Vote/Veto Classification, Ensemble Clustering and Sequence Classification for Author Identification

Roman Kern\textsuperscript{1,2}
Stefan Klampfl\textsuperscript{2}  Mario Zechner\textsuperscript{2}

\textsuperscript{1} Knowledge Management Institute - Graz University of Technology
\textsuperscript{2} Know-Center

rkern@tugraz.at
{rkern, sklampfl, mzechner}@know-center.at

PAN Workshop @ CLEF 2012 / 2012-09-20
Vote/Veto Classification

▶ Same as last year
   ⇒ Compare data-sets

▶ Three different feature-set sets
   ⇒ Compare influence of uni-grams features vs. stylometric features
Authorship Attribution - Classification

Classification Algorithm

- Combine feature-spaces via individual base classifiers
- Based on performance in training phase
- In classification phase combine results

Base Feature Spaces

- Basic statistics, token statistics, grammar statistics
- Stop-word terms, slang terms, pronoun terms
- Intro-outro terms, bigram terms, unigram terms, terms

Feature Space Combinations

- Terms
- Stylometric
- Statistics
# Authorship Attribution - Data-Sets

## Basic Statistics

<table>
<thead>
<tr>
<th></th>
<th>PAN 2011</th>
<th>PAN 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Paragraph to lines ratio</td>
<td>Number of characters</td>
</tr>
<tr>
<td>2</td>
<td>Text to lines ratio</td>
<td>Number of words</td>
</tr>
<tr>
<td>3</td>
<td>Number of lines</td>
<td>Number of lines</td>
</tr>
<tr>
<td>4</td>
<td>Empty lines ratio</td>
<td>Number of stop-words</td>
</tr>
<tr>
<td>5</td>
<td>Number of paragraphs</td>
<td>Number of tokens</td>
</tr>
</tbody>
</table>

## Token Statistics

<table>
<thead>
<tr>
<th></th>
<th>PAN 2011</th>
<th>PAN 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Likelihood of proper nouns</td>
<td>Number of tokens</td>
</tr>
<tr>
<td>2</td>
<td>Number of tokens</td>
<td>Likelihood of proper nouns</td>
</tr>
<tr>
<td>3</td>
<td>Average token length</td>
<td>Average verb length</td>
</tr>
<tr>
<td>4</td>
<td>Likelihood of infrequent word groups</td>
<td>Average token length</td>
</tr>
<tr>
<td>5</td>
<td>Likelihood of tokens of length 9</td>
<td>Likelihood of pronouns</td>
</tr>
</tbody>
</table>
Authorship Attribution - Feature Types

Comparison of configurations

![Bar chart showing comparison of feature types: Terms, Statistics, and Stylometric.](chart.png)
Authorship Clustering - Approach

Ensemble Clustering

- Multi-tier clustering
- Combine output of base clusters
- Only use stylometric features

Ensemble clustering is also known as consensus clustering or clustering aggregation
Multiple feature spaces

- Basic statistics (same as for authorship attribution)
- Stylometric features (hapax-legomena, hapax-dislegomena, yules-k, simpsons-d, brunets-w, sichels-s, honores-h, ...)
- Stem-suffixes, stop-words, pronouns
- Character 1-grams, 2-grams, 3-grams

⇒ Total of 7 feature spaces
Authorship Clustering - Clustering

Base clustering

▶ k-means clustering
▶ k-means++ seed selection
▶ Different relatedness measures for different feature spaces
  ▶ Cosine similarity
  ▶ Euclidean distance (after normalising the features)

Ensemble clustering

▶ Create a meta-space from the individual clustering solution
▶ In meta-space the distance between instances depends on the agreement of the clustering solutions
  ▶ Give different base clusters different weight
▶ k-means clustering
## Authorship Clustering - Evaluation

### Ensemble clustering results

<table>
<thead>
<tr>
<th>Feature Space</th>
<th>A vs B</th>
<th>C vs D</th>
<th>E vs F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-grams</td>
<td>51.52%</td>
<td>53.98%</td>
<td>61.87%</td>
</tr>
<tr>
<td>2-grams</td>
<td>50.91%</td>
<td>54.46%</td>
<td>56.70%</td>
</tr>
<tr>
<td>3-grams</td>
<td>50.91%</td>
<td>51.33%</td>
<td>52.37%</td>
</tr>
<tr>
<td>Stop-Words &amp; Pronouns</td>
<td>62.20%</td>
<td>50.72%</td>
<td>72.91%</td>
</tr>
<tr>
<td>Stem Suffices</td>
<td>65.85%</td>
<td>63.01%</td>
<td>54.61%</td>
</tr>
<tr>
<td>Stylometry</td>
<td>52.74%</td>
<td>59.76%</td>
<td>64.25%</td>
</tr>
<tr>
<td>Basic Statistics</td>
<td>57.01%</td>
<td>56.87%</td>
<td>65.22%</td>
</tr>
<tr>
<td><strong>Ensemble</strong></td>
<td><strong>66.10%</strong></td>
<td><strong>80.34%</strong></td>
<td><strong>78.44%</strong></td>
</tr>
</tbody>
</table>
Sexual Predator Identification - Approach

Sequence classification

- Not directly classify predators
- Classify individual messages/line in chats
- Simple features
Chat message classes/labels

- normal, predator; offending; reaction, post-offending

<table>
<thead>
<tr>
<th>Chat #1</th>
<th>Chat #2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 normal</td>
<td>1 normal</td>
</tr>
<tr>
<td>2 predator</td>
<td>2 predator</td>
</tr>
<tr>
<td>3 normal</td>
<td>3 normal</td>
</tr>
<tr>
<td>4 normal</td>
<td>4 normal</td>
</tr>
<tr>
<td>5 predator</td>
<td>5 offending</td>
</tr>
<tr>
<td>6 normal</td>
<td>6 reaction</td>
</tr>
<tr>
<td>7 predator</td>
<td>7 post-offending</td>
</tr>
<tr>
<td>8 predator</td>
<td>8 post-offending</td>
</tr>
<tr>
<td>9 normal</td>
<td>9 reaction</td>
</tr>
<tr>
<td></td>
<td>10 reaction</td>
</tr>
</tbody>
</table>
Sexual Predator Identification - Features

Simple features

- Unigrams
- Double Metaphone
- isInitialAuthor, isLastAuthor, isMostVerboseAuthor, isFewerAuthors, hasTermFromPrevious

Classification algorithm

- Maximum entropy & beam search
Conversation: cc21a4030e5be0428f302d96452a2fbd
pred: hi
pred: u found more pics
pred: can u do it tonight
pred: no
pred: do it now
pred: leave message my cell is dead
pred: 7077187918
pred: k
pred: ok
pred: <email/>
.pred: k

Conversation: 0a00d0980272a40af6e8475e88363f44
pred: i love u

Conversation: 168d5c53c212270d6cd51972abdc3511
pred: yea
pred: k
pred: huh
pred: when u call me i'll tell u what i'll be wearing
pred: wow
pred: top
pred: cool
pred: bring jacket
pred: when u leave are u going to miss me alot
pred: yees
pred: call me tommrow
pred: what time
pred: yea
pred: 10am
pred: when time she leave
pred: work tommrow
pred: sneak
pred: yea have to sneak too
pred: aways with u
pred: caall me at 10am
pred: ok
pred: what time do u want to call
pred: no i wont
pred: lol
pred: love sneaking
pred: does anyone know where u are going
## Sexual Predator Identification - Results

<table>
<thead>
<tr>
<th>Class</th>
<th>Count</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>normal</td>
<td>3,117</td>
<td>0.955</td>
<td>0.995</td>
</tr>
<tr>
<td>predator</td>
<td>29</td>
<td>0.3</td>
<td>0.103</td>
</tr>
<tr>
<td>offending</td>
<td>52</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>post-offending</td>
<td>216</td>
<td>0.871</td>
<td>0.847</td>
</tr>
<tr>
<td>reaction</td>
<td>275</td>
<td>0.959</td>
<td>0.764</td>
</tr>
<tr>
<td>Identify predators</td>
<td>2</td>
<td>0.667</td>
<td>1</td>
</tr>
</tbody>
</table>
Thank you!

Open-source code
https://www.knowminer.at/svn/opensource/projects/pan2012/trunk

Corresponding Author
Roman Kern <rkern@tugraz.at>