university of groningen / faculty of arts

A lexico-syntactic approach to deception detection on Dutch reviews

Ivo Taris

Bachelor thesis
Information Science
Ivo Taris
s2188724
December 12, 2017

ABSTRACT

As political elections and civic discourse are manipulated through 'fake news', and companies publicly face untruthful feedback, identifying deceptive writing has become an important task. Although deception is often not identifiable by humans, classification systems show the ability to detect subtle differences between truthful and deceptive writings. As the amount of research on deception detection in Dutch is neglectible, this research focused on Dutch reviews from the CLiPS Stylometry Investigation (CSI) corpus. The results demonstrate that classifying the reviews as truthful or deceptive can be improved from 72.2% to 79.3% by incorporating both text-based lexical cues and syntactic patterns as features, and by filtering for stopwords.

CONTENTS

PREFACE

I would like to express my gratitude to Barbara Plank for providing overall guidance in this research and offering comprehensive feedback. In addition, I would like to thank Ben Verhoeven and Walter Daelemans from the University of Antwerp for creating and publicizing the CLiPS Stylometry Investigation (CSI) corpus. Both enabled the formation of this research greatly.

1 INTRODUCTION

In a time where political elections and civic discourse are manipulated through 'fake news' on social media, and where companies publicly face untruthful feedback (Ott et al., 2012), identifying deceptive writings has become an important task. The internet has given rise to large-scale computer-mediated communication, part of which entails written discourse or the publishing of written statements. Such communication omits al non-verbal cues, making a recipient solely reliant upon the text in judging the truthfulness of the content. With no special aids or training, a statistical meta-analytic research shows that humans can discriminate lies from truths in only 54% of the cases (Bond and DePaulo, 2006).

On the other hand, a growing body of research suggests that automatic deception detection by applying machine-learning is more effective than humans (Hauch et al., 2015). Thus, automatic detection of deceptive writings using classification systems emerges as a prominent tool to counter potential adverse developments.

Online reviews are widely consulted in making purchasing decisions and can exert significant effects to the object of the review. The impact of negative reviews can even exceed that of positive reviews (Chang and Wu, 2014; D. Hollebeek and Chen, 2014). Due to the use of social media, information can reach a large number of readers rapidly (Pfeffer et al., 2014). Untruthful negative information can therefore have detrimental consequences for organizations and individuals.

Text-based deception can come in many forms. With respect to detecting textual deception, previous research has mainly focused on opinion spam in reviews. This research will focus on deceptive opinion spam in reviews as defined by Ott et al. (2011): 'fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader'. These deceptive opinions are typically not detectable by a human reader. Nevertheless, cues to deception could still be present in the text whereas deception 'conceals, fabricates, or distorts information' and 'involves additional mental effort' (Frank et al., 2008). Using only the text as the object of analysis requires careful examination of both and lexical and syntactic aspects of the respective text.

Most research has focused on lexical, syntactic, and behavioral features in English. The amount of research that focuses on deception detection in Dutch is neglectible. The first effort that focuses on Dutch deception detection is a use-case for a newly developed corpus, and therefore uses only word unigrams in classifying reviews as truthful or deceptive (Verhoeven and Daelemans, 2014). Hence, this paper will investigate whether the classification of Dutch product and service reviews as truthful or deceptive can be improved by focusing on both lexical and syntactic patterns.

The remainder of this paper starts with reviewing relevant literature and providing a theoretical context for the following research steps. The subsequent chapter provides a description of the used dataset and its source. Following this a detailed overview of the research methodology is given. Finally, this report concludes by presenting and discussing the research findings, elaborating on the implications, and reflecting on the limitations.

2 BACKGROUND

Online deception has been studied in various contexts. As this research revolves around written product and service reviews, the specific characteristics of these reviews need to be taken into account. This is important in order to get a grasp on how deception manifests itself this setting, and subsequently how it can be detected. Research on text-based deception detection can be roughly be divided among early quantitative studies and more recent machine learning approaches. Both provide the necessary theoretical foundation for the current investigation.

2.1 REVIEWS

When consumers consider the purchase of a product or service through e-commerce, an information asymmetry between buyers and sellers can arise. Consumers face the difficulty of making the distinction between high- and low-quality products and services, and have to determine the trustworthiness of sellers. The Nobel Prize winner Akerlof (1978) identifies the former difficulty as "one of the more important aspects of uncertainty" in prospective purchases. Oliveira et al. (2017) determined that the latter difficulty is an important factor in online purchasing intentions.

To overcome these issues consumers increasingly rely on third-party information sources such as online reviews when making purchase decisions (Kwark et al., 2014). Reviews are a form of word-of-mouth (WOM) communication, which can be defined as "informal communications between private parties concerning evaluations of goods and services" (Anderson, 1998) and is "one of the strongest influencers on adoption of new products and services" (Meuter et al., 2013). More specifically, research has shown that consumers are very susceptible to online consumer reviews (Chen and Xie, 2008) and are affected in their behaviour by it (Hennig-Thurau et al., 2003). Online reviews are characterized as being one-to-many communication (Litvin et al., 2008) and typically one-way information flows (Schindler & Bickart, 2005).

Consumers value customer-generated reviews over information provided by manufacturers, retailers, or service-providers (Smith et al., 2005) because they "deem customer-to-customer information as more reliable and less susceptible to commercial motives" (Chen et al., 2016). Nevertheless, depending on consumer-to-consumer information does not mitigate the consideration of source trustworthiness, whereas the authors of online reviews are mostly anonymous or unknown to the reader. This notion is corroborated by Cheung et al. (2008), which found that source trustworthiness has a statistically insignificant impact on the perceived usefulness of information due to its anonymous nature.

The anonymous nature of online reviews gives rise to the possibility of deception. Deception is previously defined as "a message knowingly transmitted by a sender to foster a false belief or conclusion by the receiver" (Buller and Burgoon, 1996). As mentioned earlier, deceptive online reviews are hardly detectable by a human reader (Bond and DePaulo, 2006). This

is in part due to the lack of physical and social cues. Yet there remain cues to deception embedded in the text, whereas deception "conceals, fabricates, or distorts information" and "involves additional mental effort" (Frank et al., 2008). This research will therefore perform an in-depth linguistic analysis of reviews to determine the usefulness of text-based cues to deception. The scope is limited to these cues because of the anonymous nature of online reviews, which makes the classification model independent of author identity and behavior.

2.2 QUANTITATIVE STUDIES ON TEXT-BASED DECEPTION DETECTION

By studying written or transcribed verbal statements, Newman et al. (2003) determined that there is a qualitative difference in the linguistic style of communication between truth-tellers and those who distort the truth. Deceptive communications exhibit less linguistic markers that signify cognitive complexity, show lesser self- and other-references, and increasingly use of words that reflect a negative emotion compared to thruthful communications (Newman et al., 2003).

Zhou et al. (2004) expatiates on the topic of linguistic-based cues to deception in the context of a discussion facilitated by an asynchronous electronic messaging platform. The respective research found that textual cues can be detected by performing a systematic analysis of linguistic information. However, they express their concern that "cue profiles are unlikely to apply uniformly across contexts".

Zhou and Sung (2008) subsequently studied Chinese text-based synchronous computer-mediated communications extracted from a chat in a popular Chinese game. Four main classes of cues to deception are dinsinguished:

1. Quantity: word and message counts per player
2. Syntactic and lexical complexity: represented by the average sentence length and the average word length respectively
3. Diversity: measured by lexical diversity, content word diversity, and redundancy as ratios of specific unique words to the total number of words.
4. Non-immediacy: first- and third-person pronoun counts

Only the classes 1, 2, and 3 are found to be statistically different between truthful and deceptive communications.

In contrast to conversations or sending messages directly from a sender to a receiver, Yoo and Gretzel (2009) performed a similar statistical analysis on online reviews. The research focused on the same preceding four classes of cues to deception. It used 42 deceptive hotel reviews written by students and 40 truthful travel reviews extracted from Tripadvisor. Contrary to Zhou and Sung (2008), Yoo and Gretzel (2009) found that lexical complexity is significantly different and that the quantity of lexical elements is not significantly different between truthful and deceptive reviews. More specifically, deceptive reviews showed an increased lexical complexity. Also brand name mentioning, first person pronouns use, and positive sentiment

are indicative for deception. Hence, the research states that deception detection requires a different approach in the context of online reviews. Due to its asynchronous nature, Yoo and Gretzel (2009) states that "writers can deliberate on their writings and model their writing after reviews posted by others".

The preceding papers concentrate mainly on very specific linguistic markers of deception. The focus is put on word frequencies, word and sentence lengths, occurrences of unique words, and grammatical person references. These features might be too review- or source-specific to be generalizable for a robust classification system. This research will therefore use certain lexical features and extend these with complex syntactic patterns.

2.3 MACHINE LEARNING APPROACHES TO TEXT-BASED DECEPTION DETECTION

Jindal and Liu (2008) was the first to study the trustworthiness of opinions in reviews together with the application of a classification system using text-based features. The research states that opinion spam—incorrect or irrelevant information—occurs widely online. Using 5.8 million reviews and 2.14 million reviewers from amazon.com, the aim is to classify product reviews as spam or non-spam. Opinion spam can occur in three forms:

1. Untruthful by deliberately misleading readers
2. Brand-oriented instead of focusing on the product
3. Non-reviews by conveying only irrelevant information

The researchers manually labelled 470 type 2 and type 3 spam reviews. Identifying type 1 was considered to be extremely difficult and the research concludes "other ways have to be explored in order to find training examples for detecting possible type 1 spam reviews". The features entailed three types, namely review-centric features, reviewer-centric features, and product-centric features. Specifically the review-centric features are interesting whereas these include textual features such as the ratio of positive and negative opinion-bearing words in a review and brand name mentioning. Applying a logistic regression learning model effectuated an average AUC (Area Under ROC Curve) score of 98.7%. However, it must be noted that this high score is due to using unsophisticated reviews which easily give away the class.

Ott et al. (2011) elaborates on the previously discussed type 1 of opinion spam: deceptive opinions. Ott et al. (2011) defines these as 'fictitious opinions that have been deliberately written to sound authentic'. The research addresses this type specifically because they are almost unidentifiable by humans. By using a pool of 400 Human Intelligence Tasks they effectuated a crowdsourced dataset containing 400 truthful and 400 deceptive reviews. This avoids the difficult task of manual annotation. Three automated approaches employed are:

1. Genre identification using the frequency distribution of part-of-speech (POS) tags
2. Psycholinguistic deception detection

3. Text-categorization using the Linguistic Inquiry and Word Count (LIWC) and n-grams

The latter approach is of interest to this research. The text-categorization focuses on word unigrams, bigrams, and trigrams. Combining bigrams with LIWC achieved an accuracy score of 89.8%. The preceding two papers are interesting whereas they clearly explicate the difficulty of identifying untruthful reviews which are written to deliberately mislead readers. In addition they demonstrate that focusing on specific informative words and n-grams increases the discriminative power.

Mukherjee et al. (2013a) states however that crowdsourced reviews are not real deceptive reviews for the very fact that they intentionally simulated. Mukherjee et al. (2013a) recreated the experiment of Ott et al. (2011) and effectuated similar results on crowdsourced data. However, applying this approach to real reviews from a commercial environment resulted in an accuracy score of only 67.6%. This score still indicates that n-gram features are useful, but it might imply that using crowdsourced reviews interferes with the generalization of the classification system. Mukherjee et al. (2013a) continues by developing behavioral features, which among others include the time-window of activity, the number of reviews written by an author, and review length. Performing the classification using a Support Vector Machine classifier increases the accuracy score above 80% on a set of cross-domain real reviews.

Mukherjee et al. (2013b) support these findings in a follow-up research. Behavioral features like the number of reviews, review length, content similarity, and the percentage of positive reviews achieved an accuracy score of 86% using a Support Vector Machine classifier. These behavioral features outperformed text-based features such as word n-grams and part-of-speech tags. Nevertheless, word n-grams are only one of many text-based linguistic cues that can be exploited in order to detect deception. It is interesting that Mukherjee et al. (2013a) and Mukherjee et al. (2013b) are actually one of the first in the context of automated deception detection in reviews to critically review the generalizability of previous papers' work, and to express doubts with respect to the usefulness of text-based features. This paper will therefore explore text-based linguistic features more extensively and critically assess the trade-off between the hard to identify nature of deceptive reviews and the extent to which they are real.

Most literature has focused on English reviews. Verhoeven and Daelemans (2014) is the first to develop a classification system for Dutch reviews. However, the scope of the respective research is confined to token unigrams as single feature. This leaves deception detection in Dutch relatively unexplored. As deception detection in different languages can differ significantly due to inherent structural differences, this research is the first to investigate the influence of lexical cues and syntactic patters on deception detection in Dutch.

3 DATA AND PROCESSING

The objective of this research is to develop a supervised classification system to classify written product and service reviews as being either truthful or deceptive. Henceforth, two discrete labels are distinguished:

1. Truthful
2. Deceptive

3.1 DATA

The created classification system[1] is trained and tested using the reviews from the CLiPS Stylometry Investigation (CSI) corpus (Verhoeven and Daelemans, 2014), a Dutch corpus developed by the University of Antwerp. The corpus is updated on an annual basis and currently contains 1298 reviews and 517 essays written by 661 students. Whereas only the reviews are annotated as being truthful or deceptive, this research will focus only on the 1298 reviews written by 618 authors.

Each student has written at least one truthful and one deceptive review with respect to a product or service. The students were allowed to write the deceptive and truthful reviews about different topics. The corpus is therefore not controlled for a single topic. Table 1 presents the general categories of the reviews and their respective counts. These numbers indicate that the corpus is not balanced with respect to topic.

Table 1: Distribution of topic of reviews

Topic	# reviews
Books	175
Food/drink companies	433
Movies	321
Muscicians	197
Smartphones	172
Total	1298

The corpus is evenly balanced with respect to the classification labels, containing 649 truthful and 649 deceptive reviews. The corpus is also balanced with respect to sentiment, which is shown in table 2. The reviews together count 202827 tokens, with an average length of 156 tokens per review and a standard deviation of 65 tokens.

Table 3 shows the distribution of the authors' gender and sexual orientation. The corpus is strongly skewed towards the female gender and a straight sexual orientation. Table 4 shows the origin of the authors. These concern the five Dutch speaking provinces of Belgium, the Netherlands, or other Dutch regions. As this corpus is developed by the University of

[1] The documented code can be found on https://github.com/imtaris/bscThesis2017

Table 2: Distribution of type of review and sentiment

	Positive	Negative	Total
Truth	323	326	649
Deception	319	330	649
Total	642	656	1298

Antwerp, this explains why approximately 88% of the authors originate from Belgium, and mainly from Antwerp.

Table 3: Distribution of gender and sexual-orientation of authors

Gender	Straight	LGBT	Unknown	Total
Male	85	10	32	127
Female	336	10	145	491
Total	421	20	177	618

Table 4: Distribution of origin of authors

Region	# authors	% authors
Antwerpen	404	65.37
Limburg	37	5.99
Vlaams-Brabant	21	3.40
Oost-Vlaanderen	57	9.22
West-Vlaanderen	28	4.53
The Netherlands	62	10.03
Other	9	1.46
Total	618	100

The authors' year of birth lie within the range of 1964 and 1997. However, the distribution of the authors within this range is heavily skewed to the 90's, as shown in table 5. Most authors are recent and current students of the University of Antwerp, which explains the skewness.

3.2 PROCESSING

The corpus contains a preproccesed version of the original reviews. All the product names have been replaced with a *propname* tag. This version will be used to improve the robustness of the classifier with respect to cross-domain applications.

Tokenization will focus on seperating punctuation from words while maintaining hyphens between compound words, between prefixes and words, and in hyphenated phrases. Furthermore, apostrophes are maintained as they occur in Dutch spelling. This preserves occurrences of words in which letters are omitted "m'n (mijn)", two contracted words "zo'n (zo een)", and derivatives of acronyms "sms'en". All tokens will subsequently be transformed to lowercase in order to reduce sparsity. This effects the number of once occurring tokens due to variations in capitalization.

The content of a review can both contain high and low informative words. High informative words are those which are strongly biased towards one of the classification labels, and low informative words are occur often in both (Perkins, 2010). This notion can be extended to features as well.

Table 5: Distribution of date of birth of authors

year of birth	# authors
1964	1
1965	1
1969	1
1970	1
1978	1
1982	2
1984	1
1985	1
1986	1
1987	4
1988	18
1989	17
1990	53
1991	73
1992	73
1993	91
1994	113
1995	86
1996	75
1997	5
Total	618

Infrequent words that "do not appear in enough documents to contribute to the learning process" Verhoeven and Daelemans (2014). The respective research imposed a frequency threshold of 5 on the word unigrams. Therefore a bag of words will be created that contains unigrams with a frequency of at least 5.

Frequent words in a natural language, henceforth stopwords, are considered as low informative words whereas they provide relatively low predictive power. The stopwords are drawn from a list of 101 words provided by the Natural Language Toolkit (NLTK). High information words can be identified by using the Pearson's chi-square statistic (χ^2). This measure sums the differences between observed (O) and expected values (E) in all squares of the table, scaled by the magnitude of the expected values (Manning et al., 1999), as follows:

$$\chi^2 = \sum_{i,j} \frac{O_{ij} - E_{ij}{}^2}{E_{ij}}$$

where i ranges over rows of the table and j ranges over columns. This is implemented using the 'most_informative_words()' function from featx.py Perkins (2010).

The features' elements will be filtered for stopwords and selected based on high χ^2 scores. The variation in performance of applying these methods will be tested accordingly.

4 METHOD

As mentioned earlier, Verhoeven and Daelemans (2014) developed a classification system for the CSI corpus focusing on token unigrams using a Support Vector Machine. The accuracy score of 72.2% effectuated by Verhoeven and Daelemans (2014) forms the external baseline of this research. An elementary modification of this system would be to include word bigrams. Hence, if word unigrams and word bigrams together outperform the setup of Verhoeven and Daelemans (2014), then this will function as the internal baseline of this research. The following subsections explain which features will be developed to improve the internal baseline, the qualifying learning models, and the employed evaluation metrics.

4.1 FEATURES

The total set of features comprises 11 lexical and syntactic features. These are created for each review, which is the unit of analysis. The lexical features focus only on lexical n-grams, the syntactic features concern part-of-speech tags and the dependency relations of each sentence. The features are listed in table 6 on page 13, which assigns a unique reference id to each feature to avoid ambiguity. Furthermore the table displays each feature's blueprint and provides concrete comma-seperated examples extracted from the previous example sentence and Alpino parse output displayed in figure 2. The following subsections will discuss the motivation of these features and how they are created.

4.1.1 Alpino

In order to automatically obtain syntactic information, this research used Alpino (Bouma et al., 2001), a depency parser for the Dutch language. The respective parser is created in the context of the project Algorithms for Linguistic Processing and is a NWO PIONIER research project. Alpino has many parameters that can be set to allow for customization. For example, using the web-application of alpino Bouma et al. (2011) the sentence "dit restaurant is een echte aanrader"[1] results in a dependency tree visualized in figure 1.

This research focused on the 'end_hook' parameter which specifies the course of action with respect to a specific parse. Instead of focusing on the entire dependency structure of every sentence, the parse can be narrowed down to a subset of the resulting dependency data. The respective parameter has been set to 'triples_with_frames', which produces an output as shown in figure 2. Each line contains six fields separated by the "|" character:

1. Head word of the dependency relation
2. POS tag of the head word

[1] This can be translated to English as "This restaurant is highly recommended".

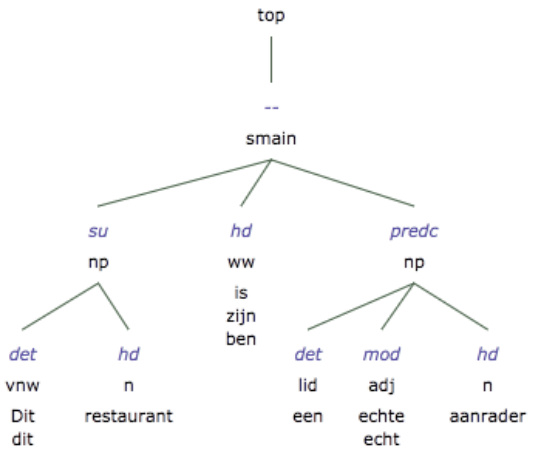

Figure 1: A dependency tree produced by the Alpino parser

3. Dependency name
4. Dependent word in the dependency relation
5. POS tag of the dependent word
6. Sentence number

The words are representend as "Stem/[Start,End]" where "Stem" is the base or root form of the word, obtained through the process of reducing morphological variations due to different grammatical categories. "Start" and "End" are the string positions of the respective words (Bouma et al., 2001).

```
aanrader/[5,6]|noun|hd/det|een/[3,4]|det|1
aanrader/[5,6]|noun|hd/mod|echt/[4,5]|adj|1
ben/[2,3]|verb|-- / --|./[6,7]|punct|1
ben/[2,3]|verb|hd/predc|aanrader/[5,6]|noun|1
ben/[2,3]|verb|hd/su|restaurant/[1,2]|noun|1
restaurant/[1,2]|noun|hd/det|dit/[0,1]|det|1
top/top|top|top/hd|ben/[2,3]|verb|1
```

Figure 2: Alpino 'triples_with_frames' parse output

4.1.2 N-grams

The lexical features focus only on word and stem n-grams and contribute two more features to the setup of (Verhoeven and Daelemans, 2014). Together the set entails word (F1) and stem (F2) unigrams and word bigrams (F3). N-grams are all contiguous sequences of N elements in a text. This enables an examination of the lexical diversity and the quantity of these lexical elements. The parameter 'triples_with_frames' of the Alpino parser yields stems as unit of analysis. Using the stem of words allows to investigate word occurences and lexical diversity regardless of their morphological variations. This can be useful because although words occur in numerous different forms due to different grammatical contexts, these do not imply an increased lexical diversity. A stem can therefore be regarded as being more robust when investigating lexical diversity between multiple texts.

4.1.3 Part-of-speech tags

Part-of-speech tagging is performed automatically by the Alpino parser. It is the process of word-category disambiguation by assigning descriptive tags to lexical elements of a text corresponding to a particular part of speech. Part-of-speech tags encode information with regard to the grammatical structure of sentences. As these might differ between truthful and deceptive writings, part-of-speech tags are operationalized as features by including them as unigrams (F4) and in combination with dependency relations.

4.1.4 Dependency relations

Dependency grammars encode the syntactic structure of sentences by identifying binary asymmetrical relations between its lexical elements. The syntactic structure concerns, among others things, the order of lexical elements in a sentence. An analysis of dependency relations therefore can involve all preceding feature elements. The dependencies itself can be used as unigrams (F5). Furthermore, the binary asymmetrical relation between both stems (F6) and part-of-speech tags (F7) can be used as features.

The latter two features can be futher broken down into other features. This means taking only the head stem (F8) or POS tag (F9) in combination with the relation, or taking only the relation and its dependent stem (F8) or POS tag (F9). To allow for less specific instances of the preceding two features, the dependency relation can be substituted for a general placeholder in the form of a specific character (F10 and F11). For instance, this could imply using an underscore as placeholder.

4.2 CLASSIFICATION AND EVALUATION

Two supervised learning models are considered to be the most suitable for the binary classification task of this research. The respective algorithms are:

1. Support Vector Machine
2. Logistic Regression

Verhoeven and Daelemans (2014) attained the score of 72.2% by using a Support Vector Machine from the LibSVM package. A Support Vector Machine is a supervised learning model that attempts to identify a border between the datapoints that reflects the best seperation of the two categories. Subsequently, it seeks to optimally fit a hyperplane based on the data points at the edge of each class and evaluates each new data point against this plane. Logistic Regression is also widely used. Logistic regression measuring the probabilistic relation between the features and the two categories by using a logistic function.

The difference in performance between the two learning models will be tested by means of a 10-fold cross-validation performed 20 times. Additionally, the performance of filtering for infrequent unigrams and high informative unigrams, and the imposition of a frequency thresfold of 5 will be included in this test. Performing a 10-fold cross-validation 20 times effectuates more consistent results by approaching the central tendency. With respect to the features this test comprises only word unigrams and word

bigrams—the internal baseline. The mean of the accuracy and F1 scores together with the matching standard deviation will serve as decisive motivator for either of the learning models, whether to use a filter for high informative features, and if a word frequency threshold should be imposed.

In order to prevent under- and overfitting of the features on the resulting learning model, 10-fold cross validation is applied 20 times while testing the features. Feature selection is effected by firstly testing the inclusion of all features, subsequently inluding only all stem related features, next all part-of-speech related features, and finally all depedency related features. Following this test, specific combinations of features will be made out of which the best performing combinations will be presented. The perfomance measurements focus on recall, precision, F1-score, and accuracy scores.

Table 6: The list of features used in this research

ID	Name	Blueprint	Example
F1	word unigram	word1, word2, ..	dit, restaurant, is, een, echte, ..
F2	word bigram	word1 word2, word2 word3, ..	dit restaurant, restaurant is, ..
F3	stem unigram	stem1, stem2, ..	dit, restaurant, ben, een, ..
F4	POS tag unigram	POS tag1, POS tag2, ..	noun, det, ..
F5	dependency relation unigram	dependency1, dependency2, ..	hd/det, hd/mod, ..
F6	stem dependency relation triple	head-stem dependency dependent-stem, ..	aanrader hd/det een, ..
F7	POS tag dependency relation triple	head-POS dependency dependent-POS, ..	noun hd/det det, ..
F8	stem dependency relation double	head-stem dependency, dependency dependent-stem, ..	aanrader hd/det, hd/det een, ..
F9	POS tag dependency relation double	head-POS dependency, dependency dependent-POS, ..	noun hd/det, hd/det det, ..
F10	stem dependency relation placeholder double	head-stem placeholder, placeholder dependent-stem, ..	aanrader _, _een, ..
F11	POS tag dependency relation placeholder double	head-POS placeholder, placeholder dependent-POS, ..	noun _, _ det, ..

5 RESULTS AND DISCUSSION

Based on the performance test of the Logistic Regression and Support Vector Machine classifiers, one of them will be used for executing the classification task. This performance test also establishes which filters will be applied throughout the classification task. Subsequently, each individual feature will be added to the internal baseline to determine its contribution. Finally the best performing feature combinations will be presented. The highest scores are emphasized by a bold typeface.

5.1 LEARNING MODELS

Table 7 displays the results of the performance comparison of the Logistic Regression (LR) and the Support Vector Machine (SVM) learning models without a word frequency threshold. The results demonstrate that Logistic Regression outperforms the Support Vector Machine learning model on all accuracy and F1 scores. Additionally, filtering for stopwords effectuates the highest performance. The χ^2 filter performs poorly, effectuating near chance accuracy scores. Overall it can be stated that the standard deviation of the accuracy scores is low.

Table 7: Performance comparison of the LR and SVM learning models without a frequency threshold

	No filter			Stopwords filter			χ^2 filter		
	A	A_σ	F1	A	A_σ	F1	A	A_σ	F1
LR	77.1	0.1	76.6	**79.2**	0.6	**78.7**	57.2	1.2	54.4
SVM	73.7	0.4	73.0	74.5	0.4	73.3	57.8	1.7	53.4

Table 8 displays the results of the performance comparison of the Logistic Regression (LR) and the Support Vector Machine (SVM) learning models with a word frequency threshold of 5. The results are consistent with table 7 in that Logistic Regression realizes higher accuracy and F1 scores. However, filtering for stopwords does not result in the highest performance. Furthermore it is observed that imposing a word frequency threshold of 5 on the unigrams lowers the overall performance.

Table 8: Performance comparison of the LR and SVM learning models with a frequency threshold of 5

	No filter			Stopwords filter			χ^2 filter		
	A	A_σ	F1	A	A_σ	F1	A	A_σ	F1
LR	**74.5**	0.5	**73.2**	72.7	0.5	69.4	54.8	1.8	49.4
SVM	69.5	0.6	66.6	65.2	0.4	54.0	54.6	1.3	45.7

The currently established baseline scores are shown in table 9. As opposed to the external baseline set by the research of Verhoeven and Daelemans (2014), a reasonable improvement has already been achieved by the internal baseline.

Table 9: Research baselines

Baseline	Learning Model	Filter	A	R	P	F1
Internal	LR	Stopwords	79.2	78.0	80.0	78.7
External	SVM	Frequency$_{\geqslant 5}$	72.2	72.2	72.2	72.2

5.2 FEATURE PERFORMANCE

Table 10 displays the results of the inclusion of all features (F1-F11) and the addition of each individual feature (F3-F11) to the internal baseline setup. The inclusion of all features results in an accuracy score of 76.7% when filtering for stopwords. Hence, using all features underperforms with respect to the internal baseline. When adding individual features to the internal baseline, all but one feature unperform. Only by adding dependency relation unigrams (F5) increases the accuracy to 79.3%. Additionally, the recall and F1 score are increased as well and precision is slightly lower.

Table 10: Performance of adding all features and individual features to the internal baseline

Setup	No filter				Stopwords filter			
Internal baseline	A	R	P	F1	A	R	P	F1
+ All features	76.1	75.1	76.4	75.8	76.7	76.3	77.0	76.5
+ F3	76.8	76.0	77.4	76.5	78.0	77.6	78.3	77.8
+ F4	76.3	74.5	77.4	75.7	78.7	79.0	78.6	78.7
+ F5	76.8	76.0	77.4	76.5	**79.3**	**79.3**	**79.5**	**79.2**
+ F6	76.2	74.8	77.1	75.8	78.7	78.2	79.2	78.5
+ F7	76.3	74.5	77.4	75.8	78.4	77.9	78.8	78.2
+ F8	76.6	75.1	77.5	76.2	77.5	77.3	77.8	77.4
+ F9	75.8	74.4	76.6	75.3	76.8	76.1	77.2	76.5
+ F10	**77.0**	**76.2**	**77.6**	**76.7**	77.3	77.1	77.5	77.2
+ F11	76.4	74.7	77.4	75.9	78.1	77.9	78.2	77.9

The preceding performance scores provide a starting base for combining features into a robust classification system. Table 11 presents the performance scores of combining several features. Firstly, features are grouped based on corresponding aspects. This implies the creation of four feature groups based on being related to POS tags (F4, F7, F9, and F11), stems (F3, F6, F8, and F10), dependency relations (F4-F11), and unigrams (F1-F5). All four groups are inferior in performance compared to the internal baseline.

After devising numerous possible combinations of features the best performing combination is the addition of POS tag unigrams and dependency relation unigrams, which realized an average accuracy score of 78.5%.

Table 11: Performance of feature combinations

Setup	No filter				Stopwords filter			
Internal baseline	A	R	P	F1	A	R	P	F1
+ F4, F7, F9, F11	75.5	74.1	76.3	75.0	76.4	75.4	77.0	76.1
+ F3, F6, F8, F10	**76.8**	**76.2**	**77.2**	**76.5**	77.2	76.9	77.5	77.0
+ F4-F11	759	75.1	76.5	75.6	76.5	76.0	76.9	76.3
+ F1-F5	76.5	75.6	77.1	76.2	77.3	76.8	77.6	77.0
+ F4, F5	76.2	75.3	76.8	75.9	**78.5**	**79.0**	**78.2**	**78.5**

5.3 DISCUSSION

Although Verhoeven and Daelemans (2014) and other relevant previous studies (Ott et al., 2011; Mukherjee et al., 2013a,b) predominantly use Support Vector Classifiers, this study has found Logistic Regression to be more effective.

The χ^2 measure was to a limited extent indicative of feature performance. Using this measure it was determined that POS tag related features (F4, F7, F9, and F11) offered no high informative items at all. Overall, using the χ^2 measure attains near-chance performance scores and was therefore excluded from subsequent feature testing. As the CSI corpus contains cross-domain reviews, it might be possible that selecting high informative items creates very review-specific and therefore non-generalizable features. On the other hand, filtering for stopwords generally improves performance scores. It therefore seems that removing low informative words improves deception detection.

With respect to the setup using the stopwords filter, adding all POS tag related features (F4, F7, F9, and F11) resulted in the lowest accuracy of 76.5%. This is followed by 76.7% by adding all features related to dependency relations (F4-F11) and by 76.7% by adding all features. This does not imply that POS tags and depedency relations inherently detiorate the internal baseline setup. On the contrary, adding both POS tag unigrams (F4) and dependency relation unigrams (F5) results in an accuracy score of 78.5%, which approaches the internal baseline. Appending only dependency relation unigrams (F5) to the internal baseline setup even attains the highest accuracy score of 79.3%. It appears that an increasing specificity of a feature decreases the discriminative power of the classification system. Although all features attain a decent accuracy score, n-gram features (F1-F5) both lexical and syntactic outperform more complex features (F6-F11). The discovered usefulness of word unigrams and bigrams is contrary to the findings of Mukherjee et al. (2013a) and Mukherjee et al. (2013b). This might be due to differences in language, data, or review topics.

6 CONCLUSION

This paper has performed the first in-depth investigation into deception detection on Dutch reviews by focusing on text-based lexical and syntactic features. A previous research has shown the usefulness of the Dutch CLiPS Stylometry Investigation (CSI) corpus in developing classification systems for deception detection. Using the CSI corpus, this research has demonstrated that removing specific words that have a high frequency in Dutch—stopwords—improves the overall quality of deception detection by consistently effectuating higher accuracy, recall, precision, and F1 scores. Furthermore, this research has established that using text-based lexical and syntactic features improves deception detection. Using word unigrams, word bigrams, depedency unigrams, and filtering for stopwords results in an accuracy score of 79.3%. Although not all lexical and syntactic features necessarily improve deception detection compared to the internal baseline of 78.8%, it can be stated that they do provide a useful dicriminative power.

The research has several theoretical contributions. Firstly, it has been discovered that the χ^2 measure proves to be unusable in deception detection by effectuating only near-chance results. Furthermore it has found that the lexical context (word bigrams) and the syntactic structure of sentences do encode valuable cues to detect deceptive writings. However, these contributions might be limited because the CSI corpus consists of simulated reviews. As the research did not cover real online reviews, this might interfere with the generalizability of the findings. This might be a starting point for subsequent research. Furthermore, future research could extend the work in this research by focusing on behavioral features of the respective Dutch reviews. Providing a weighted framework for lexical and syntactic features in combination with behavioral features might result in a improved robust system for deception detection in Dutch. Subsequent research could also focus on the development of a language independent classification system. Using the findings of papers that focus on different languages might result in general applicable features. Such systems could focus on similarities in grammar formalisms.

BIBLIOGRAPHY

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in Economics*, pp. 235–251. Elsevier.

Anderson, E. W. (1998). Customer satisfaction and word of mouth. *Journal of service research 1*(1), 5–17.

Bond, C. F. and B. M. DePaulo (2006, Aug). Accuracy of deception judgments. *Personality and Social Psychology Review 10*(3), 214–234.

Bouma, G., G. Van Noord, and R. Malouf (2001). Alpino: Wide-coverage computational analysis of dutch. *Language and Computers 37*(1), 45–59.

Bouma, G., G. Van Noord, and R. Malouf (2011). Alpino web-application. http://www.let.rug.nl/vannoord/bin/alpino.

Buller, D. B. and J. K. Burgoon (1996). Interpersonal deception theory. *Communication theory 6*(3), 203–242.

Chang, H. H. and L. H. Wu (2014). An examination of negative e-wom adoption: Brand commitment as a moderator. *Decision Support Systems 59*, 206–218.

Chen, J., L. Teng, Y. Yu, and X. Yu (2016). The effect of online information sources on purchase intentions between consumers with high and low susceptibility to informational influence. *Journal of Business Research 69*(2), 467–475.

Chen, Y. and J. Xie (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management science 54*(3), 477–491.

Cheung, C. M., M. K. Lee, and N. Rabjohn (2008). The impact of electronic word-of-mouth: The adoption of online opinions in online customer communities. *Internet research 18*(3), 229–247.

D. Hollebeek, L. and T. Chen (2014). Exploring positively-versus negatively-valenced brand engagement: a conceptual model. *Journal of Product & Brand Management 23*(1), 62–74.

Frank, M. G., M. A. Menasco, and M. O'Sullivan (2008). Human behavior and deception detection. *Wiley Handbook of Science and Technology for Homeland Security*.

Hauch, V., I. Blandón-Gitlin, J. Masip, and S. L. Sporer (2015, Nov). Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review 19*(4), 307–342.

Hennig-Thurau, T., G. Walsh, and G. Walsh (2003). Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the internet. *International journal of electronic commerce 8*(2), 51–74.

Jindal, N. and B. Liu (2008). Opinion spam and analysis. *In Proceedings of the international conference on Web search and web data mining*, 309–319.

Kwark, Y., J. Chen, and S. Raghunathan (2014). Online product reviews: Implications for retailers and competing manufacturers. *Information systems research 25*(1), 93–110.

Manning, C. D., H. Schütze, et al. (1999). *Foundations of statistical natural language processing*, Volume 999. MIT Press.

Meuter, M. L., D. B. McCabe, and J. M. Curran (2013). Electronic word-of-mouth versus interpersonal word-of-mouth: are all forms of word-of-mouth equally influential? *Services Marketing Quarterly 34*(3), 240–256.

Mukherjee, A., V. Venkataraman, B. Liu, and N. Glance (2013a). Fake review detection: Classification and analysis of real and pseudo reviews. *Technical Report UIC-CS-2013–03, University of Illinois at Chicago, Tech. Rep.*.

Mukherjee, A., V. Venkataraman, B. Liu, and N. S. Glance (2013b). What yelp fake review filter might be doing? In *ICWSM*.

Newman, M. L., J. W. Pennebaker, D. S. Berry, and J. M. Richards (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin 29*(5), 665–675.

Oliveira, T., M. Alhinho, P. Rita, and G. Dhillon (2017). Modelling and testing consumer trust dimensions in e-commerce. *Computers in Human Behavior 71*, 153–164.

Ott, M., C. Cardie, and J. Hancock (2012). Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International Conference on World Wide Web*, pp. 201–210. ACM.

Ott, M., Y. Choi, C. Cardie, and J. Hancock (2011). Finding deceptive opinion spam by any stretch of the imagination. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 219–230.

Perkins, J. (2010). *Python text processing with NLTK 2.0 cookbook*. Packt Publishing Ltd.

Pfeffer, J., T. Zorbach, and K. M. Carley (2014). Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications 20*(1-2), 117–128.

Smith, D., S. Menon, and K. Sivakumar (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of interactive marketing 19*(3), 15–37.

Verhoeven, B. and W. Daelemans (2014). Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. *In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland.*.

Yoo, K.-H. and U. Gretzel (2009). Comparison of deceptive and truthful travel reviews. *Information and communication technologies in tourism 2009*, 37–47.

Zhou, L., J. K. Burgoon, J. F. Nunamaker, and D. Twitchell (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation 13*(1), 81–106.

Zhou, L. and Y.-w. Sung (2008). Cues to deception in online chinese groups. In *Hawaii international conference on system sciences, proceedings of the 41st annual*, pp. 146–146. IEEE.