CLiPS Stylometry Investigation (CSI) corpus

A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text

Ben Verhoeven & Walter Daelemans
{Ben.Verhoeven;Walter.Daelemans}@uantwerpen.be

Presented at LREC 2014
Reykjavik, Iceland
Introduction

• Stylometry
  - Quantitative study of text style in order to learn about its author

• Computational stylometry
  - Authorship identification/verification
  - Author profiling: prediction of age, gender, personality, region, mother tongue, ...
Few existing resources
(especially for Dutch)

Issues
- Authorial profile can be hard to get
- Not all freely available
  - Non-disclosure agreements
  - Anonymization problems
- None have more than 2 kinds of meta-data
Why do we want all meta-data?

- All aspects have an influence on the author’s writing style
- More importantly: these aspects are reflected in the same kind of features
  - E.g. pronouns (Pennebaker, 2011)
- Solutions:
  - control for some aspects
  - balance the data
  - take all aspects into account
CSI Corpus Summary

- Corpus in two genres: essays and reviews
- Large amount of meta-data
- Multitude of purposes
  - Mostly in computational stylometry
- Freely available
- Yearly expansion
Corpus source

- Students in Linguistics & Literature
- Native speakers of Dutch

- Collect their essays, papers and make them write reviews
- Now in the third year, presenting the statistics of the past two years
Author Meta-Data

- Age
- Gender: male/female
- Sexual orientation*: straight or LGBT
- Region of origin: Belgian provinces or The Netherlands
- Personality profile: Big Five and MBTI*

* Provided optionally
Personality Profile

Big Five
- **Openness to experience**
- **Conscientiousness**
- **Extraversion**
- **Agreeableness**
- **Neuroticity**

Score 0-100 per trait

MBTI (Myers-Briggs Type Indicator)
- Extravert – Introvert
- Thinking – Feeling
- Sensing – iNtuition
- Judging – Perceiving

Dichotomy with score 0-100
Genre

- Essays, papers: written for Dutch proficiency course (university level), formal text
- Reviews: special assignment
Reviews

• Two reviews per person
  - Truthful: reflects true opinion of author on existing product
  - Deceptive: about a fictional product

• Balanced for sentiment: positive/negative

• Topic (categories) available
  - Food chains, movies, musicians, smartphones, books

Universiteit Antwerpen
## Corpus Statistics

- **Document statistics per genre**

<table>
<thead>
<tr>
<th>Genres</th>
<th># docs</th>
<th># tokens</th>
<th>Avg. length</th>
<th>Std.dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reviews</td>
<td>540</td>
<td>69,132</td>
<td>128</td>
<td>74</td>
</tr>
<tr>
<td>Essays</td>
<td>209</td>
<td>235,400</td>
<td>1126</td>
<td>757</td>
</tr>
<tr>
<td>Total</td>
<td>749</td>
<td>304,532</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Corpus Statistics

- Distribution of reviews over types

<table>
<thead>
<tr>
<th>Type</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth</td>
<td>136</td>
<td>134</td>
</tr>
<tr>
<td>Deception</td>
<td>119</td>
<td>151</td>
</tr>
</tbody>
</table>

- More or less balanced
Corpus Statistics

- **Author age**

<table>
<thead>
<tr>
<th>Average</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.5</td>
<td>18</td>
<td>47</td>
<td>2.87</td>
</tr>
</tbody>
</table>

- **Number of documents per author**

<table>
<thead>
<tr>
<th>Average</th>
<th>Minimum</th>
<th>Maximum</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.25</td>
<td>1</td>
<td>9</td>
<td>0.88</td>
</tr>
</tbody>
</table>
**Corpus Statistics**

**Region of author**

- **Antwerpen**: 228
- **Limburg**: 18
- **Oost-Vlaanderen**: 13
- **Vlaams-Brabant**: 18
- **West-Vlaanderen**: 30
- **The Netherlands**: 32
- **Other**: 6

**Gender and sexual orientation**

- **Straight**: 146
- **LGBT**: 8
- **Unknown**: 25

**Universiteit Antwerpen**
Corpus Statistics

- Average Big Five personality

<table>
<thead>
<tr>
<th>Openness</th>
<th>Conscientiousness</th>
<th>Extraversion</th>
<th>Agreeableness</th>
<th>Neuroticity</th>
</tr>
</thead>
<tbody>
<tr>
<td>50.7</td>
<td>45.2</td>
<td>49.8</td>
<td>41.6</td>
<td>54.7</td>
</tr>
</tbody>
</table>

-> Rather balanced

(slightly neurotic students?)

Universiteit Antwerpen
License

- Creative Commons Attribution-NonCommercial-ShareAlike 3.0

- www.clips.uantwerpen.be/datasets
Case Study: Deception Detection

- Task of automatically classifying text as being either truthful or deceptive, in our case by examining writing style of author
- Product reviews ≈ deceptive opinion spam

“Fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader” (Ott et al, 2011)
Related Research

- Text classification problem
  - Truth vs. deception
- Frequently used features:
  - Token unigrams
  - LIWC lexicon words
- No computational research for Dutch so far
Experimental Setup

• Supervised ML (LibSVM), 10-fold cv
• Features:
  - only token unigrams (threshold 5)
  - without domain-specific words (such as product names)
• 3 experiments:
  - All data
  - Negative reviews
  - Positive reviews

Universiteit Antwerpen
<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
<th>Baseline</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Data</td>
<td>72.2</td>
<td>72.2</td>
<td>72.2</td>
<td>72.2</td>
<td>50.0</td>
</tr>
<tr>
<td>Positive</td>
<td>69.7</td>
<td>69.7</td>
<td>69.3</td>
<td>69.3</td>
<td>53.3</td>
</tr>
<tr>
<td>Negative</td>
<td>71.5</td>
<td>71.4</td>
<td>71.4</td>
<td>71.4</td>
<td>53.0</td>
</tr>
</tbody>
</table>
Result discussion

- Comparable to state-of-the-art results of Mihalcea & Strapparava (2009) for English opinion texts (~70%)

- Ott et al. (2011) achieve higher performances (up to 89%), but truthful and deceptive instances come from different sources: TripAdvisor and Amazon Mechanical Turk
Summary

Advantages
- Multiple purposes
- Yearly expansion
- Text from similar sources (within each genre)
- Enables cross-genre experiments

Disadvantages
- Opportunistic nature (restricted to authors at hand) influences balance of meta-data
Future Work

Planned additions
- Third genre: bachelor dissertations
- More meta-data, e.g. grades for papers and dissertations
  -> enables automatic grading

Anyone interested in building a similar corpus for their language?

Universiteit Antwerpen
Thank you!

For suggestions and/or questions:

**Ben Verhoeven & Walter Daelemans**
CLiPS – Computational Linguistics Group
University of Antwerp

{Ben.Verhoeven;Walter.Daelemans}@uantwerpen.be