- Future work
  - Analyze the use of languages in tweets
  - Examine the use of separate feature sets for the two tasks
  - Combination of topic modelling with emotions detected in tweets

References