For the authorship attribution task we performed experiments with a document occurrence representation using a standard classification-based approach. Results obtained with this approach were mixed: in the small data sets distributional representations resulted very helpful, although in the large data sets a simple bag-of-words approach outperformed the document occurrence approach. For the authorship verification task we adopted a classification-based approach and proposed a modification to Ensemble Particle Swarm Model Selection (EPSMS) for selecting classification models for each task. This approach obtained acceptable performance in two out of the three data sets.

**Authorship attribution**

A neural network classifier was used under the one-vs-all formulation. Documents were represented under the document occurrence representation (DOR).

```
Fig. 1. OVA approach to authorship attribution
```

**The document occurrence representation**

DOR is a distributional term representation in which a document is represented by a distribution of occurrences over other documents in the same corpus. Intuitively, a document is represented by its context. Each term in the vocabulary is first represented as a distribution of occurrences over documents.

\[
\mathbf{w}^{\text{dor}}(t, \mathbf{d}) = df(t, \mathbf{d}) \times \log \left( \frac{V_t}{N_d} \right)
\]

Next, each document is then represented by a combination of the representations of terms that occur in the document.

\[
df(t, \mathbf{d}) = \begin{cases} 
1 + \log(\#(t, \mathbf{d})) & \text{if } \#(t, \mathbf{d}) > 0 \\
0 & \text{otherwise}
\end{cases}
\]

**Experimental results at PAN’11**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MA-P</th>
<th>MA-R</th>
<th>MA-F1</th>
<th>MI-P</th>
<th>MI-R</th>
<th>MI-F1</th>
<th>Sum-ranks</th>
<th>Overall rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small</td>
<td>0.676</td>
<td>0.381</td>
<td>0.387</td>
<td>0.709</td>
<td>0.709</td>
<td>0.709</td>
<td>19</td>
<td>3rd out of 19</td>
</tr>
<tr>
<td>Small+</td>
<td>0.65</td>
<td>0.201</td>
<td>0.193</td>
<td>0.578</td>
<td>0.578</td>
<td>0.578</td>
<td>16</td>
<td>2nd out of 16</td>
</tr>
<tr>
<td>Large</td>
<td>0.608</td>
<td>0.294</td>
<td>0.303</td>
<td>0.508</td>
<td>0.508</td>
<td>0.508</td>
<td>48</td>
<td>8th out of 48</td>
</tr>
<tr>
<td>Large+</td>
<td>0.53</td>
<td>0.203</td>
<td>0.191</td>
<td>0.446</td>
<td>0.446</td>
<td>0.446</td>
<td>29</td>
<td>9th out of 29</td>
</tr>
</tbody>
</table>

**Table 1.** Official results obtained with the proposed approach.

Results were very competitive, the entries were above the average performance among other participants. Results were particularly positive in the Small data sets and for the “+” Version.

**Does DOR really help?**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Accuracy</th>
<th>Macro-F1</th>
<th>Micro-F1</th>
<th>Sum-ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>DOR</td>
<td>79.91</td>
<td>67.68</td>
<td>0.387</td>
<td>0.418</td>
</tr>
<tr>
<td>BOW</td>
<td>70.08</td>
<td>67.68</td>
<td>0.387</td>
<td>0.418</td>
</tr>
<tr>
<td>DOR</td>
<td>79.91</td>
<td>67.68</td>
<td>0.387</td>
<td>0.418</td>
</tr>
<tr>
<td>BOW</td>
<td>70.08</td>
<td>67.68</td>
<td>0.387</td>
<td>0.418</td>
</tr>
</tbody>
</table>

**Table 2.** Comparison of the DOR and bag-of-words (BOW) representations.

We compared the performance of the DOR representation with a straight BOW formulation. DOR outperformed BOW for the Small data sets (accuracy). Although, BOW outperformed DOR in the Large datasets.

**Authorship verification**

We faced the verification problem as a classification task (docs. written by the author vs. docs. written by any other author). Documents were represented with the BOW formulation and EPSMS was used to select the classifier.

**Ensemble Particle Swarm Model Selection**

EPSMS is a method for the automatic selection of binary classification models. In a nutshell, EPSMS searches for the best ensemble method that can be generated by using the methods available in a machine learning toolbox.

In order to obtain stable predictions, the results obtained with five runs of EPSMS were combined to obtain the final prediction of the model.

**Experimental results at PAN’11**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>MA-P</th>
<th>MA-R</th>
<th>MA-F1</th>
<th>MI-P</th>
<th>MI-R</th>
<th>MI-F1</th>
<th>Sum-ranks</th>
<th>Overall rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verify-1</td>
<td>0.1</td>
<td>0.333</td>
<td>0.154</td>
<td>17</td>
<td>8th out of 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verify-2</td>
<td>0.4</td>
<td>0.8</td>
<td>0.533</td>
<td>11</td>
<td>1st out of 10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verify-3</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>30</td>
<td>9th out of 10</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Official verification results obtained with the proposed approach.

The results obtained with the proposed formulation are interesting and give evidence that the classification approach to AV can be very effective. We believe the proposed method has potential for this and other tasks, although we would like to conduct an extensive evaluation of it in order to detect what factors influence the performance of the proposed technique.

**Conclusions**

- For AA we found that the use of DOR was partially beneficial in a classification-based approach.
- We confirmed that the BOW is a strong baseline.
- In general, the classification-based approach to AA and AV is very competitive nowadays.
- In AV we found EPSMS is competitive although it still can further improved. We are studying was of tailoring EPSMS for the AA and AV tasks.

**References**