CLiPS Stylometry Investigation (CSI) Corpus

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

Ben Verhoeven and Walter Daelemans

CLiPS - Computational Linguistics Group, University of Antwerp, Belgium

CLiPS Computational Linguistics & Psycholinguistics University of Antwerp

{ben.verhoeven;walter.daelemans}@uantwerpen.be
www.clips.uantwerpen.be

Summary

CSI is a freely available Dutch corpus designed to serve a multitude of purposes, mostly in computational stylometry. The corpus provides textual data in two genres with large amounts of meta-data and will be expanded on a yearly basis.
Successful experiments on the detection of deception already illustrate its usefulness.

Author Info

Students taking Dutch proficiency courses at University of Antwerp

Available meta-data:
- Age
- Gender
- Region of origin
- Personality scores
    - Big Five
    - MBTI *
- Sexual orientation *

* Provided optionally

Document Info

Two genres:

Essays
- Written for Dutch proficiency course
- Rather formal text

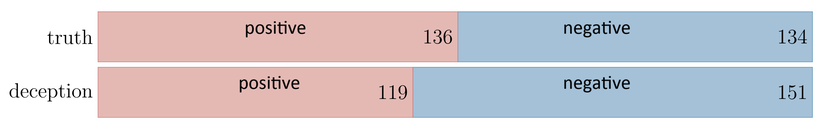
Reviews
- Special assignment
- Two reviews per person
    - Truthful: real opinion on real product
    - Deceptive: fictional product
- Corpus balanced for sentiment
    - Positive
    - Negative
- Topics available

Corpus Statistics

1. Document statistics per genre

Genres	# docs	# tokens	Avg. length	Std.dev.
Reviews	540	69,132	128	74
Essays	209	235,400	1126	757
Total	749	304,532		

2. Distribution of reviews over types



3. Age of authors

Average	Minimum	Maximum	Std.Dev.
20.5	18	47	2.87

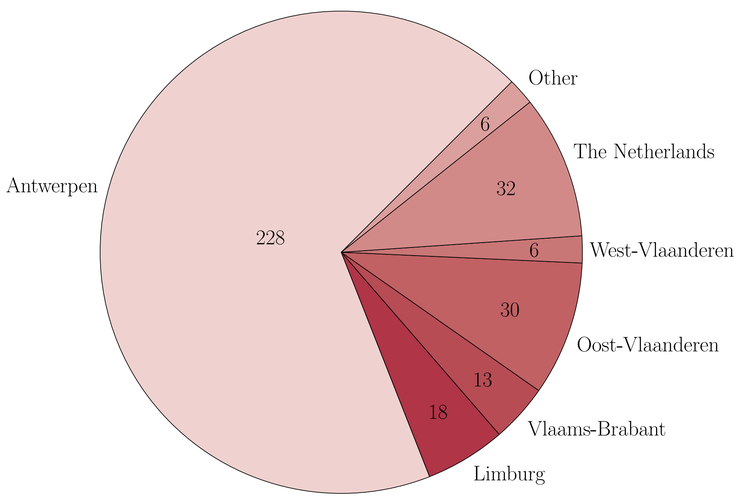
4. Number of documents per author

Average	Minimum	Maximum	Std.Dev.
2.25	1	9	0.88

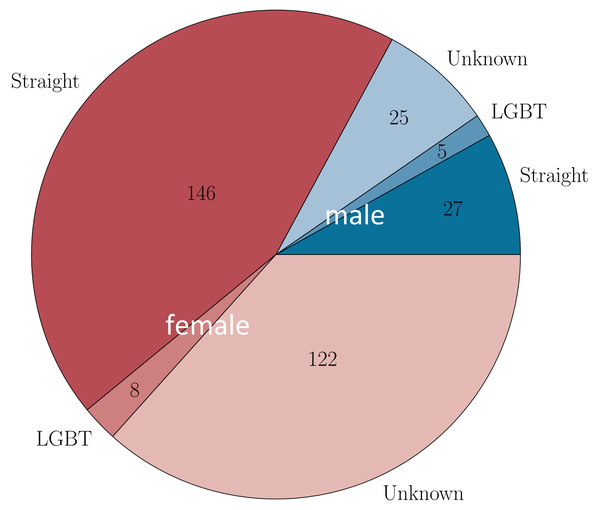
5. Average Big Five personality profile of the authors in the corpus

Openness	Conscientiousness	Extraversion	Agreeableness	Neuroticity
50.7	45.2	49.8	41.6	54.7

6. Distribution of origin of author (Dutch-speaking Belgian provinces, The Netherlands, or other)



7. Distribution of author gender and sexual orientation (LGBT = Lesbian, Gay, Bisexual or Transgender)



Case Study: Deception Detection

- Classifying text as truthful or deceptive by examining writing style of author
- Related to spam detection, false reviews are deceptive opinion spam:

"fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader" (Ott *et al.*, 2011)

Setup
- Supervised ML, 10-fold cv
- LibSVM (Chang & Lin, 2011)
- Features:
    - Token unigrams of training data
    - Threshold: 5
    - Filter out domain-specific words
- Three experiments:
    - All data
    - Negative reviews
    - Positive reviews

Case Study: Results

	Accuracy	Precision	Recall	F-Score	Baseline
All Data	72.2	72.2	72.2	72.2	50.0
Positive	69.7	69.7	69.3	69.3	53.3
Negative	71.5	71.4	71.4	71.4	53.0

Results for different classifiers on deception detection, the baseline is the majority class frequency

- 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 these are contested results because positive and negative examples come from different sources: TripAdvisor and Amazon Mechanical Turk

Discussion

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

Disadvantages
- Opportunistic nature influences balance of meta-data
- Not all meta-data available for all authors

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

This research was funded by the Flemish government agency for Innovation by Science and Technology (IWT) through Strategic Basic Research (SBO) project AMiCA.

@clipsua