Automated detection of health websites' HONcode conformity: can N-gram tokenization replace stemming?

Célia Boyer¹, Ljiljana Dolamic¹, Natalia Grabar¹,*
¹Health on the Net Foundation, Geneva, Switzerland; ²Université Lille 3, France
celia.boyer@healthonnet.org

**HONCode – most widely utilized healthcare website Code of conduct**
- Parts of webpages (extracts) are stored in the database by the expert indicating the website’s compliance with one of the HONcode principles
- 8300 certified websites mostly in English, French, Spanish and German
- Large scale manual certification is not feasible

**Could it be done in an automated manner?**
- Quality health related information in their native language is crucial for end-users worldwide
- Would the automated system result in similar performance quality regardless the underlying language
- Linguistic treatment has proven to be a powerful tool, especially for morphologically complex languages
- Language independent approach required, we propose character n-gram

**Automated system for detection of HONcode conformity:**

- Stop-words removal (174)+
  - Word (W1, baseline)
  - Stem (W1p, porter)
  - n-gram (C3, C4, C5)
- Feature selection:
  - Document frequency (DF)
  - Chi-square (C2)
  - Z-score (ZS) – threshold limit 2

**HONcode principles**
- Authority
- Complementarity
- Privacy
- Attribution
- Justifiability
- Contact details
- Financial disclosure
- Advertising policy

**9 separate classifiers**
- Attribution (Reference+ Date)
- Privacy
- Justifiability
- Authority
- Complementarity
- Advertiser
- Financial disclosure
- Advertising policy

**Classification results**
- Precision (P), recall (R), F measure
- Best performance marked in bold
- Statistically significant differences marked by *

**Character n-gram**
- A viable alternative to stemming
- Might results in better performance than stemming for more complex languages
- Very good classification results for all HONcode criteria can be achieved by selectively determining the correct set of parameters
- “Correct” dimensionality reduction algorithm can improve the classification results

**Relative difference in precision between W1p (baseline) and C5:**
- from -7.25% (“Justifiability”) to 3.45% (“Reference”)
- “Justifiability” - hardest criteria to detect
- Small number of training documents
  - 872 for “Justifiability” vs. 2683 for “Privacy”
- Similar tendencies in precision and recall regarding dimensionality reduction with different percentages of features kept for W1, W1p and C5
- DF and Z-score significantly outperform the Chi-square

**5-gram tokenization, 10 fold average**

<table>
<thead>
<tr>
<th>Ctr.</th>
<th>Privacy</th>
<th>Attribution</th>
<th>Justifiability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Date</td>
<td>F1</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>R</td>
<td>F1</td>
</tr>
<tr>
<td>DF</td>
<td>0.89*</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>C2</td>
<td>0.90*</td>
<td>0.99</td>
<td>0.94</td>
</tr>
<tr>
<td>ZS</td>
<td>0.91</td>
<td>0.98</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Character n-gram**
- A viable alternative to stemming
- Might results in better performance than stemming for more complex languages
- Very good classification results for all HONcode criteria can be achieved by selectively determining the correct set of parameters
- “Correct” dimensionality reduction algorithm can improve the classification results