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

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. 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. 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. A major advantage of the corpus is its planned yearly expansion with each year's new students. By 2016, the corpus will have tripled in size. The corpus is available on the CLiPS website (www.clips.uantwerpen.be/datasets) and can freely be used for academic research purposes.

For this corpus creation, we took advantage of the availability of data at our university. A lot of text is written during a student's school career, which is usually left unused. We collected these essays and papers and gathered metadata from the students. Authors provided us with their birth date, gender, region of origin, and personality information (Big Five). They could optionally specify their sexual orientation and do a second personality test (MBTI). The rather formal essay genre was supplemented with a specific review writing assignment. Reviews were balanced for sentiment (positive/negative) and veracity (truthful/deceptive) and dealt with products from five categories: smartphones, books, movies, musicians and food chains.

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%.