Authorship Verification with neural networks via stylometric feature concatenation

Authorship Verification. PAN at CLEF 2021
22 -23 September 2021

Name
Antonio Menta Garuz (amenta@invi.uned.es)
Ana Garcia Serrano (agarcia@lsi.uned.es)

NLP & IR Research Group (UNED), Spain
Our Approach

**Features:** Based on Stylometric feature extraction process.
- Character-level n-grams.
- Punctuation marks.

**Model input:**
- feature vector difference between texts as input for NN.

**Model:** Binary classifier based on neural networks.
- One NN for each feature.
- Concatenate output vector representation into a final NN for decision making.
Features

Objective:

Reduce high dimensionality of character-ngrams features vs punctuation marks.

- Dimension reduction techniques: PCA.  
- Solution: Obtain a latent vector for each feature.

Neural Network

Character-level n-grams:
  • 6 Fully Connected Layers
  • From dimension 45000 to 6

Punctuation marks:
  • 3 Fully Connected Layers
  • From dimension 32 to 6

Concatenate layers:
  • 3 Fully Connected Layers

Results

Results obtained at Authorship Verification shared task:

<table>
<thead>
<tr>
<th>Set</th>
<th>AUC</th>
<th>c@1</th>
<th>F1</th>
<th>F0.5u</th>
<th>Brier</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>0.9635</td>
<td>0.9024</td>
<td>0.8990</td>
<td>0.9186</td>
<td>0.9155</td>
<td>0.9198</td>
</tr>
<tr>
<td>Small</td>
<td>0.9385</td>
<td>0.8662</td>
<td>0.8620</td>
<td>0.8787</td>
<td>0.8762</td>
<td>0.8843</td>
</tr>
</tbody>
</table>

“Authorship Verification with neural networks via stylometric feature concatenation” (Menta and Garcia-Serrano, 2021)

CODE: https://github.com/Hisarlik/Authorship-verification
Conclusion

• Stylometric features can achieve competitive results.

• Neural Networks work well with both datasets. (small and large)

• The Importance of each feature can be modified by varying its output vector.

<table>
<thead>
<tr>
<th>Character-level n-grams</th>
<th>Punctuation marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.67 0.12 ... 0.14</td>
<td>2.67 0.12</td>
</tr>
</tbody>
</table>

Size : 6

Size : 6
Next Steps

• Increase the number and type of features used:
  0 Lexical features: Vocabulary richness
  0 Syntactic features: Part-of-speech, phrase structure
  0 Structural features: Average frequencies of word length, paragraph length,…

• Improve hyperparameters tuning of the neural network.
  0 Automated hyperparameter optimization methods
Authorship Verification with neural networks via stylometric feature concatenation

Authorship Verification. PAN at CLEF 2021
22 -23 September 2021

Name
Antonio Menta Garuz (amenta@invi.uned.es) NLP & IR Research Group (UNED), Spain
Ana Garcia Serrano (agarcia@lsi.uned.es) NLP & IR Research Group (UNED), Spain