Authorship Attribution in Greek Tweets Using Author’s Multilevel N-gram Profiles

George K. Mikros & Kostas Perifanos
National and Kapodistrian University of Athens, Greece
A brief typology of authorship research

- **Authorship attribution**: Closed problem. We assume that one of 1, 2, 3… n candidates is the real author of a text.

- **Author verification**: Open problem. We assume an open set of authors and each text should be attributed to its real author without reference to any corpus from other authors.

- **Author profiling**: Closed problem. We assume that specific extralinguistic characteristics (gender, age, psychological profile etc.) of the author(s) can be traced in his/her texts.
Aims of the present research

- To perform authorship attribution experiments in tweets written in Modern Greek.
  - Create the first Modern Greek Tweets Corpus (GTC) in order to use it as a reference corpus for studying social media language including authorship attribution, sentiment analysis and linguistic variation.

- Explore the effectiveness of a specific document representation called Author’s Multilevel N-gram Profile (AMNP), which comprises of a combined vector of increasing size and different level n-grams

- Investigate alternative ways to construct training sets for authorship attribution in Twitter data. More specifically:
  - Is it better to use single tweets for testing or do we need to merge tweets producing bigger text units?
  - In the case of merging tweets, what is the text size that produces the best attribution results?
# The Greek Twitter Corpus

<table>
<thead>
<tr>
<th>Authors</th>
<th>No of Tweets</th>
<th>Total size (words)</th>
<th>Average size (words)</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>500</td>
<td>5,378</td>
<td>10.75</td>
<td>5.42</td>
</tr>
<tr>
<td>B</td>
<td>918</td>
<td>10,515</td>
<td>11.45</td>
<td>5.52</td>
</tr>
<tr>
<td>C</td>
<td>2,065</td>
<td>32,098</td>
<td>15.54</td>
<td>6.73</td>
</tr>
<tr>
<td>D</td>
<td>455</td>
<td>7,451</td>
<td>16.57</td>
<td>5.48</td>
</tr>
<tr>
<td>E</td>
<td>1,347</td>
<td>9,822</td>
<td>7.29</td>
<td>5.01</td>
</tr>
<tr>
<td>F</td>
<td>535</td>
<td>3,692</td>
<td>6.90</td>
<td>4.93</td>
</tr>
<tr>
<td>G</td>
<td>1,277</td>
<td>9,412</td>
<td>7.37</td>
<td>5.63</td>
</tr>
<tr>
<td>H</td>
<td>2,306</td>
<td>26,212</td>
<td>11.36</td>
<td>5.86</td>
</tr>
<tr>
<td>I</td>
<td>2,986</td>
<td>18,720</td>
<td>6.26</td>
<td>4.28</td>
</tr>
<tr>
<td>J</td>
<td>584</td>
<td>7,618</td>
<td>13.06</td>
<td>6.74</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12,973</strong></td>
<td><strong>130,918</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Author’s Multilevel N-gram Profile - AMNP
Experimental methodology

- We used GTC to create 4 different datasets which contained merged tweets of increasing text sizes (25, 50, 75 and 100 words).
- As classification algorithm we used multiclass support vector classification (LIBLINEAR library).
- We tested the authorship attribution accuracy with each feature group separately and compared it with AMNP. Accuracy figures were calculated on two different conditions:
  - a) 10-fold cross-validation (cv) in the merged tweets text units
  - b) External test set which contained 500 single tweets not included in the training set (35-60 per author).
Accuracy vs. text size chunks using two different validation methods

- 10-fold cv
- External Dataset

Axis Title
Impact of text size and feature representation method in authorship attribution accuracy using cross-validation

Accuracy vs. Text Size

- AMNP
- Char2grams
- char3grams
- word2grams
- word3grams
Impact of text size and feature representation method in authorship attribution accuracy using external dataset

Text Size vs. Accuracy

- AMNP
- Char2grams
- char3grams
- word2grams
- word3grams
Conclusions

- Authorship attribution in tweets of Modern Greek is a feasible task. Our top performance (0.952 accuracy in 10-fold cv using 100-word text chunks) is a good indication that the tweet’s linguistic structure is a significant carrier of authorship information.

- AMNP representation is based on a solid linguistic – semiotic theoretical background and proved highly efficient compared to single n-gram feature groups in all text sizes.

- The obtained results indicated that optimal performance is achieved when both training and testing sets for authorship attribution contained merged tweets.
Thank you...