Can We Hide in the Web?
Age and Gender Author Profiling in Social Media

...most of this sugar comes from high fructose corn syrup which is the chief ingredient in chips, cereals or breads. And just because it is "all natural", it does not mean it's good for you. To the body, it's all sugar!

System Setup – DKPro Lab Framework

PAN Challenge Data
- from web environment (chats, blogs, fora)
- 226,600 documents in English
- 75,900 documents in Spanish
- 3 age groups (13-17, 23-27, 33-47 years)
- male and female authors
- over 200 mil. words

Classification
DKPro Text Classification & WEKA
- Multiclass classification for six classes (one-against-all approach)
- Logistic regression with ridge estimates (WEKA)
- Information Gain filter

Extracting Features
DKPro Text Classification
- Surface: Long/short words, words per sentence, number of hyperlinks, number of smileys, type-token ratio, text length...
- Readability: Flesch, Kincaid, Coleman-Leu, SMOG, FOG, LIX
- Content: Emotion words (e.g. anger), topic words (e.g. school)
- Syntax: POS ratios, Contextuality measure, plurals, modals
- Punctuation: Inner punctuation, questions, exclamations
- Lexical: Emotional endings (e.g. –ous, –ly…)

Age and gender differences are related

Style becomes more "male" with age - we get more descriptive while showing less emotions. For relevant features, such as frequency of smileys, the difference between adult men and women (upper plot) is similar to the one between teenage and adult men (lower plot).

Content-based features outperform style

Features based on word lists (mainly teenage slang and emotions) contributed to the overall performance more than stylistic features. However, they were more successful in determining age than gender.

<table>
<thead>
<tr>
<th>Word list</th>
<th>Words</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teenage words</td>
<td>113</td>
<td>Bro, geez, brb, lol</td>
</tr>
<tr>
<td>Person words</td>
<td>134</td>
<td>Relative, team-mate, friend</td>
</tr>
<tr>
<td>Work words</td>
<td>287</td>
<td>Employee, bonus, recruiter</td>
</tr>
<tr>
<td>Positive words</td>
<td>297</td>
<td>Cheerful, unusual, joyful</td>
</tr>
<tr>
<td>Negative words</td>
<td>521</td>
<td>Meanable, scared, stressed</td>
</tr>
</tbody>
</table>

Performance (classifier accuracy)

<table>
<thead>
<tr>
<th>Subsystem</th>
<th>EN</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maj class baseline</td>
<td>0.17</td>
<td>0.17</td>
</tr>
<tr>
<td>Human evaluation</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>Surface features only</td>
<td>0.20</td>
<td>0.21</td>
</tr>
<tr>
<td>Syntactic &amp; punct features</td>
<td>0.23</td>
<td>0.30</td>
</tr>
<tr>
<td>Content &amp; lexical features</td>
<td>0.27</td>
<td>0.33</td>
</tr>
<tr>
<td>Syntax, punct, &amp; content &amp; lex.</td>
<td>0.29</td>
<td>0.38</td>
</tr>
<tr>
<td>All features combined</td>
<td>0.29</td>
<td>0.38</td>
</tr>
</tbody>
</table>

User study
- 20 participants,
- 20 random texts from the PAN challenge,
- Age accuracy 0.5: random decisions
- Human prediction based on stereotypes, fails on neutral topics

Conclusions - Gender classification (a = .62)

Men
- use longer words, more articles and hyperlinks, and talk more often about computers.
- Women
- use more emotional words, smileys, exclamations and "love" words

Conclusions - Age classification (a = .55)

Older authors
- write less readable longer posts, use longer words, commas, links, talk more about work and god.
- Teenagers
- use more pronouns and smileys, less nouns and articles, speak with more emotional words, neologisms and slang, talk more about people and computers and often violate the spelling rules.

References
Lightweight Framework for Reproducible Parameter Sweeping in Information Retrieval

Can We Hide in the Web? Large Scale Simultaneous Age and Gender Author Profiling in Social Media - Notebook for PAN at CLEF 2013

© DIPF 2013