# Using Named Entities for Computer-Automated Verbal Deception Detection

Kleinberg, B.; Mozes, M.; Arntz, A.; Verschuere, B.

**DOI**
10.1111/1556-4029.13645

**Publication date**
2018

**Document Version**
Final published version

**Published in**
Journal of Forensic Sciences

**License**
Article 25fa Dutch Copyright Act

Link to publication

**Citation for published version (APA):**
Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018). Using Named Entities for Computer-Automated Verbal Deception Detection. *Journal of Forensic Sciences*, *63*(3), 714-723. https://doi.org/10.1111/1556-4029.13645

JOURNAL OF FORENSIC SCIENCES

*J Forensic Sci*, May 2018, Vol. 63, No. 3
doi: 10.1111/1556-4029.13645
Available online at: onlinelibrary.wiley.com

PAPER

GENERAL

*Bennett Kleinberg*,[1] *M.Sc.; Maximilian Mozes*,[1,2]*; Arnoud Arntz*,[1] *Ph.D.; and Bruno Verschuere*,[1] *Ph.D.*

# Using Named Entities for Computer-Automated Verbal Deception Detection

**ABSTRACT:** There is an increasing demand for automated verbal deception detection systems. We propose named entity recognition (NER; i.e., the automatic identification and extraction of information from text) to model three established theoretical principles: (i) truth tellers provide accounts that are richer in detail, (ii) contain more contextual references (specific persons, locations, and times), and (iii) deceivers tend to withhold potentially checkable information. We test whether NER captures these theoretical concepts and can automatically identify truthful versus deceptive hotel reviews. We extracted the proportion of named entities with two NER tools (spaCy and Stanford's NER) and compared the discriminative ability to a lexicon word count approach (LIWC) and a measure of sentence specificity (speciteller). Named entities discriminated truthful from deceptive hotel reviews above chance level, and outperformed the lexicon approach and sentence specificity. This investigation suggests that named entities may be a useful addition to existing automated verbal deception detection approaches.

**KEYWORDS:** forensic science, computational linguistics, deception detection, named entity recognition, linguistic inquiry and word count, reality monitoring, criteria-based content analysis

With an increased demand for security systems like airport border control, researchers and practitioners alike have identified the need for applications to detect deception on a large scale (1). For example, airport security settings preclude many tools used in deception research (e.g., polygraphy, extensive interviewing) due to their limited applicability (e.g., real-time data analysis; scalability). The method of verbal deception detection seems promising as it is rooted in the assumption that the content of a statement contains information about the statement's veracity. By looking at the content of what is said or written, the equipment needed is minimal compared to the psychophysiological deception detection toolkit. For applied purposes (e.g., airport security), verbal deception detection would offer a viable alternative to currently applied, yet scientifically questionable techniques, such as detecting deception based on suspicious behavior (2,3). Oberlader et al.'s (4) meta-analysis concludes that "content-based techniques are [. . .] among the best available empirically validated methods for the veracity assessment of statements" (p. 14). However, in its current state, verbal deception detection is not fit for large-scale applications for at least two reasons. First, a massive system requires near real-time analysis of a text to aid practitioners, for example, in determining the subsequent procedure. An applied system also needs to be scalable (i.e., be able to process vast numbers of people within a short time) to meet the demands of settings such as airport security with tens of thousands of passengers per day. The most significant impediment herein is that verbal deception detection relies on trained human judges who score each statement on a set of criteria (e.g., the richness of detail). This process is time-consuming as it involves face-to-face interviewing, transcribing of interviews, and individual scoring and is therefore less suitable for large-scale systems. Second, human judges' scoring is never entirely reliable (5,6). Typically, two or more trained coders read a statement and rate it on a set of criteria. The coders' agreement (i.e., inter-rater reliability) often indicates that there is considerable variation between two independent judgments (7) which poses a threat to the validity of any scoring method's reliability (for alternative human coding methods see [6]). Novel paths in computational linguistics might offer a solution to these two limitations of verbal deception detection.

### *Computer-Automated Verbal Deception Detection*

Several studies have examined how the verbal content of statements can be analyzed automatically (8). The aim of automated approaches to verbal deception detection is substituting the human coding of statements (i.e., the extraction and counting of features/cues) with the fast and reliable algorithmic extraction of cues to deception. Some studies on computer-automated analysis for verbal deception detection have used a lexicon-based approach, typically using the Linguistic Inquiry and Word Count software (LIWC) (9–11). Text statements processed with the LIWC return proportions of word categories occurring in the text. Each word category is intended to model psycholinguistic variables. For example, the LIWC category "affect" models emotional processes by counting the occurrences of words in a text that match a large dictionary of words intended to represent emotional processes (e.g., happy, sad). For deception detection specifically, a LIWC-analysis revealed that, for example, truthful statements contained more first-person pronouns and self-references (e.g., "mine," "our") than false statements, whereas false

[1]Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129 D, 1018 WS, Amsterdam, The Netherlands.
[2]Department of Informatics, Technical University of Munich, Boltzmannstr. 3, Garching near, Munich, Germany.

Received 27 April 2017; and in revised form 24 Aug. 2017; accepted 28 Aug. 2017.

statements contained more words referring to certainty (e.g., “totally,” “truly”) and to other-references (e.g., “they,” “themselves”) (12).

Other studies have used supervised machine learning to build classifiers of multiple features (=psychological cues and *n*-grams) that learn to differentiate between false and truthful statements. Ott et al. (10,11) applied linear Support Vector Machine (SVM) classifiers using occurrences of frequent two-word units (i.e., bi-grams) on deceptive and truthful hotel reviews. They found that the classifiers (89% accuracy) outperformed the human judge performances (58% accurate for positive reviews; 69% for negative reviews). Bachenko et al. (13) examined 275 unique propositions (e.g., “I just feel hopeless,” p. 43) uttered in real criminal cases. They were able to classify these propositions through psychological and linguistic features with a sensitivity of 76% and specificity of 74%.

The emerging body of computer-automated verbal deception research indicates that computer automation not only performs equal to human-annotated statements (12,13), but it also allows for a finer level of analysis (e.g., single utterances rather than whole statements only) (14), and is increasingly used to model more nuanced variables (e.g., jargon) (15). However, one limitation of automated verbal deception detection, predominantly done with machine learning classification, is that of the poor generalization across multiple contexts; that is, the high classification accuracies might be overestimations obtained through training and testing a classifier in the same domain (9,12). In contrast to data-driven approaches, investigations based on theoretical verbal deception principles could be more likely to generalize across domains. As theoretical principles are formulated to grasp the core mechanisms involved in (verbal) deception *in general* one can argue that, at least to some degree, these fundamental mechanisms should be at play in multiple deception contexts (a point we come back to in the Discussion; and for a combination of data-driven and theory-led approaches see [16]). Therefore, the promise of computer-automated deception detection could be further improved by including algorithmic operationalizations of theoretical constructs.

### *Named Entities: A Tool for Automated Verbal Deception Detection?*

The focus of this study is how named entity recognition (NER) can help bridge the gap between computer-automated deception detection and verbal deception detection theory. NER is a subfield from the areas of natural language processing and information extraction that deals with identifying and extracting so-called *named entities* from a text (17). The term “*named entity”* was initially defined during the sixth the Message Understanding Conference (MUC-6) in 1996 (18) and comprised the identification of people, organizations, and locations in texts. In its most basic form, NER aims to extract information from text (e.g., single words or phrases) and to classify them into predefined categories (e.g., persons, locations, organizations, currencies) whereby a mixture of methods is used to extract these categories (19).

While early concepts mainly relied on the use of rule-based algorithms, more recent approaches utilize underlying probabilistic models, unsupervised and supervised statistical learning algorithms (17). The following example is a common way to represent text with annotated named entities: “We met yesterday [DATE] at 11:30 am [TIME] on Coronado [GPE] beach, then went to Starbucks [ORG] and paid $3.50 [MONEY] for a coffee” (GPE = geopolitical entity; ORG = organization.). The specific category labels differ per NER algorithm. Throughout the study, we use the named entity recognizer from the natural language processing tool spaCy – written in the Python programming language (20) (see Table 2 for keys to all named entities extracted with spaCy). There are at least three theoretical rationales why we consider NER to be an attractive candidate for automated verbal deception detection.

### *Theoretical Rationale 1: Richness of Detail*

The theory of Reality Monitoring (used interchangeably here with interpersonal Reality Monitoring) (21) suggests that the source of one’s memory determines how a memory is recalled (22). Genuine memories have been obtained through sensory experiences whereas fabricated memories were constructed through cognitive operations. Consequently, the narratives of genuine memories should be richer in sensory information (e.g., perceptual, spatial, temporal information), whereas narratives of fabricated memories should contain more references to cognitive operations (23). This theoretical framework has been extended to deception, with truthful statements expected to contain more perceptual details, more temporal details, and more spatial details. Meta-analytical research supports the theoretical predictions from RM (23,24), showing that truthful statements contain more specific information (esp. temporal and spatial) than false statements, and offering a potential application of named entities.

### *Theoretical Rationale 2: Contextual Embeddings*

The idea behind the popular criteria-based content analysis (CBCA) is that truthful statements differ from false statements in quality and content because the process through which the statements are constructed is different. Like RM, CBCA also considers the quantity of detail (25) as well as contextual embedding (*“Events being placed in time and location, and actions being connected with other daily activities and/or customs”*; (24), p. 8) as a sign of veracity. These imply references to concrete information about an activity or about events. It is widely corroborated that truthful statements contain more contextual embeddings than false statements (25). Named entities might be an automated means to approximate the cue of contextual embeddings.

### *Theoretical Rationale 3: The Verifiability of Details*

While the richness of detail and the contextual embeddings point to characteristics of *truthful* statements, the Verifiability Approach (VA) (26) explains verbal features of *deceptive* statements. Inclined to appear truthful, deceivers realize they have to come across as forthcoming and talkative (i.e., providing a statement with sufficient detail to sound convincing). The deceivers’ dilemma is that, at the same time, they have to avoid giving information that an interviewer or conversation partner could potentially verify (26).

For example, an answer like “I spoke to my friend James in the Vondelpark” might be a detail that theoretically could be checked by the interviewer (e.g., by consulting James), whereas “I spoke to someone in the park” would not be verifiable. Several studies (26,27) indeed found that the amount of verifiable information discriminates deceivers from truth tellers, at least when instructed to mention as much verifiable information as possible. The working definition of verifiable information includes any activity that (i) has been done with an identifiable person, (ii) has been witnessed by an identifiable person, or (iii) has been recorded through technology (e.g., CCTV, email, social networks). These three criteria suggest that references to persons and locations are key; both of which can potentially be operationalized through named entities.

### *The Current Study*

A scoring method that meets the large-scale applicability requirements of being fast and automated but that at the same time encapsulates the theoretical frameworks supported by a vast body of research would benefit verbal deception research. The primary objective of this paper is to examine whether named entities are suitable to grasp the postulated difference between truthful and deceptive statements. Consequently, our main hypothesis is that *the number of named entities is higher in truthful statements than the number of named entities in deceptive statements*.

We compare named entities with indicators derived from two existing computer-automated tools (LIWC, '[28]; and the speciteller tool [29]). Although the LIWC is the most popular tool in computational linguistics deception research, we reason that the NER approach is more capable of grasping the difference in statement specificity. Named entities tap into categories that match the criteria of the richness of detail, contextual embeddings, and the verifiability notion, such as "persons" or "locations." Furthermore, NER does not rely on lexicons and should therefore be more flexible toward unseen words that are not in predefined lexicons. We also compare two NER tools to the *speciteller* tool, which models sentence specificity. The speciteller originates from the observation that two propositions can be similar in meaning but differ in the degree of specificity. Like the NER tools, the speciteller tool was not designed with verbal deception theory in mind.

For the course of this paper, we use the hotel review datasets provided by Ott et al. (10,11). Not only do these data offer an exceptional corpus of truthful and deceptive hotel reviews of both positive and negative valence, but they are also suitable in size (1600 statements) for reliable statistical analysis. We compare the truthful and deceptive positive reviews on NER, LIWC, and speciteller. To assess the generalizability of the findings, we evaluate the results of negative and positive reviews separately. A machine learning head-to-head comparison between the complete LIWC and the named entity approach is available in the online supplementary material at *anonymouslink*.

## Materials and Methods

The named entity extraction code (python), the analysis code (R), and the data used in the current investigation are available at https://osf.io/2qjs4/.

### *Dataset*

The dataset consists of 1600 positive and negative, truthful and deceptive reviews on 20 hotels in Chicago. Ott et al. (11) provided the first publicly available opinion spam dataset containing 800 gold-standard positive hotel reviews. Truthful data were gathered by selecting reviews of the 20 most popular hotels in the Chicago area listed on TripAdvisor in 2011. The selection criteria for each review were that (i) its author has published an opinion on TripAdvisor before, (ii) it was written in the English language, and (iii) it was a five-star review with a minimum length of 150 characters. The authors selected 400 reviews ($M_{\text{length}}$ = 123.63, $SD_{\text{length}}$ = 68.33) of equal length distributed evenly on the 20 hotels, resulting in 20 reviews for each hotel. Deceptive hotel reviews were obtained via Amazon Mechanical Turk. Four hundred reviews ($M_{\text{length}}$ = 116.24, $SD_{\text{length}}$ = 61.69) were collected by instructing participants to review the hotel realistically and positively from a customer's perspective (i.e., to write a fake positive review).

Ott et al. (10) extended the 2011 dataset by adding 800 negative hotel reviews about the same hotels. The procedure was identical to Ott et al. (11). Contrary to the positive reviews, participants were instructed to write a fake negative review about a competitor's hotel ($M_{\text{length}}$ = 178.16, $SD_{\text{length}}$ = 93.60). Genuine negative reviews ($M_{\text{length}}$ = 179.49, $SD_{\text{length}}$ = 100.57) were 1- or 2-star rated reviews collected from six hotel review websites (Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, Yelp).

### *Named Entity Recognition (NER)*

We operationalize the occurrence of named entities recognized with spaCy's named entity recognizer. spaCy is an open-source library providing natural language processing tools for the Python programming language (Version 1.3.0) (20). Research by Jiang et al. (30) shows that spaCy's NER tool performs second best among four well-established open-source NER tools regarding accuracy and that it is the fastest in processing speed.

spaCy's NER tool extracts named entities in eighteen categories: persons, nationalities or religious groups, facilities, organizations, geopolitical entities, locations, products, events, works of art, law documents, languages, dates, times, percentages, money, quantities, ordinals, and cardinals (see [30]). From a technical perspective, spaCy utilizes a set of well-established entity recognition models that are based on statistical learning methods to identify named entities in texts. The technical details and statistical models of the NER algorithm are beyond the scope of this paper, but source-code can be consulted via https://github.com/explosion/spaCy.

All categories fit conceptually with the notion that (i) they are specific and therefore potentially checkable, and (ii) that deceivers might be more inclined than truth tellers to avoid mentioning them. The proportion of named entities is obtained by adding the number of unique occurrences (i.e., counting a recurring entity only once per statement) of named entities divided by the overall word count per hotel review. As verbal content-based scoring tools (RM and CBCA) do typically not include repetitions of details, we use the proportion of unique named entities as the primary dependent variable for information specificity.

As spaCy's NER is a rather new open-source tool, we compare it to a second, often-used NER system, namely the Stanford Named Entity Recognizer (Stanford NER) (31). Stanford NER is a publicly available software tool written in Java that is capable of identifying named entities of the seven categories locations, persons, organizations, money, percent, date, and time. The technical details of the underlying recognition algorithm of Stanford NER can be obtained via https://nlp.stanford.edu/software/CRF-NER.shtml. The NER evaluation paper by Jiang et al. (30) concluded that Stanford's NER is the most accurate overall. Thus, we test our key hypothesis with the two NER systems that were best evaluated regarding speed and accuracy.

Box 1 shows an annotated example of reviews high and low in information specificity using both spaCy's as well as Stanford's NER system.

BOX 1—*Annotated example with the spaCy and Stanford NER systems of a truthful and a deceptive positive review of the same hotel (excerpt).*

| Truthful (high frequency of named entities) | Deceptive (low frequency of named entities) |
|---|---|
| Spacy's named entity recognizer | |
| We stayed at the *Hard Rock* on January 27th, 2009 **[DATE]** for $125/night **[MONEY]**. It is located on Michigan Ave **[PERSON]**[a], just two **[CARDINAL]** blocks from the *Mag Mile*, two **[CARDINAL]** blocks from Millennium Park **[PERSON]**, and five **[CARDINAL]** blocks from the Chicago Art Institute **[ORGANIZATION]**. | My husband and I recently stayed at the *Hard Rock Hotel* Chicago **[GPE]** and we can't wait to go back! The hotel is located in downtown Chicago **[GPE]** and seems to be at the heart of the city, we were close to everything. The *Hard Rock Hotel* is forty **[CARDINAL]** stories high and the view from our room was simply breathtaking. [. . .] |
| Stanford's named entity recognizer | |
| We stayed at the *Hard Rock* on January 27th **[DATE],** 2009 **[DATE]** for $125/night **[MONEY].** It is located on Michigan **[LOCATION]** Ave, just *two* blocks from the *Mag Mile*, *two* blocks from Millennium Park **[LOCATION]**, and *five* blocks from the Chicago Art Institute **[ORGANIZATION]**. [. . .] | My husband and I recently stayed at the Hard Rock Hotel Chicago **[ORGANIZATION]** and we can't wait to go back! The hotel is located in downtown Chicago **[LOCATION]** and seems to be at the heart of the city, we were close to everything. The *Hard Rock Hotel* is *forty* stories high and the view from our room was simply breathtaking. [. . .] |

Recognized named entities are underlined and their category labels are bold in **[brackets]**. Nonrecognized named entities are in *italics*.
[a]Note the misclassification of "Michigan Ave" (should be a [GPE]) as a [PERSON] (see Discussion).

### *Sentence Specificity*

Two sentences can convey the same content but might vary in the specificity that these propositions are embedded in. This observation led to the development of *speciteller* (29), a machine learning-based classifier written in Python that gives the specificity of a sentence ranging from 0 (lowest) to 1 (highest). Li and Nenkova (29) had five independent annotators code a sample of 885 sentences from the Wall Street Journal, New York Times, and Associated Press. This annotation was used to build a classifier with shallow surface features (e.g., the number of words, estimated number of named entities) and dictionary features (e.g., subjective words, concreteness).

Using machine learning techniques (supervised logistic regression, semi-supervised, and co-training classification), they derived a final classifier released under the name speciteller. We calculated the sentence specificity for each sentence per hotel review and divided the sum of sentence specificity by the number of sentences per hotel review. We used the NLTK sentence boundary detector to split the reviews into sentences (32). The resulting variable is referred to as *average sentence specificity*.

### *Linguistic Inquiry and Word Count (LIWC)*

Previous research proposed that word categories offered by LIWC might function as a proxy of classic RM variables (33). Specifically, the richness of detail was modeled with the LIWC categories "perceptual processes," "space references," and "time references" and was shown to be acceptable for modeling individual RM scoring. Using a logistic regression classifier yielded a sensitivity of 71.1% and a specificity of 64.5%, thereby performing well above chance level and outperforming human lie detectors.

In the current study, we use the LIWC to model richness in detail by summing the proportions of words belonging to the categories *percept* (=perceptual processes; incl. the subcategories *see*, *hear*, and *feel*; e.g., saw, touch, heard), *space* (=spatial references; e.g., down, in), and *time* (=temporal references; e.g., until, end) (33). The sum of proportions of these three categories is henceforth referred to as *LIWC richness of detail*.

## Results

### *Analytical Plan*

The statistical analysis consists of three separate 2 (Valence: negative vs. positive) by 2 (Veracity: deceptive vs. truthful) between-subjects ANOVAs on (i) the proportion of named entities, (ii) the speciteller average sentence specificity, and (iii) the LIWC richness of detail. We use Cohen's $f$ to denote the magnitude of effects for the statistical tests (with 0.10, 0.25, and 0.40 for a small, moderate, and substantial effect, respectively) (34). A significant main effect of Veracity would support our hypothesis. To assess the diagnostic accuracy of the single variables and to compare variables with each other, we conduct Receiver Operating Characteristics (ROC) analysis.

### *Main Analysis*

*Proportion of Named Entities*—Using the spaCy NER system, the 2 (Veracity: deceptive vs. truthful) by 2 (Valence: positive vs. negative) ANOVA on the proportion of unique named entities revealed the predicted significant main effect of Veracity, $F(1, 1596) = 137.95$, $p < 0.001$, $f = 0.29$. There was also a significant main effect of Valence, $F(1, 1596) = 97.44$, $p < 0.001$, $f = 0.25$, and a significant Veracity*Valence interaction, $F(1, 1596) = 4.65$, $p = 0.031$, $f = 0.05$. The interaction revealed that the difference in the proportion of unique named entities between truthful and deceptive hotel reviews was stronger for positive, $F(1, 798) = 75.64$, $p < 0.001$, $f = 0.31$, than for negative reviews, $F(1, 798) = 63.63$, $p < 0.001$, $f = 0.28$ (Table 1).

Using Stanford's NER system, the 2 (Veracity: deceptive vs. truthful) by 2 (Valence: positive vs. negative) ANOVA on the proportion of unique named entities revealed the predicted significant main effect of Veracity, $F(1, 1596) = 15.15$, $p < 0.001$, $f = 0.10$. There was also a significant main effect of Valence, $F(1, 1596) = 187.41$, $p < 0.001$, $f = 0.34$; and a significant Veracity*Valence interaction, $F(1, 1596) = 7.73$, $p = 0.003$, $f = 0.07$. The interaction showed that the truthful-deceptive difference in the proportion of unique named entities between was only significant for positive, $F(1, 798) = 16.02$, $p < 0.001$, $f = 0.14$, but not for negative reviews, $F(1, 798) = 1.01$, $p = 0.314$, $f = 0.04$ (Table 1).

TABLE 1—*Means (SDs), effect size and AUC for the dependent variables per valence and veracity.*

| | Positive Hotel Reviews | | | | Negative Hotel Reviews | | | |
|---|---|---|---|---|---|---|---|---|
| | M Truthful (SD) | M Deceptive (SD) | *f* | AUC [95% CI] | M Truthful (SD) | M Deceptive (SD) | *f* | AUC [95% CI] |
| % unique named entities (spaCy) | 4.14 (2.36) | 2.87 (1.73) | 0.30 | 0.67 [0.63–0.71] | 3.04 (1.77) | 2.17 (1.31) | 0.28 | 0.65 [0.61–0.69] |
| % unique named entities (Stanford) | 2.39 (1.79) | 1.93 (1.40) | 0.14 | 0.57 [0.52–0.61] | 1.26 (1.15) | 1.18 (0.98) | 0.04$^{\text{ns}}$ | 0.51 [0.47–0.55] |
| Sentence specificity | 19.11 (16.06) | 15.03 (12.92) | 0.14 | 0.59 [0.55–0.63] | 19.71 (13.30) | 16.02 (11.98) | 0.15 | 0.60 [0.56–0.64] |
| Detailedness (LIWC) | 18.48 (4.17) | 17.43 (4.09) | 0.13 | 0.57 [0.53–0.61] | 18.29 (4.03) | 18.24 (4.01) | 0.01$^{\text{ns}}$ | 0.50 [0.46–0.54] |

$^{\text{ns}}$Nonsignificant at $p < 0.05$.

*Average Sentence Specificity*—The 2 (Veracity: deceptive vs. truthful) by 2 (Valence: positive vs. negative) ANOVA on the average sentence specificity revealed a significant main effect of Veracity, $F(1, 1596) = 32.44$, $p < 0.001$, $f = 0.14$. There was no significant main effect of Valence, $F(1, 1596) = 1.37$, $p = 0.243$, $f = 0.03$; nor was there a significant Veracity*Valence interaction, $F(1, 1596) = 0.08$, $p = 0.771$, $f = 0.01$ (Table 1).

*LIWC: Richness of Detail*—The 2 (Veracity: deceptive vs. truthful) by 2 (Valence: positive vs. negative) ANOVA on LIWC richness of detail revealed a significant main effect of Veracity, $F(1, 1596) = 7.32$, $p = 0.007$, $f = 0.07$. There was no significant main effect of Valence, $F(1, 1596) = 2.36$, $p = 0.124$, $f = 0.04$. The significant Veracity*Valence interaction, $F(1, 1596) = 5.93$, $p = 0.015$, $f = 0.06$, indicated that the difference between truthful and deceptive hotel reviews was significant for positive reviews, $F(1, 798) = 12.85$, $p < 0.001$, $f = 0.13$, but not for negative reviews, $F(1, 798) = 0.04$, $p = 0.846$, $f = 0.01$ (Table 1).

*Receiver Operating Characteristics*—As a measure of the diagnostic efficiency of the dependent variables, we calculated the area under the curve (AUC). The AUC is the surface area under the graph resulting from plotting the sensitivity (i.e., the true positives) against 1-sensitivity (i.e., the true negatives) for all observed criterion values (i.e., the dependent variables: named entities, average sentence specificity, LIWC richness of detail) (35). Theoretically, the AUC can assume values between 0 and 1, whereby a value of 0.50 represents random classification accuracy. The closer the AUC is to 1, the better the discriminatory power of the criterion. All ROC calculations were conducted with the pROC R package (36). In contrast to performance metrics such as accuracy, recall, and precision, the AUC does not require a specific (often arbitrary) cutoff value to classify a statement as truthful or deceptive. Rather it represents the diagnostic power of a criterion variable across all possible cutoff values (i.e., how good are named entities, for example, to tell truthful from false hotel reviews in general).

The AUC for the proportion of unique named entities, irrespective of the reviews' valence, was 0.66 [95%CI: 0.63–0.68] with spaCy's NER system, and 0.54 [0.51–0.56] with Stanford's NER system (Table 1). For the average sentence specificity, the AUC was 0.60 [0.57–0.62], and for the LIWC richness of detail, the AUC was 0.53 [0.51–0.56]. To test whether the criteria differed in their diagnostic efficiency, we used Venktraman's AUC comparison test (37). The AUC for the proportion of unique named entities with spaCy's named entity recognition outperformed that of Stanford's named entity recognition, $E = 149630$, bootstraps = 2000, $p < 0.001$.

As the NER with the spaCy software seems superior for deception detection than Stanford's NER, the remainder of the analysis focuses on the results obtained with spaCy. The AUC for the proportion of unique named entities was significantly larger than that of the average sentence specificity, $E = 82970$, bootstraps = 2000, $p < 0.001$; and also larger than that of the LIWC richness of detail, $E = 159860$, bootstraps = 2000, $p < 0.001$. Lastly, the AUC for the average sentence specificity was significantly larger than that of the LIWC richness of detail, $E = 80878$, bootstraps = 2000, $p < 0.001$.

### *Exploratory Analyses*

*Verifiable Named Entities*—Although all eighteen named entity categories are related to the notion of the richness and verifiability of detail, some categories (e.g., persons, locations, times) may fit these theoretical lines better than others (e.g., works of art, language references). To explore that idea, we selected only those named entities that could, in principle, lead to a verifiability of the given information. That is, we calculated the proportion of unique named entities referring to persons, facilities, geopolitical entities, locations, organizations, events, dates, times, money. We found a significant main effect of Veracity, $F(1, 1596) = 72.11$, $p < 0.001$, $f = 0.21$; a significant main effect of Valence, $F(1, 1596) = 155.11$, $p < 0.001$, $f = 0.31$; and a significant Veracity*Valence interaction, $F(1, 1596) = 13.25$, $p < 0.001$, $f = 0.09$. The interaction revealed that the difference between truthful reviews was more pronounced for positive reviews ($M_{\text{truthful}} = 3.33$, $SD_{\text{truthful}} = 2.10$; $M_{\text{deceptive}} = 2.37$, $\text{SD}_{\text{deceptive}} = 1.56$), $F(1, 798) = 53.01$, $p < 0.001$, $f = 0.26$, than for negative reviews ($M_{\text{truthful}} = 2.06$, $\text{SD}_{\text{truthful}} = 1.33$; $M_{\text{deceptive}} = 1.68$, $\text{SD}_{\text{deceptive}} = 1.11$), $F(1, 798) = 19.24$, $p < 0.001$, $f = 0.16$. Compared to the results of the overall proportion of named entities, we conclude that using only "verifiable" named entities did not increase the deceptive-truthful difference.

*Most Frequent Named Entities*—As many named entities do occur only rarely in the hotel reviews (Table 3), we also calculated the proportion of unique named entities that occurred in at least 10% of the reviews. This criterion resulted in the inclusion of persons, facilities, dates, times, money, ordinals, and cardinals. The analysis revealed a significant main effect of Veracity, $F(1, 1596) = 219.44$, $p < 0.001$, $f = 0.37$, and a significant Veracity*Valence interaction, $F(1, 1596) = 9.80$, $p = 0.002$, $f = 0.08$. The main effect of Valence was not significant, $F(1, 1596) = 3.57$, $p = 0.059$, $f = 0.05$. The interaction indicated that the difference between truthful reviews was more pronounced for positive reviews ($M_{\text{truthful}} = 2.48$, $\text{SD}_{\text{truthful}} = 1.85$; $M_{\text{deceptive}} = 1.22$, $\text{SD}_{\text{deceptive}} = 1.17$; $f = 0.46$), $F(1, 798) = 132.05$, $p < 0.001$, $f = 0.41$) than for negative reviews ($M_{\text{truthful}} = 2.13$, $\text{SD}_{\text{truthful}} = 1.42$; $M_{\text{deceptive}} = 1.31$, $\text{SD}_{\text{deceptive}} = 1.02$), $F(1, 798) = 87.41$, $p < 0.001$, $f = 0.33$).

Table 2 shows the descriptive statistics for each named entity category as well as the *f* effect sizes for the Veracity main effect. These results suggest that dates (spaCy's NER only), references to money, and the occurrence of ordinals and cardinals (highlighted in bold in Table 2) were consistently significant predictors of a review's veracity.

*Other LIWC Categories*—Although the LIWC category selection was based on previous research (33), the three categories "percept," "space," and "time" represent only a subset of all LIWC categories. Table 3 shows those LIWC categories that resulted in significant truthful-deceptive differences (see online supplementary material on the Open Science Framework at https://osf.io/2qjs4/ for all 92 categories). These findings show that punctuation was more pronounced in truthful than in deceptive reviews (11). Deceptive reviews also contained more function words, more pronouns overall, more personal pronouns, more first-person pronouns, and more verbs than truthful statements. Interestingly, the inclusion of numbers was bigger in truthful than in deceptive statements, which corroborates the findings for ordinals and cardinals as named entities.

## Discussion

The current investigation set out to examine how named entities can be used to synthesize verbal deception theory with a

TABLE 2—*Descriptive statistics of the proportion unique named entities (M, SD) per veracity and valence.*

| | | | Positive | | | Negative | | | |
|---|---|---|---|---|---|---|---|---|---|
| Named entity category | Example | *f* | Truthful | Deceptive | *f* | Truthful | Deceptive | *f* | Sparsity |
| Persons (spaCy) | "Janice" | 0.16* | 54.98 (90.40) | 25.08 (53.35) | 0.20* | 19.78 (39.48) | 13.34 (32.70) | 0.09 | 72.44 |
| **Persons** (Stanford) | "Peter" | 0.13* | 16.24 (46.70) | 7.29 (25.76) | 0.12* | 7.92 (22.07) | 2.12 (11.78) | 0.16* | 89.38 |
| Nationalities | "Egyptian" | 0.03 | 5.83 (26.89) | 4.42 (20.88) | 0.03 | 2.99 (15.47) | 2.00 (11.89) | 0.04 | 94.75 |
| Facilities | "North Bridge" | 0.04 | 16.22 (41.88) | 9.84 (29.10) | 0.09 | 5.06 (17.50) | 6.44 (23.84) | 0.03 | 88.06 |
| Organizations (spaCy) | "McDonald's" | 0.06 | 79.11 (96.77) | 70.22 (95.85) | 0.05 | 48.18 (65.74) | 39.09 (53.37) | 0.08 | 48.25 |
| Organizations (Stanford) | "Hard Rock Hotel Chicago" | −0.07 | 70.62 (88.63) | 75.82 (81.51) | −0.03 | 34.42 (60.11) | 48.63 (53.59) | −0.12* | 47.00 |
| Geopolitical entities | "Chicago" | −0.11* | 59.97 (71.64) | 77.44 (74.52) | 0.12* | 27.87 (44.41) | 37.87 (51.53) | 0.10 | 46.50 |
| Locations (spaCy) | "Caribbean" | 0.06 | 13.72 (42.67) | 8.42 (29.20) | 0.07 | 4.45 (21.29) | 2.83 (15.70) | 0.04 | 91.88 |
| Locations (Stanford) | "Michigan" | −0.04 | 84.10 (86.02) | 88.14 (85.99) | −0.02 | 28.54 (55.47) | 38.46 (63.97) | −0.06 | 41.00 |
| Products | "Diesel Ford Excursion" | 0.05 | 0.96 (7.43) | 0.14 (2.89) | 0.07 | 0.57 (7.20) | 0.26 (5.15) | 0.02 | 99.25 |
| Events | "New Years" | 0.08* | 1.90 (12.90) | 0.46 (4.34) | 0.07 | 1.52 (12.28) | 0.00 (0.00) | 0.09 | 98.63 |
| Works of art | "The Magnificent Mile" | 0.02 | 2.27 (15.15) | 1.15 (10.94) | 0.04 | 1.07 (8.53) | 1.25 (9.51) | 0.01 | 98.06 |
| Documents of law | "21st century standards" | 0.01 | 0.10 (1.98) | 0.09 (1.74) | 0.00 | 0.19 (2.35) | 0.31 (4.36) | 0.01 | 99.56 |
| Language | "English" | 0.03 | 0.25 (4.95) | 0.00 (0.00) | 0.04 | 0.24 (3.44) | 0.11 (2.22) | 0.02 | 99.75 |
| **Date** (spaCy) | "recent week" | 0.14* | 51.00 (76.94) | 32.14 (59.33) | 0.17* | 46.70 (62.24) | 30.43 (45.55) | 0.15* | 59.94 |
| Date (Stanford) | "January" | 0.13* | 24.66 (54.51) | 11.96 (37.21) | 0.14* | 14.00 (35.15) | 7.55 (22.69) | 0.11 | 83.40 |
| Time (spaCy) | "7:30 in the morning" | 0.09* | 27.10 (52.49) | 11.73 (31.47) | 0.18* | 30.63 (46.13) | 30.01 (47.40) | 0.01 | 70.19 |
| Time (Stanford) | "4:30 in the morning" | 0.08* | 6.62 (22.97) | 2.75 (14.55) | 0.10 | 8.59 (22.13) | 5.74 (19.13) | 0.07 | 90.38 |
| Percent (spaCy) | "100%" | 0.04 | 0.94 (10.96) | 0.40 (4.70) | 0.03 | 1.20 (10.14) | 0.49 (5.35) | 0.04 | 98.81 |
| Percent (Stanford) | "100%" | 0.03 | 0.67 (10.06) | 0.53 (5.33) | 0.01 | 0.85 (7.24) | 0.06 (1.13) | 0.08 | 99.13 |
| **Money** (spaCy) | "$15" | 0.23* | 28.56 (65.43) | 2.13 (13.88) | 0.28* | 22.20 (47.83) | 8.35 (28.63) | 0.18* | 84.13 |
| **Money** (Stanford) | $100 | 0.22* | 29.27 (66.53) | 2.17 (14.06) | 0.28* | 21.49 (47.34) | 9.19 (30.16) | 0.16* | 84.06 |
| Quantity | "13 inch" | 0.01 | 1.22 (8.28) | 2.22 (13.64) | 0.04 | 3.17 (14.13) | 1.47 (10.38) | 0.07 | 96.56 |
| **Ordinal** | "first" | 0.13* | 20.74 (43.49) | 9.93 (30.12) | 0.14* | 21.02 (41.26) | 12.66 (28.62) | 0.12* | 78.75 |
| **Cardinal** | "one" | 0.20* | 49.49 (77.47) | 31.38 (66.21) | 0.13* | 67.49 (86.82) | 29.72 (47.70) | 0.27* | 58.00 |

Means and SDs are multiplied by 100 for interpretability.
Sparsity = % of zero counts.
Negative *f*-values indicate larger values for deceptive than for truthful hotel reviews.
*Significant main effect of Veracity at $p < 0.002$ (Bonferroni-corrected: 0.05/25 comparisons = 0.002).

TABLE 3—*Means (SDs) and effect sizes (per veracity and valence) for LIWC categories with significant veracity effects.*

| | | | Positive | | | Negative | | |
|---|---|---|---|---|---|---|---|---|
| LIWC category | Explanation | *f* | Truthful | Deceptive | *f* | Truthful | Deceptive | *f* |
| Analytic | analytical thinking | 0.15 | 77.33 (17.17) | 70.56 (20.25) | 0.18 | 69.99 (17.58) | 65.66 (18.83) | 0.12 |
| Dic | dictionary words | 0.23 | 87.86 (4.62) | 89.49 (4.31) | 0.20 | 87.97 (3.77) | 89.87 (3.35) | 0.27 |
| function | total function words | 0.26 | 51.09 (5.62) | 53.61 (4.84) | 0.24 | 54.02 (4.42) | 56.40 (3.88) | 0.29 |
| pronoun | total pronouns | 0.26 | 9.46 (3.64) | 11.85 (3.92) | 0.31 | 11.72 (3.69) | 13.11 (3.50) | 0.19 |
| ppron | personal pronouns | 0.28 | 5.98 (2.89) | 8.13 (3.34) | 0.34 | 7.53 (3.19) | 8.86 (3.06) | 0.21 |
| i | 1st person singular (e.g., I, mine) | 0.36 | 2.51 (2.33) | 4.86 (3.27) | 0.41 | 3.28 (2.56) | 4.98 (3.13) | 0.30 |
| verb | common verbs | 0.20 | 13.82 (3.44) | 15.08 (3.44) | 0.18 | 15.61 (3.02) | 16.93 (2.87) | 0.22 |
| number | number (e.g., thousand, second) | 0.29 | 1.85 (1.82) | 0.92 (1.09) | 0.31 | 2.16 (1.57) | 1.43 (1.12) | 0.27 |
| AllPunc | all punctuation | 0.37 | 15.97 (5.91) | 12.39 (3.58) | 0.37 | 15.60 (5.38) | 12.25 (3.63) | 0.37 |
| Period | periods | 0.24 | 7.53 (3.56) | 6.14 (2.34) | 0.23 | 7.07 (3.44) | 5.71 (1.62) | 0.25 |
| Dash | dashes | 0.18 | 0.88 (1.49) | 0.44 (0.84) | 0.18 | 0.92 (1.51) | 0.48 (0.87) | 0.18 |
| Parenth | pairs of parentheses | 0.27 | 0.84 (1.38) | 0.21 (0.71) | 0.29 | 0.79 (1.28) | 0.27 (0.67) | 0.25 |
| OtherP | other punctuation | 0.30 | 0.70 (1.23) | 0.11 (0.44) | 0.32 | 0.52 (0.89) | 0.14 (0.40) | 0.28 |

The reported effect sizes are significant at the Bonferroni-corrected alpha level of 0.05/92 = 0.0005.

computer-automated computational linguistics approach. Specifically, we assessed how named entities differentiated between deceptive and truthful hotel reviews. The use of named entities was motivated by the theoretical principles that truthful statements are richer in detail, contain more contextual embeddings and more verifiable information. The aim for a computational operationalization of these theoretical lines was motivated by the need for empirically validated and scalable methods for purposes such as airport security settings. It was predicted that truthful statements would contain more references to specific information than false statements. Based upon these predictions, we used two named entity recognition algorithms to operationalize this prediction and to discriminate truthful from deceptive hotel reviews.

### *Named Entities for Verbal Deception Detection*

The results indicate support for our central hypothesis that there are more named entities in truthful than in deceptive hotel reviews. In addition, by not relying on fixed lexicons, the named entity recognition approach is better capable of identifying *unseen* (groups of) words than the LIWC. Named entity recognition also offers the categories that resemble the concept of richness and verifiability of detail and contextual embeddings. We tested our central hypothesis with two named entity recognition systems (spaCy and Stanford). Indeed, named entities performed better in discriminating truthful from deceptive hotel reviews than the LIWC richness of detail. However, this key finding was moderated by the choice of named entity recognition system: using spaCy we found strong support for our hypothesis and, furthermore, our results suggest that named entities were also better predictors than the average sentence specificity. This discrepancy might be due to the domain for which the speciteller tool was created, namely the specificity of news headlines. It might be that the speciteller approach is applicable for deception detection but simply did not include the principal dynamics of verbal deception theory.

The expected effect of named entities extracted with Stanford's NER system was smaller than with spaCy. That difference can be explained by the number of named entity categories of both tools. spaCy extracts named entities of 18 categories, whereas Stanford's tool only provides seven categories. More specifically, Stanford's NER system is primarily built for the identification of persons, organizations, and locations (https://nlp.stanford.edu/software/CRF-NER.shtml). Consequently, more named entities will be extracted with spaCy than with Stanford's tool. For example, among the best individual named entity categories in the present study were cardinals and ordinals which Stanford's tool does not recognize. Moreover, Stanford's tool recognized fewer named entities than spaCy (see Online Supplementary Material) and those entities that Stanford's tool did detect were classified differently than with spaCy. For example, "January 27th, 2009" was recognized as one date entity with spaCy but as two date entities ("27th January," "2009") with Stanford's tool. The latter capitalizes the named entities so that even mere mentions of a date appear as two distinct pieces of concrete information. Although this is not a drawback of Stanford's tool in general and does not diminish the accuracy of either tool, the chunking of, for example, date entities in spaCy's tool appears to be more suitable for deception detection. From a theoretical perspective, the more restrictive count of spaCy's named entity recognition "rewards" uniquely added information. The capitalizing count of Stanford's named entity recognition might benefit deceptive statements and the deceiver more as it overestimates the unique information provided through named entities. In general, the comparison of the two systems suggests (i) that both yielded similar results albeit of larger magnitude with spaCy's NER, (ii) that a higher number of entity categories is preferable to a smaller one, and (iii) that a more restrictive identification and chunking of named entities is better suited for deception detection than a liberal, potentially over-capitalizing one.

Exploratory analyses revealed that excluding sparse named entities increased the difference between truthful and deceptive reviews. The results corroborate the key role of highly specific information and show that some named entities play a more significant role in deception than others. The findings for individual categories of named entities reveal that truthful statements contained more dates, mentions of money, ordinals, cardinals, and person references (esp. when using Stanford's NER). These indications might open questions for future research on the psycholinguistic processes involved in deception (i.e., why are some named entities more important than others?) as well as on the accuracy of named entity recognition (i.e., how does the accuracy of named entity recognition affect deception detection?). Similar to the verifiability notion, deceptive reviews may have been affected by the deceiver's dilemma of not including that information that could potentially unmask their deceit (see [38]). One would further expect person references and location references to be related to the verifiability notion. The findings partly support this for overall deceptive/truthful reviews and positive reviews, but not for negative ones. References to persons and dates are aligned with the deceivers' information withholding strategy (i.e., avoiding potentially damaging information), but the predictive power of money references might be context-specific. It is rather typical to mention amounts of money for hotel reviews (e.g., for the good room price, the expensive cocktail) but less typical in other contexts (e.g., false testimonies, attitudes).

Finally, classifiers based on *n*-grams (i.e., the occurrence of frequent *n*-word units; e.g., "We met," "met yesterday," "yesterday at," are examples of bi-grams) were shown to achieve among the highest classification accuracies on the hotel review dataset (10). While the *n*-gram classifiers were purely data-driven, it is interesting to explore possible psychological dynamics underlying its successful classification. For instance, it could be that the predictive power of *n*-gram analyses might be driven by the same mechanism as the named entities, that is, that the *n*-grams that have the most power to differentiate the true from false reviews are in fact named entities. Therefore, future investigations could use our theoretical argument and test whether there is overlap between *n*-grams and named entities.

### *Named Entities Versus Lexicon Approaches*

The results from this current study allow for a comparison between named entity recognition and lexicon-based approaches for verbal deception detection. Both methodological approaches provide useful predictors of deception in hotel reviews. Although both pertain to surface features of a text, they differ in the core mechanisms through which they analyze the input text. The lexicon-based approach uses a database of words belonging to different categories (e.g., "perceptual processes"). All tokens (i.e., words and punctuation) in the text are allocated to the predefined categories, counted, and then standardized for document length. Named entity recognition, on the other hand, does not rely on lexicons but is primarily built on machine learning classifiers. Just as for the lexicon, tokens are allocated into named entity categories. However, rather than comparing each word with a database, the algorithm decides probabilistically for previously unseen tokens to which category they belong (i.e., “Janice” is most likely a person although “Janice” was never presented to the NER algorithm in a learning phase).

Both approaches have advantages and limitations. While the lexicon approach identifies 100% of the words that are in the connected database, it will fail to categorize any word that is not in the database. Therefore, it is highly sensitive to *unseen* words. Named entity recognition is flexible toward unseen words because it is based on rules that determine the probability of a word belonging to a class. However, this implies some misclassifications and nonclassifications.

This investigation indicates that named entities can grasp some concepts better than lexicon approaches. For example, both the LIWC and the named entities include the category “money.” Whereas the LIWC identifies words *related* to money (e.g., “budget,” “cash,” “underpaid”), the named entity recognition captures more *direct references* to money (e.g., “$15,” “$69/per,” “$240.43”). The latter has been shown to be a better predictor than the LIWC category. Similarly, the named entity recognition provides a distinction between ordinals (e.g., “second,” “sixth”) and cardinals (e.g., 8, 5, two), which are subsumed under the “number” LIWC category. The latter might obscure the nuances in the occurrence of references to numbers (i.e., here it seems to be driven by cardinals). The current investigation might function as an impetus toward the integration of named entities and lexicons in automatic verbal deception detection. Named entities seem to be a valuable addition to existing lexicon approaches.

## Limitations

### *Communication Context and Discourse*

Although the data support the proposed use of named entities, certain shortcomings merit attention.

The findings obtained here might be unique to the domain of hotel reviews. For example, deception detection could be more domain-dependent than previously assumed. That is, results found in area A (e.g., criminal intentions) do not necessarily apply to area B (e.g., lying about political preferences) (12). It is important to acknowledge the question of the discourse. Communication in general, and deceptive communication specifically, can occur on diverse topics (e.g., attitudes, testimonies, opinions), in diverse formats (e.g., free narrative, interrogation, computer-mediated communication), on diverse temporal dimensions (e.g., present, past, future events), in different production modes (e.g., spoken, hand-written, typed), and in various lengths (e.g., essays, brief reviews, yes/no answers), to name but a few (5). One deception theory that appreciates the role of context is Levine’s Truth-Default-Theory (TDT) (39). According to TDT knowing the communication context is often a precondition for deception detection as it offers baselines of similar situations against which a verbal account can be compared. For example, the occurrence of money references is rather typical in hotel reviews but less common in opinions about abortion or the death penalty. Consequently, the discourse and context of verbal accounts might be one of the key moderators of deception detection models.

Contrary to approaches that rely on data-driven insights rather than theory-led analysis, we tried to stay close to the theoretical lines of content-based analysis tools. Moreover, we adopted the deceivers’ dilemma put forward by Nahari et al. (26) and incorporated named entities for the reason that they are likely to mirror the concept of richness and verifiability of detail as well as contextual embeddings. Psychological processes such as the specificity of a memory trace and the strategic avoidance of potentially verifiable information (i.e., for deceivers) might contribute to the generalizability of the current findings. Consequently, we would expect to find more named entities in truthful than in false statements in various other contexts as well. For example, the richness of detail was a valid predictor of veracity for children as well as adults, for victims as well as perpetrators, and for real sex crimes as well as innocuous laboratory games (24). In the current investigation, we found that they generalize to hotel reviews. Future research would need to further assess domain-specificity.

### *Accuracy of Named Entity Recognition Systems*

Despite the speed and reliability of the named entity operationalization, it is not perfect. The example in Box 1 hints at inaccuracies in the named entity recognition system: some entities are misclassified (e.g., an organization is classified as a person) and other entities are not recognized at all. Jiang et al. (30) found that NER accuracy is highly dependent on the category. When evaluating four widely used named entity recognition tools on the identification accuracies for persons, organizations, and locations, they found that spaCy’s NER performs well in identifying persons (recall: 73.25%; precision: 72.86%) but poorly for organizations (recall: 28.73%; precision: 33.46%). The latter implies that organizations are often misclassified or not recognized. A way to address the accuracy problem is to train NER systems with a supervised machine learning task. Although higher NER accuracy is desirable, we believe that this has not affected the predictive power of information specificity because truthful and deceptive reviews were examined with the same (in)accurate NER system. In fact, our data support the notion that accuracy is less important than the number of named entity categories. When we used Stanford’s NER system, the named entities were not as a good a veracity predictor as with spaCy. Future research could explore how an increased accuracy affects concepts such as the information specificity.

### *Other Psycholinguistic Cues to Deception*

The broad application of richness of detail, contextual embeddings and the verifiability of details aside, these are only some variables that content-based analysis tools use to tell truthful from deceptive statements. For example, the plausibility of a statement is often found to be a valid discriminator between truthful and false statements (40). However, automated approaches are not easily applied to semantic constructs like the plausibility or logical structure of a statement. Arguably, these concepts would need to be modeled differently with more advanced word representations such as *word embeddings*. Word embeddings represent words as real vectors in a dense and low-dimensional vector space and are seen as a suitable method for measuring semantic relationships between different words (41). Word embeddings learned by artificial neural networks seem to be a promising method to capture syntactic and semantic regularities in language (42).

### *Human Versus Computer-Based Deception Detection*

Some skepticism exists toward automated approaches of verbal deception detection (43). Content-based analysis tools rely heavily on interpretation and context of phrases within the statement. Human-coders are in general thought of being better able to grasp the meaning of information in context than algorithms do. The argument is that computers fail to pick up the subtleties needed for detecting deception in verbal statements. However, this criticism warrants relativization. First, accuracy rates of computer-automated detection—predominantly through machine learning classification—are equal to or better than those of human-coded analysis (5,23). Ott et al. (10,11) found human judges to be less accurate than computerized analysis, findings that the current investigation supports. As the human judges in Ott et al.'s studies were not applying any coding strategy (e.g., scoring the plausibility or richness of detail), a direct comparison of trained human annotation and automated analysis is needed to examine the human vs. machine issue. Second, the current investigation, as well as the study by Bond and Lee (33), suggest that concepts motivated by content-based tools can indeed be modeled with computational methods. Future research in verbal deception detection should further test out the boundaries of computer-automated approaches to make this promising strand of deception research more applicable. Experimental studies in the future might want to focus on methodological elements of the coding procedure (human vs. machine) on dimensions such as the speed of judgments, the reliability of assessments, and the applicability of both approaches on a large scale.

## Conclusions

This study showed that named entities can be used to model variables useful for verbal deception detection. Based on verbal deception theory, named entities offer a viable addition to the psycholinguistic features used in computer-automated verbal deception detection. We encourage others to explore named entities for verbal deception detection further and to work on the synthesis of verbal deception detection theory and computational linguistics.

## References

1. Honts C, Hartwig M. Credibility assessments at portals. In: Raskin DC, Honts CR, Kircher JC, editors. Credibility assessment: scientific research and applications. San Diego, CA: Academic Press, 2014.
2. Ormerod TC, Dando CJ. Finding a needle in a haystack: toward a psychologically informed method for aviation security screening. J Exp Psychol Gen 2015;144(1):76–84.
3. Weinberger S. Airport security: intent to deceive? Nature 2010;465 (7297):412–5.
4. Oberlader VA, Naefgen C, Koppehele-Goseel J, Quinten L, Banse R, Schmidt AF. Validity of content-based techniques to distinguish true and fabricated statements: a meta-analysis. Law Hum Behav 2016;40(4):440–57.
5. Hauch V, Blandón-Gitlin I, Masip J, Sporer SL. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Pers Soc Psychol Rev 2015;19(4):307–42.
6. Nahari G. When the long road is the shortcut: a comparison between two coding methods for content-based lie-detection tools. Psychol Crime Law 2016;22(10):1000–14.
7. Vrij A. Criteria-based content analysis: a qualitative review of the first 37 studies. Psychol Public Policy Law 2005;11(1):3–41.
8. Fitzpatrick E, Bachenko J, Fornaciari T. Automatic detection of verbal deception. San Rafael, CA: Morgan & Claypool Publishers, 2015.
9. Newman ML, Pennebaker JW, Berry DS, Richards JM. Lying words: predicting deception from linguistic styles. Pers Soc Psychol Bull 2003;29(5):665–75.
10. Ott M, Cardie C, Hancock JT. Negative deceptive opinion spam. In: Proceedings of 2013 NAACL-HLT Conference; 2013 June 9-14; Atlanta, GA. Stroudsburg, PA: Association for Computational Linguistics, 2013;497–501.
11. Ott M, Choi Y, Cardie C, Hancock JT. Finding deceptive opinion spam by any stretch of the imagination. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies – Volume 1; 2011 June 19-24; Portland, OR. Stroudsburg, PA: Association for Computational Linguistics, 2011;309–19.
12. Mihalcea R, Strapparava C. The lie detector: explorations in the automatic recognition of deceptive language. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers; 2009 Aug 2-7; Suntec, Singapore. Stroudsburg, PA: Association for Computational Linguistics, 2009;309–12.
13. Bachenko J, Fitzpatrick E, Schonwetter M. Verification and implementation of language-based deception indicators in civil and criminal narratives. In: Proceedings of the 22nd International Conference on Computational Linguistics – Volume 1; 2008 Aug 18-22; Manchester, United Kingdom. Stroudsburg, PA: Association for Computational Linguistics, 2008;41–8.
14. Fornaciari T, Poesio M. Automatic deception detection in Italian court cases. Artif Intell Law 2013 Sep;21(3):303–40.
15. Markowitz DM, Hancock JT. Linguistic obfuscation in fraudulent science. J Lang Soc Psychol 2016;35(4):435–45.
16. Zhou L, Burgoon JK, Nunamaker JF, Twitchell D. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group Decis Negot 2004 Jan;13(1):81–106.
17. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investig 2007;30(1):3–26.
18. Grishman R, Sundheim B. Message understanding conference-6: a brief history. In: Proceedings of the 16th Conference on Computational Linguistics – Volume 1; 1996 Aug 5-9; Copenhagen, Denmark. Stroudsburg, PA: Association for Computational Linguistics, 1996;466–71.
19. Nothman J, Ringland N, Radford W, Murphy T, Curran JR. Learning multilingual named entity recognition from Wikipedia. Artif Intell 2013;194:151–75.
20. Honnibal M. SpaCy, 2016; https://spacy.io/(accessed January 20, 2017).
21. Nahari G, Vrij A. Are you as good as me at telling a story? Individual differences in interpersonal reality monitoring. Psychol Crime Law 2014;20(6):573–83.
22. Johnson MK, Raye CL. Reality monitoring. Psychol Rev 1981;88(1):67–85.
23. Masip J, Sporer SL, Garrido E, Herrero C. The detection of deception with the reality monitoring approach: a review of the empirical evidence. Psychol Crime Law 2005;11(1):99–122.
24. Vrij A. Verbal lie detection tools: statement validity analysis, reality monitoring and scientific content analysis. In: Granhag PA, Vrij A, Vershuere B, editors. Detecting deception: current challenges and cognitive approaches. Chichester, West Sussex, UK: John Wiley & Sons Ltd, 2015;3–35.
25. Köhnken G. Statement validity analysis and the 'detection of the truth'. In: Granhag PA, Stromwall LA, editors. The detection of deception in forensic contexts. Cambridge, UK: Cambridge University Press, 2004;41–63.
26. Nahari G, Vrij A, Fisher RP. Exploiting liars' verbal strategies by examining the verifiability of details. Leg Criminol Psychol 2014 Sep;19 (2):227–39.
27. Harvey AC, Vrij A, Nahari G, Ludwig K. Applying the verifiability approach to insurance claims settings: exploring the effect of the information protocol. Leg Criminol Psychol 2017;22(1):47–59.
28. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015, 2015; https://repositories.lib.utexas.edu/handle/2152/31333 (accessed January 5, 2017).
29. Li JJ, Nenkova A. Fast and accurate prediction of sentence specificity. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence; 2015 Jan 25-30; Austin, TX. Palo Alto, CA: AAAI Press, 2015;2281–7.
30. Jiang R, Banchs RE, Li H. Evaluating and combining named entity recognition systems. In: Proceedings of the Sixth Named Entity

Workshop, Joint with 54th Association for Computational Linguistics; 2016 Aug 12; Berlin, Germany. Stroudsburg, PA: Association for Computational Linguistics, 2016;21–7.

31. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics; 2005 June 25-30; Ann Arbor, MI. Stroudsburg, PA: Association for Computational Linguistics, 2005;363–70.
32. Bird S. NLTK: the natural language toolkit. In: Proceedings of the COLING/ACL on Interactive Presentation Sessions; 2006 July 17-21; Sydney, Australia. Stroudsburg, PA: Association for Computational Linguistics, 2006;69–72.
33. Bond GD, Lee AY. Language of lies in prison: Linguistic classification of prisoners' truthful and deceptive natural language. Appl Cogn Psychol 2005;19(3):313–29.
34. Cohen J. Statistical power analysis for the behavioral sciences. New York, NY: Academic Press, 1988.
35. National Research Council, Committee to Review the Scientific Evidence on the Polygraph, Division of Behavioral and Social Sciences and Education. The polygraph and lie detection. Washington, DC: The National Academies Press, 2003.
36. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 2011;12(1):77.
37. Venkatraman ES. A permutation test to compare receiver operating characteristic curves. Biometrics 2000;56(4):1134–8.
38. Kleinberg B, Nahari G, Verschuere B. Using the verifiability of details as a test of deception: a conceptual framework for the automation of the verifiability approach. In: Proceedings of the 2016 NAACL-HLT Conference; 2016 June 12-17; San Diego, CA. Stroudsburg, PA: Association for Computational Linguistics, 2016;18–25.
39. Levine TR. Truth-Default Theory (TDT): a theory of human deception and deception detection. J Lang Soc Psychol 2014 Sep;33(4):378–92.
40. Leal S, Vrij A, Warmelink L, Vernham Z, Fisher RP. You cannot hide your telephone lies: providing a model statement as an aid to detect deception in insurance telephone calls. Leg Criminol Psychol 2015;20 (1):129–46.
41. Fu R, Guo J, Qin B, Che W, Wang H, Liu T. Learning semantic hierarchies via word embeddings. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014 June 23-25; Baltimore, MD. Stroudsburg, PA: Association for Computational Linguistics, 2014;1199–209.
42. Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 NAACL-HLT Conference; 2013 June 9-14; Atlanta, GA. Stroudsburg, PA: Association for Computational Linguistics, 2013;746–51.
43. Vrij A. Reality monitoring. In: Vrij A, editor. Detecting lies and deceit: pitfalls and opportunities, 2nd edn. Chichester, West Sussex, UK: John Wiley & Sons, 2008.

Additional information and reprint requests:
Bennett Kleinberg, M.Sc.
Department of Psychology
University of Amsterdam
Nieuwe Achtergracht 129 D
1018 WS Amsterdam
The Netherlands
E-mail: b.a.r.kleinberg@uva.nl