CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text

Ben Verhoeven, Walter Daelemans

CLiPS - Computational Linguistics Group
University of Antwerp, Belgium
{ben.verhoeven;walter.daelemans}@uantwerpen.be

Abstract

We present the CLiPS Stylometry Investigation (CSI) corpus, a new Dutch corpus containing reviews and essays written by university students. It is designed to serve multiple purposes: detection of age, gender, authorship, personality, sentiment, deception, topic and genre. Another major advantage is its planned yearly expansion with each year's new students. The corpus currently contains about 305,000 tokens spread over 749 documents. The average review length is 128 tokens; the average essay length is 1126 tokens. The corpus will be made available on the CLiPS website (www.clips.uantwerpen.be/datasets) and can freely be used for academic research purposes.

An initial deception detection experiment was performed on this data. Deception detection is the task of automatically classifying a text as being either truthful or deceptive, in our case by examining the writing style of the author. This task has never been investigated for Dutch before. We performed a supervised machine learning experiment using the SVM algorithm in a 10-fold cross-validation setup. The only features were the token unigrams present in the training data. Using this simple method, we reached a state-of-the-art F-score of 72.2%.

Keywords: computational stylometry, text classification, deception detection

1. Introduction

Research in computational stylometry has always been constrained by the limited availability of training data, since collecting textual data with the appropriate meta-data requires a large effort. For every text, the characteristics of the author have to be known. At the moment, there exist a number of Dutch corpora for the detection of age, gender (Peersman et al., 2011; Nguyen et al., 2013), authorship and personality (Luyckx and Daelemans, 2008). Yet, not all of these corpora are freely available (e.g. because of non-disclosure agreements and anonymization problems) and none of these corpora contain information on all relevant characteristics. Other issues may arise when different classification systems are used for some of these characteristics, e.g. MBTI (Briggs Myers and Myers, 1980) vs. Big Five (Goldberg, 1990) in personality detection. The situation is similar for other languages. Although more corpora exist for English, most of them are not available for other researchers (Celli et al., 2013).

Having large amounts of data remains the key to reliable results in computational stylometry. In this paper, we present the CLiPS Stylometry Investigation (CSI) corpus, a freely available Dutch corpus that can be used for stylometry research and many other applications

2. Corpus Description

The CSI corpus contains essays and reviews written by Linguistics & Literature students taking Dutch proficiency courses (for native speakers) at the University of Antwerp. Since there are new students every year, we have the opportunity to continue collecting data over several years. One of the major advantages of the corpus is its yearly expansion. The current corpus contains data from the past two years (2012-2013). In order to avoid confusion over the characteristics and statistics of the expanding corpus, overviews of each version with the corresponding meta-data are available on the corpus website[1].

The entire corpus has been anonymized and all authors have explicitly given us permission to include their submissions and profile information in a corpus for research purposes.

2.1. Characteristics

A unique aspect of our corpus is the breadth of the meta-data. There is meta-data available on both the authors and the documents included in the corpus.

2.1.1. Author Meta-Data

For each author, we have information on age, gender, region of origin and personality scores on the Big Five scale. The authors can optionally also provide extra meta-data: their sexual orientation, and personality scores on the MBTI scale.

Age Authors have provided us with their birth date. Given the timestamp of a document, we can compute the age of the author at the moment of writing.

Gender Authors were asked to indicate their gender, choosing 'male' or 'female'.

Region of origion The region of origin is defined as the region where the author grew up in, i.e. where the author lived between ages 2 and 10. Default regions are the Dutch-speaking Belgian provinces (Antwerpen, Limburg, Vlaams-Brabant, West-Vlaanderen, Oost-Vlaanderen) and The Netherlands. When the default options do not apply, authors are able to select their country from a list.

[1] http://www.clips.uantwerpen.be/datasets/csi-corpus/

Personality We used two systems of personality measurement. All students (from the year 2013 onward) were required to take an online Big Five personality test[2] (Goldberg, 1990). This personality test provides a score (0-100) on five traits: openness to experience (OPN), conscienctiousness (CON), extraversion (EXT), agreeableness (AGR), and neuroticity (NEU).
Optionally, students could also complete an online MBTI (Myers-Brigss Type Indicator) personality test[3] (Briggs Myers and Myers, 1980). The MBTI test provides scores (0-100) on four dichotomies: Extraversion-Introversion, Thinking-Feeling, Sensing-iNtuition and Judging-Perceiving.

Sexual orientation Authors can optionally specify their sexual orientation by selecting 'straight' or 'LGBT'[4]. The label is 'Unknown' when this information is not available.

This information can be used for a number of interesting experiments. For example, we can investigate the influence of someone's sexual orientation on the detection of stylometric features. Having both Big Five and MBTI personality scores allows us to compute the relation between these personality frameworks.

2.1.2. Document Meta-Data

We have so far mainly discussed the author characteristics of the corpus. Here we describe the kind of documents we have at our disposal.
The corpus contains documents of two genres: essays/papers and reviews. The essays are rather formal texts written by our students as assignments for their Dutch proficiency course. In their first year, they write a shorter text, here called 'essay'. In their second year, they write a longer text, here called 'paper'.
The reviews are a special assignment for the students. Participants in the review collection did not know the purpose of the review writing. Everyone has to write two reviews, a truthful and a deceptive one. The reviews are balanced for sentiment (negative and positive), which also makes this corpus an interesting dataset for sentiment detection. Deception is implemented here by asking the author to write a convincing review (either positive or negative) about a fictional product, thus pretending to know about the product while actually making up the review. The truthful reviews reflect the author's real opinion on an existing product. Truthful and deceptive reviews are written about products from the same five categories: smartphones, musicians, food chains, books, and movies. The category and product of a review are included in the metadata.
Since we have both truthful and deceptive texts of the same author, we can compare the writing style in these two circumstances with more authority than previous research using texts from different sources (e.g. comparing real reviews with reviews collected through Amazon Mechanical Turk (Ott et al., 2011)).

[2]http://www.outofservice.com/bigfive/
[3]http://www.humanmetrics.com/cgi-win/jtypes2.asp
[4]Lesbian, gay, bisexual or transgender

2.2. Statistics

Statistics of this corpus are by definition temporary due to its yearly expansion; you will find the statistics for the 2013 version in tables 1 to 5 and figures 1 to 3.

Because we took advantage of the data and author availability at our university, some characteristics of the authors may be under- or overrepresented.

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

Table 1: Document statistics per genre (length is in words including punctuation, also known as 'tokens').

Projecting these statistics about corpus size to the future returns an expected corpus size of about 1200 reviews and 550 essays in three years, depending on the number of students enrolling in these courses. The size of the future corpus in tokens is estimated to be at least 620,000 for the essays and 120,000 for the reviews.

	Positive	Negative	Total
Truth	136	134	270
Deception	119	151	270
Total	255	285	540

Table 2: Distribution of reviews over types

Table 2 shows us that there is a (more or less) balanced distribution of sentiment and veracity in our reviews.

The distribution of the topics of the reviews over their veracity is, however, slightly skewed for the topics 'musicians' and 'books' (Figure 1).

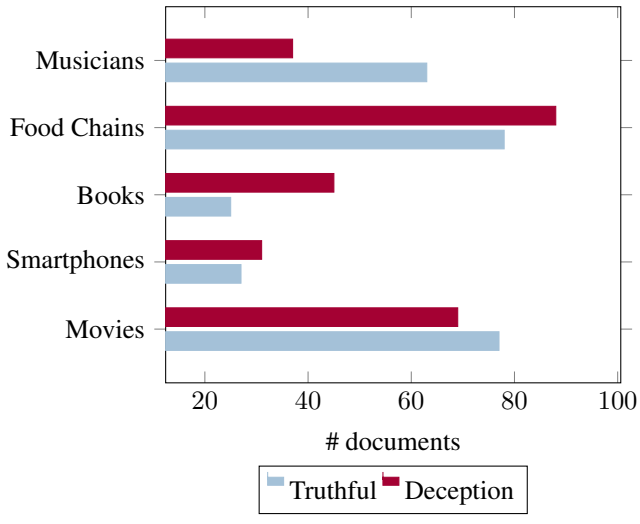

Figure 1: Distribution of review topics over veracity

Average	Minimum	Maximum	Std.Dev.
2.25	1	9	0.88

Table 3: Number of documents per author

In Table 3 we find that there are multiple documents per author in our corpus. This allows our corpus to be used for authorship verification experiments, where the task is to verify whether a certain document is written by the same author as a given document. In fact, an adapted version of our corpus will be used for the PAN 2014 shared task on authorship verification[5].

Average	Minimum	Maximum	Std.Dev.
20.5	18	47	2.87

Table 4: Age of authors.

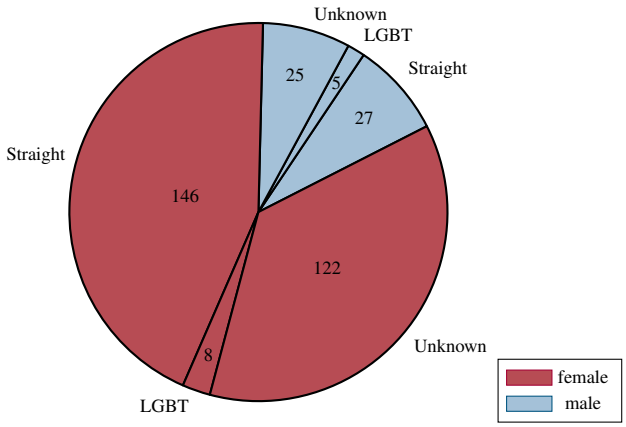

Figure 2: Distribution of author gender and sexual orientation

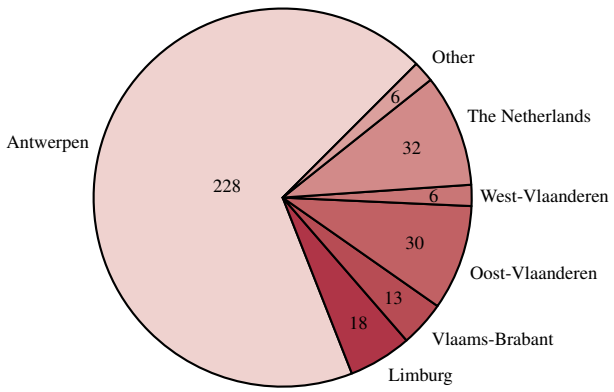

Figure 3: Distribution of region of origin of author

[5]http://pan.webis.de

Openness	50.7
Conscientiousness	45.2
Extraversion	49.8
Agreeableness	41.6
Neuroticity	54.7

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

Given these statistics, a typical author in our corpus is a 20 year old woman that grew up in the Belgian province Antwerpen.

2.3. License

The CSI corpus is licensed under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported (CC BY-NC-SA 3.0)[6]. This means you are free to use, copy, share and adapt the corpus for non-commercial purposes under the following restrictions: give credit to the author, indicate changes made, and distribute derivations of the material under the same license. Any resulting publications should cite this paper.

3. Case Study: Deception Detection

To illustrate the usefulness of this corpus, we have performed a basic experiment on the detection of deception in Dutch reviews.

Deception detection is the task of automatically classifying a text as being either truthful or deceptive, in our case by examining the writing style of the author. The deceptive texts at hand, product reviews, can be considered deceptive opinion spam: 'fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader' (Ott et al., 2011). The detection task we are performing is thus opinion spam detection or fake review detection, which is a more specific variant of deception detection.

Deception detection (in the framework of computational linguistics) is usually conceived as a text classification problem where our system should classify an unseen document as either truthful or deceptive. Such a system is first trained on known instances of deception. Frequently used features are token unigrams and LIWC lexicon words.

Although there has been one paper using Dutch data for research on deception (Schelleman-Offermans and Merckelbach, 2010), this was a psychological experiment with analysis of the participants' writings, focusing on the connection between deception and fantasy proneness. Therefore, our case study is the first experiment on deception detection for Dutch, to our knowledge.

For a more thorough background on deception detection, see Ott et al. (2011) and Zhou et al. (2004) and references therein.

3.1. Setup

A supervised machine learning experiment using tenfold cross-validation with the SVM algorithm from the LibSVM package (Chang and Lin, 2011) was set up with as only

[6]http://creativecommons.org/licenses/by-nc-sa/3.0/

features the token unigrams present in the training data. A frequency threshold of 5 was imposed on these unigrams because infrequent unigrams do not appear in enough documents to contribute to the learning process. The unigrams were also cleared of domain-specific words, i.e. we removed the names of the real and fictional products since they would show a one-to-one relationship with their category. The thresholding approach makes sure that misspellings of these product names are also disregarded.
We investigated deception in this data in three different ways, using tenfold cross-validation.

- A classifier was trained using all the reviews.
- A classifier was trained using the negative reviews.
- A classifier was trained using the positive reviews.

This allows us to compare with previous research only investigating deception on single-sentiment data (Ott et al., 2011).

3.2. Results

We present the results of our three experiments in table 6. We provided a majority baseline for comparison. This baseline indicates the performance of a system that would classify all instances as belonging to the most frequent class.

	Acc.	Prec.	Rec.	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

Table 6: Results for different classifiers on deception detection

With these features, the classical approach of using all the data and building one binary classifier seems to be the most succesful one.
When taking a closer look at the 100 most important features (with highest X^2), we notice that about 90% of those are functors (function words) and punctuation. This is an indication that our system uses stylistic features as a basis for its decision. In order to check whether the somewhat skewed distribution of topics over the veracity has an influence on our results, we tested our system on each topic separately. No significant differences in performance were found.
Our results are comparable with the state-of-the-art results of Mihalcea and Strapparava (2009) for English opinion texts. They also achieved a performance of around 70%. Although Ott et al. (2011) achieve even higher performances (up to 89%), their results are somewhat contested because their positive and negative training examples come from different sources (truthful reviews from TripAdvisor and deceptive reviews collected through Amazon Mechanical Turk); they may thus be performing 'platform recognition' instead of deception detection. This suspicion is strengthened by Mukherjee et al. (2013) who report a widely different word distribution between those fake and true reviews.

4. Conclusion and Future Work

In this paper, we have presented a new text corpus for stylometric research and we have demonstrated its usefulness by performing promising experiments on the automatic detection of deceptive text.
This corpus has many advantages: it serves multiple purposes (detection of age, gender, authorship, personality, sentiment, deception and genre); it will be expanded yearly; and all texts come from similar sources (within their genre) for optimal comparability. Some disadvantages of the corpus are: its opportunistic nature (we are restricted to the authors at hand) which influences the balance of some of the meta-data; and that not all meta-data is available for all authors.
In the nearby future, we will integrate more (meta-)data into this corpus. A number of our students also write bachelor dissertations in Dutch; these will be included in the corpus as a third genre with (much) longer texts than the other data. For the essays and dissertations, grades were given by the professors that are an indication whether they are well-written or badly written texts. We will incorporate the grades for these texts as meta-data in our corpus to add another purpose to our corpus, namely automatic grading.

5. Acknowledgements

We would like to express our gratitude to Katrien Verreyken, Shanti Verellen, Sarah Van Hoof, Dominiek Sandra and Reinhild Vandekerckhove (University of Antwerp) as well as our former master students Mario De Groof, Jimmy Michiels and Suzanne Mpouli for their help in collecting the data.
The research described in this paper was performed in the context of the AMiCA project, which is funded by the Flemish Agency for Innovation through Science and Technology (IWT).

6. References

Isabel Briggs Myers and Peter B. Myers. 1980. *Gifts differing: Understanding personality type*. Davies-Black Publishing, Mountain View, CA.

Fabio Celli, Fabio Pianesi, Michael Stillwell, and David Kosinski. 2013. Workshop on computational personality recognition (shared task). In *Proceedings of the Workshop on Computational Personality Recognition (Shared Task)*, Boston, MA, 07/2013.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Kim Luyckx and Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.*, Marrakech, Morocco. European Language Resources Association.

Rada Mihalcea and Carlo Strapparava. 2009. The lie detector: explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009*

Conference Short Papers, pages 309–312, Stroudsburg, PA, USA. Association for Computational Linguistics.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013. Fake review detection: Classification and analysis of real and pseudo reviews. Technical Report UIC-CS-2013-03, University of Illinois at Chicago.

Dong Nguyen, Ilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "How old do you think I am?": A study of language and age in Twitter. In *Proceedings of International Conference on Weblogs and Social Media (ICWSM)*. Association for the Advancement of Artificial Intelligence.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computaional Linguistics: Human Language Technologies*, volume 1, pages 309–319, Stroudsburg, PA, USA. Association for Computational Linguistics.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC2011)*, Glasgow, UK. ACM Digital Library.

Karen Schelleman-Offermans and Harald Merckelbach. 2010. Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, 7:247–260.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. 2004. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group Decision and Negotiation*, 13(1):81–106.