Detecting Deception through Linguistic Cues: from Reality Monitoring to Natural Language Processing

Riccardo Loconte [1*], Chiara Battaglini[2+], Stéphanie Maldera[3+], Pietro Pietrini[1], Giuseppe Sartori [3], Nicolò Navarin [4], Merylin Monaro [3]

+ **Equal contribution**

Email addresses: riccardo.loconte@imtlucca.it (R. Loconte)*, chiara.battaglini@iusspavia.it (C. Battaglini), stephanie.maldera@studenti.unipd.it (S. Maldera), pietro.pietrini@imtlucca.it (P. Pietrini), giuseppe.sartori@unipd.it (G. Sartori), nicolo.navarin@unipd.it (N. Navarin), merylin.monaro@unipd.it (M. Monaro)

[1] Molecular Mind Lab, IMT School of Advanced Studies Lucca, Lucca, Italy

[2] Neurolinguistics and Experimental Pragmatics (NEP) Lab, Department of Humanities and Life Sciences, University School for Advanced Studies IUSS, Pavia, Italy

[3] Department of General Psychology, University of Padova, Padova, Italy

[4] Department of Mathematics "Tullio Levi-Civita", University of Padova

*** Corresponding author. IMT School of Advanced Studies Lucca, Piazza San Francesco 19, Lucca (LU) 55100. Email: riccardo.loconte@imtlucca.it**

Abstract

Detecting deception in interpersonal communication is a pivotal issue in social psychology, with significant implications for court and criminal proceedings. In this study, four experiments were designed to compare the performance of natural language processing (NLP) techniques and human judges in detecting deception from linguistic cues in a dataset of 62 transcriptions of video-taped interviews (32 genuine and 30 deceptive). The results showed that machine-learning (ML) algorithms significantly outperform naïve (accuracy=54.7%) and expert judges (accuracy=59.4%) when trained on features from the reality monitoring (RM) and cognitive load (CL) frameworks (accuracy=69.4%) or on features automatically extracted through NLP techniques (accuracy=77.3%) but not when trained on the RM criteria alone. This evidence suggests that NLP algorithms, due to their ability to handle complex patterns of linguistic data, might be useful for better disentangling truthful from deceptive narratives, outperforming traditional theoretical models.

Keywords: deception, Reality Monitoring, Natural Language Processing, lie detection, deception linguistic cues

1. Introduction

Detecting deception in interpersonal communication is a pivotal issue in social psychology, with significant implications for criminal investigations and court and criminal proceedings. For example, assessing the credibility of a suspect during interrogation is relevant, as false or distorted information may lead the investigation in the wrong direction. Moreover, in some cases, such as in the Italian court, allegations of sexual harassment are often based on the victims' declarations, which are treated as evidence and most often without clear and independent evidence of the offence. In such situations, a police officer or a judge is tasked with determining the veracity of the information provided and of the alleged accusation (Oberlader et al., 2021). However, humans exhibit inherent biases when it comes to detecting lies. In the absence of any prior knowledge of the context, human intuitive judgment in deception detection has been proven to be just slightly above the chance level (Bond & DePaulo, 2006; Ekman & O'Sullivan, 1991). Even experts in the field, such as police officers, tend to commit false-negative and false-positive errors (Elaad, 2009; Vrij, 2008). For naïve judges, the truth bias (i.e., the intrinsic human inclination to presume others as honest) has been proposed as one possible explanation for this poor performance (Levine, 2014; Street & Masip, 2015). For expert judges, instead, the debate is still open on whether the problem stems from the identification of the cue (i.e., during the evaluation process) or the difficulty of combining several cues (due to the limited cognitive resources) to then make a straightforward decision (Verschuere et al., 2023).

Recently, researchers have increasingly relied on automated approaches to deception detection based on machine learning (ML) techniques, i.e., computational methods that enable computer algorithms to identify patterns in data and make predictions based on these patterns (see Constâncio et al., 2023 for a review). One of the most exploited techniques is natural language processing (NLP), a field of artificial intelligence (AI) focused on enabling machines to interpret, analyze, and respond to human language. NLP has been applied especially in detecting deception online, such as identifying fake reviews or misinformation (Alonso et al., 2021; Salminen et al., 2022). Typically, NLP-based approaches are heavily data-driven. This means that they rely on extracting features directly from textual data, such as word frequencies, syntactic patterns, and word embeddings (i.e., numerical representations of word co-occurrences), without necessarily incorporating insights from psychological theories that have been used to identify cues of deception in language.

In this context, NLP techniques can be leveraged to transform textual data into numerical features based on theoretical frameworks that will then be used to feed ML models trained to identify subtle verbal indicators in datasets where truthful and deceptive examples are already labelled (i.e., supervised learning). The advantage of this approach is that, after the training is complete, a good ML model can be used to predict whether new statements are likely deceptive or truthful based on learned patterns, and evaluate the efficacy of a specific psychological theory of deception.

1.1 Investigating the veracity of verbal content using reality monitoring

Deception may imply reporting fabricated details or intentionally omitting relevant information conveyed in such a way as to seem truthful to the interlocutor (Newman et al., 2003). The Undeutsch hypothesis suggests that deceitful information is qualitatively different in form and content from truthful information (Vrij et al., 2000). Nevertheless, there is no verbal cue that is inherently associated with deception (Vrij, 2008).

Among the several verbal lie detection tools, reality monitoring (RM; Johnson & Raye, 1981) was developed on the basis of evidence from cognitive psychology and currently stands out in the literature for its theoretical robustness, being the most commonly employed approach by researchers. This approach is grounded in the notion that memories of actual experiences exhibit stronger connections to external stimuli than memories of imagined events. Accordingly, memories originating from perceptual experiences should feature contextual, sensory, and affective details, whereas internally generated memories, stemming from thoughts or imagination should be marked by references to cognitive processes. Eight criteria were outlined to distinguish between these two types of memories, with the presence of cognitive operations being the only lie criterion (Vrij et al., 2007). Research indicates that when scores are combined from these eight criteria, the average accuracy RM rate is comparable to that of other content-based techniques (e.g., the criteria-based content analysis), ranging between 61% and 83%, with an average of 69% (Vrij, 2008). Among the individual criteria, contextual factors, such as temporal and spatial criteria, appear to hold the greatest diagnostic value (Masip et al., 2005).

Systematic reviews (Masip et al., 2005; Vrij, 2005, 2008) and meta-analyses (Amado et al., 2015, 2016; Hauch et al., 2017; Oberlader et al., 2016) have demonstrated that RM exhibits satisfactory inter-rater reliability and empirical validity across diverse study designs and populations, possibly because of its straightforward application (Sporer, 1997; Strömwall et al., 2004; Vrij et al., 2000). Indeed, RM is not time-consuming, involves less subjective decision-making (Oberlader et al., 2016), and holds precise criteria that are easy to operationalize. Despite these findings, caution is advised when using RM due to the lack of an objective decision rule, namely a numerical cutoff for scores that allow the user to classify a narrative as honest or faked (Amado et al., 2015, 2016).

1.2 Investigating the veracity of verbal content by imposing cognitive load

In interviews, lie-tellers are known to consume more cognitive resources because they need to fabricate responses that are congruent with other fabricated information while maintaining their credibility during the examination (Vrij et al., 2008). This cognitive load (CL) is often reflected in several indices (e.g., behavioural, physiological, verbal, and neural, among others) that can be leveraged to distinguish truthful from deceptive statements (Vrij et al., 2008).

The “imposing cognitive load approach” (Vrij et al., 2017) takes advantage of this vulnerability in lie-tellers and leverages interviewing strategies that have the effect of further depleting their cognitive resources while keeping the demand manageable for truth-tellers (Walczyk et al., 2013). These strategies may involve asking the examinee to perform a second task during the interview or to continuously switch between two tasks, imposing time restrictions to answer questions,

recalling the events in reverse order, or asking the examinee unexpected questions (Walczyk et al., 2013; Vrij et al., 2009).
Among these, the strategy of asking unexpected questions has proved to be effective, reaching accuracy rates from 80% to 95% (Monaro et al., 2018; Monaro et al., 2017; Lancaster et al., 2013). It involves the examiner initially asking anticipated questions, i.e., questions that the lie-tellers expect and prepare in advance, and then switching to questions that cannot be foreseen and for which the responses were not prepared. For example, in the context of faked identities, lie-tellers may prepare answers about the name, surname, and date of birth of the stolen identity, but they unlikely prepare the answer for their zodiac sign (Monaro et al., 2017). Responses to unexpected questions in lie-tellers are associated with slower reaction times and a higher number of inconsistencies than truth-tellers (Melis et al., 2024; Monaro et al., 2017). Lie detection approaches based on imposing cognitive load produce higher accuracy rates compared to standard approaches in spotting deception (Vrij et al., 2017).

1.3 Investigating the veracity of verbal content using natural language processing

The advent of NLP for the analysis of human language has provided new opportunities for the automatic detection of deception (Tomas et al., 2022). These techniques allow to extract features at different level of granularity: - the *n*-gram model breaks the text into linear sequences of tokens and reveal frequent patterns; - part-of-speech (POS) tagging assigns grammatical categories (nouns, verbs, and adjectives) to words, thus informing on shallow syntactic text structure; - word/sentence length and number are extracted to evaluate text complexity; - the Linguistic Inquiry and Word Count (LIWC) categorizes words into psychological, social, and emotional dimensions (e.g., positive/negative affect, social words, etc) providing the semantic content of the text; - named-entity recognition (NER), automatically identifies and categorizes proper nouns (such as names of people, places, and organizations) within texts.

Moreover, these computational techniques can be applied with different methodologies. Data-driven approaches are based on statistical procedures to perform feature extraction and selection. Theory-driven approaches investigate samples of features derived from psychological theoretical models of deception. Last, hybrid models rely on theory to select variables that are restricted to those that are found to be statistically significant.

The RM and CL frameworks have been proven suitable for investigation using this computational perspective. Recent studies highlighted how manual coding is not necessarily superior to automated coding of RM (Schutte et al. 2021, preprint; Deeb et al., 2024). In addition, a meta-analysis by Hauch et al. (2015) demonstrated RM effectiveness in detecting verbal deception when using LIWC features. The same meta-analysis also found evidence in favor of the CL theory, showing that lie-tellers produce shorter, less elaborate, and less complex statements (Hauch et al., 2015). These characteristics can be automatically captured in statements through linguistic features such as the number of words, number of sentences, average sentence length, type-token ratio, word

length, and use of exclusive words (e.g., but, except, without). Furthermore, a recent study found that verbal cues of CL effectively distinguished truthful from deceptive statements in a mixed dataset that included different contexts of deception (i.e., personal opinions vs. autobiographical memories vs. future intentions), suggesting the potential for these features to serve as more generalizable cues compared to others (Loconte et al., 2023).

Parallel to these theory-driven approaches, the detection of verbal deception has also been investigated using a data-driven approach. For example, Mihalcea & Strapparava (2009) and Ott et al., (2011) detected deceptive opinions by training a naïve Bayes and a support vector machine model on n-grams and a combination of n-grams and LIWC features, respectively. Pérez-Rosas et al. (2015) investigated an open-domain dataset using n-grams, shallow and deep syntactic features (using POS tagging), semantic features from LIWC, and readability and syntactic complexity metrics. Finally, Kleinberg et al. (2018a, 2018b) employed NER as a proxy for automated scoring of details and verifiable details and accurately classified truthful and deceptive hotel reviews and future intentions.

A recent review of studies from Constâncio et al. (2023) on ML and NLP techniques for lie detection showed that automatic verbal analysis significantly outperformed the chance and human level in a variety of datasets (see Table 1). Furthermore, a short review of the literature on the use of LIWC software for lie detection (Van Der Zee et al., 2022) showed that studies using a data-driven approach reached an accuracy rate ranging from 65% to 74%, whereas studies using a theory-driven approach reached an accuracy rate ranging from 51% to 69%. Studies that employed a hybrid approach, training ML models only on statistically significant theoretical variables, reached an accuracy rate ranging from 50% to 69%. These findings showed that selecting the most appropriate approach is paramount because it may influence ML models' performance (Van Der Zee et al., 2022).

Table 1.
Studies employing ML techniques on verbal content for lie-detection tasks reported in Constâncio et al. (2023).

Authors	Dataset	Features	Algorithm	Accuracy
Rubin and Conroy (2011)	Stories written by volunteers	LIWC categories, Lexical measures	Support Vector Machine (SVM)	0.65
Fentg et al. (2012)	Reviews of 35 Italian restaurants	Bigrams, POS tags, Syntax complexity, Unigrams	SVM	0.91
Fornaciari et al. (2012)	DeCour corpus	LIWC categories, Lexical measures, N-grams, POS tags	SVM	0.69
Rubin and Conroy (2012)	Stories written by volunteers	LIWC categories, Lexical measures	Decision Tree	0.65

Authors	Dataset	Features	Algorithm	Accuracy
Perez-Rosas et al. (2013)	Video recordings from volunteers	Unigrams	SVM	0.74
Briscoe et al. (2014)	Statements provided by volunteers in a mock chat room	Emoticons, Informality, Sentiment, Syntax complexity	Gradient Boosting	0.91
Pak and Zhou (2015)	Communications during sessions of the online Mafia game	LIWC categories, Syntax complexity, Unigrams	Decision Tree	0.98
Kleinberg et al. (2018)	Interviews on weekend plans collected from volunteers	LIWC categories, Named entities, Psychological processes	SVM	0.77
Mbaziira et. al (2018)	Combination of four publicly available datasets	Syntax complexity	Neural Network	0.80
Barsever et al. (2020)	Ott Deceptive Opinion Spam Corpus	BERT embeddings	Neural Network	0.94
Kleinberg et al. (2021)	True and deceptive statements collected by a web application from volunteers	LIWC categories, POS tags	Random Forest	0.69

Note. The type of dataset, the features extracted, the algorithm employed and the highest accuracy reached in the test set are reported as they appear in the original study, rounded to two digital places.

1.2 The current study

This study examines deception detection through the theoretical lenses of RM and CL in the context of transcripts of interviews with unexpected questions. Through four experiments, this study compared the performances of naïve judges, expert judges trained on RM, and both theory-driven and data-driven ML models in detecting deception, offering critical insights into the relative effectiveness and reliability of each approach.

Specifically, in **Experiment 1,** we compared the performance of laypeople (i.e., naïve judges) to that of individuals with expertise in the forensic field (i.e., expert judges). Expert judges were trained in the application of RM criteria for lie detection. The main hypothesis (Hyp. 1a) postulates expert judges achieve higher accuracy than naïve judges because studies have demonstrated RM criteria's effectiveness in distinguishing truth from deception in verbal content (Vrij et al., 2022; Gancedo et al., 2021; Amado et al., 2016; Vrij, 2008).

However, we saw expert judges performing poorly in Experiment 1. Therefore, in **Experiment 2,** we aimed to investigate the reasons behind this poor performance, trying to disentangle whether it was associated with limitations in the effectiveness of RM when applied to specific datasets or whether the problem relied on experts' limitations associated with their evaluation and decision-

making skills. We employed a computational approach for this purpose. Specifically, four ML models were trained on two sets of RM ratings: those assigned to each transcription by expert judges in Experiment 1 and those provided by leveraging NLP techniques. Three alternative hypotheses were defined for this study:

Hyp. 2a) Expert judges performed poorly in Experiment 1 because they were poor evaluators. Specifically, expert judges may not effectively assess RM criteria. If this hypothesis is true, a poor performance is expected from ML models trained on expert ratings, similar to that obtained by forensic experts in Experiment 1. Moreover, they are expected to show lower accuracy than those trained on ratings derived through NLP techniques.

Hyp. 2b) Expert judges performed poorly in Experiment 1 because RM criteria were poorly informative for this type of dataset. If this hypothesis is true, ML models trained on expert and NLP-based ratings of RM are expected to underperform, with an accuracy similar to that obtained by forensic experts in Experiment 1.

Hyp. 2c) The RM criteria were informative, and expert judges were effective evaluators but may have struggled in effectively combining all the information to derive a final decision. If this hypothesis is true, ML models trained on either expert or NLP-based ratings of RM are expected to outperform the accuracy obtained by expert judges in Experiment 1.

Finally, Experiments 3 and 4 concerned ML models' performance. Specifically, we examined the effectiveness of theory-driven vs. data-driven approaches in feature extraction.

Experiment 3 employed linguistic features from two theoretical frameworks: RM and CL. This procedure was adopted for two reasons: i) a previous meta-analysis on deception detection indicated that the CL approach yields higher accuracy rates than standard approaches (Vrij et al., 2017); ii) the dataset under analysis was specifically designed to increase CL in lie-tellers by posing unexpected questions (Monaro et al., 2020).

The first hypothesis for this experiment posits that adding linguistic features associated with the CL framework potentially increases ML models' accuracy in detecting verbal deception compared to models trained solely on features from RM (Hyp. 3a).

The second hypothesis postulates that various interview phases (free recall vs. unexpected questions vs. full text) influence ML models' performance. Specifically, ML models trained on features extracted from unexpected questions (or full text) will yield higher accuracy than those trained on features extracted solely from free speech, based on the assumption that CL features are more prevalent in responses to unexpected questions (Hyp. 3b).

Experiment 4 was designed to explore whether a data-driven approach could surpass the performance of the previous theory-based methods. To this end, NLP techniques to extract a broad set of linguistic features and a data-driven feature selection strategy (Chandrill, 2022) were employed to identify a subset of highly informative features. The main hypothesis for this study postulates that a data-driven approach might outperform classical theory-driven approaches, particularly in scenarios in which theory-based methods have already shown limited effectiveness.

Specifically, a data-driven method is hypothesized to achieve superior performance by extrapolating rules directly from data rather than relying on general theories (Hyp. 4a).

Experiments 1 to 4 altogether allow us to test a final hypothesis for this study, which posits that theory-driven and data-driven ML approaches are expected to outperform naïve and expert human judges in identifying deception (Hyp. 4b). This hypothesis stems from ML models' computational ability to integrate and analyze complex datasets more comprehensively than humans.

2. Experiment 1: Naïve vs Expert judges

2.1 Methods and Materials

2.1.1 Participants

The sample size was determined through an a priori power analysis using G*Power (Faul et al., 2007). For the sample of naïve judges, the results indicated that a sample size of 42 is sufficiently large to achieve a statistical power (1-β) = 0.8 in a one-tailed Wilcoxon signed-rank test (one sample case), given a significance level α = 0.05 and a medium effect size (d = 0.4; Bond & DePaulo, 2006). Since we had access to a larger sample size, we collected a significantly higher number of participants, establishing a minimum of 10 judges per statement, replicating the original recruiting and evaluation procedure as in Monaro et al., (2022). Therefore, the sample of naïve judges consisted of 121 Italian-speaking participants (75 females). Age ranged from 18 to 62 years (M = 33.92, SD = 12.40), with years of education ranging from 8 to 21 (M = 15.84, SD = 2.40). Participants were volunteers recruited following a snowball sampling procedure. One participant did not complete the task and was excluded from the analysis.

For the sample of expert judges, the results of the power analysis indicated that a sample size of 23 is sufficiently large to achieve a statistical power (1-β) = 0.8 in a one-tailed Wilcoxon signed-rank test (one sample case), given a significance level α = 0.05 and a medium effect size (d = 0.55; Gangedo et al., 2021). Since having only 23 participants would result in fewer than two expert judgments per stimulus, we decided to expand the sample size to ensure at least three judges for each statement. Therefore, the sample of expert judges resulted in 36 Italian-speaking participants (27 females). Experts were recruited among psychology students attending the Master's course in Forensic Psychology and among the authors' professional network. Participation in the study was on a voluntary basis. Age ranged from 22 to 55 years (M = 30.17, SD = 7.77), with years of education ranging from 13 to 21 (M = 18.86, SD = 2.18). Nineteen experts were psychology students, and 17 were practitioners (i.e., psychologists, psychotherapists, psychiatrists, and lawyers). Experts were also asked to specify their level of expertise in forensic psychology via a multiple-choice question: 11.1% had only completed a course in forensic psychology, 27.8% had

undergone additional training in the field, 47.2% held a master's degree in forensic psychology, and the remaining 13.9% were currently employed in the field.

All participants provided their informed consent before starting the experiment. The Board of the School of Psychology, University of Padova, approved the experimental procedure.

2.1.2 Dataset

The dataset consisted of 62 videotaped interviews of Italian participants (43 females, age range 20-29, who voluntarily participated in the study) in a low-stake scenario (recalling a holiday experience). The dataset was collected by Monaro et al. (2020) in a previous study and was analyzed to detect deception through blink rate (Monaro et al., 2020) and facial expressions (Monaro et al., 2022).
The dataset comprised 32 participants allocated to the truthful condition and instructed to describe a real holiday experience that occurred within the preceeding 12 to 18 months. Thirty participants were assigned to the deceptive condition and were required to describe an entirely fabricated holiday. Each videotaped interview comprised three distinct phases:

1. **Baseline**, in which the interviewee provided their autobiographical information
2. **Free speech**, in which the interviewee freely recalled their holiday experience for approximately two minutes
3. **Unexpected Questions**, in which the interviewer asked unexpected questions about the holiday experience to increase the interviewee's CL (e.g., "Did a particular event occur during the holiday that made it necessary to revise the initial plans?").

The average length of the videos was 9.56 min. A more detailed description of the dataset is reported in Monaro et al., 2020 and in the Supplementary Material.

2.1.3 Narratives transcription procedure

Interviews were manually transcribed verbatim. Then, a linguistic expert checked and modified raw transcriptions, following the guidelines in CLIPS (Savy, 2006). Adaptations tailored for the present study were made to ensure readability for the naïve readers. Regionalisms were substituted with the standard Italian alternative. Hesitations, false starts, and repetitions were transcribed as accurately as possible. False starts were reported with the symbol +, following Savy (2006): "abbiamo spos+, cioè abbiamo trovato" ("we have mov+, I mean we have found"). Pauses were signaled using punctuation. Sentence boundaries, signaled with a full stop and commas, were inserted using the standard Italian rules for punctuation. Hesitations and laughter were reported using standardized formulas (see Table 1S in Supplementary Material).

2.1.4 Reality monitoring scoring

Following Sporer (1997, 2004), the RM framework consisted of eight criteria, as outlined in Table 2. Previous research employed three primary units of measurement to evaluate a statement

according to the RM criteria: rating scales, categorical measures (presence vs. absence), and frequency/density counts (see Gangedo et al., 2021 for a meta-analytical review on the use of RM in the forensic context). For this experiment, RM criteria were evaluated using a 7-point rating scale (1=none, 7=very much). Previous findings showed that frequency counts may offer better performance and reliability than rating scales (Nahari, 2016). However, rating scales present other advantages: they require less training for human raters, are quick to apply, and provide a clear minimum and maximum score, helping the raters understand whether the obtained score falls within a higher or lower range. This is helpful, especially considering that the RM approach does not rely on a standardized cutoff to finally evaluate an account as truthful or deceptive (Amado et al., 2015, 2016). Moreover, using rating scales (or categorical measures) ensures consistency across all RM criteria. Indeed, when we use relative or absolute frequencies, only five out of eight criteria can be scored by frequency (i.e., perceptual information, spatial information, temporal information, affective information, and cognitive operations) whereas three criteria (i.e., vividness, reconstructability, and realism) are commonly evaluated on a rating or a categorical scale (see Table 2). The formula for calculating the overall RM score is reported in formula (1) (Sporer, 2004). The total RM score ranged from 0 to 48, with higher scores indicating higher narrative genuineness.

(1) **RM score =** *Vividness* **+** *Realism* **+** *Reconstructability* **+** *Perceptual information* **+** *Sensory information* **+** *Temporal information* **+** *Affective information* **-** *Cognitive Operation*

Table 2
List of Reality Monitoring criteria adapted from Sporer (2004).

Reality Monitoring criteria	Automatic score	Human scoring
Vividness	-	rating scales, categorical measures (presence vs. absence)
Realism	-	
Reconstructability	-	
Temporal information	LIWC "Tempo" + NER "DATE" + NER "TIME" + NER "EVENT"	rating scales, categorical measures (presence vs. absence), absolute/relative frequency
Spatial information	LIWC "Spazio" + NER "GPE" + NER "LOC" + NER "FAC"	
Perceptual information	LIWC "Proc_Sen"	
Affective information	LIWC "Affett"	
Cognitive operations	LIWC "Mec_Cog"	

Note. LIWC and NER features selected for Experiment 2 to automatically compute RM criteria are provided in the second column. The general human scoring procedure for each criterion is provided in the third column.
[a] LIWC = Linguistic Inquiry and Word Count
[b] NER = Named-Entity Recognition

2.1.5 Experimental procedure

Twelve questionnaires were created on the Qualtrics platform. The transcriptions of the 62 videoclips were randomly distributed among the 12 questionnaires, as in Monaro et al. (2022), to balance the number of truthful and deceptive transcriptions for each questionnaire. Therefore, each questionnaire consisted of 5 transcriptions, with the exception of only two questionnaires with 6 transcriptions.

All participants provided informed consent and demographic information before starting the questionnaire. Moreover, expert judges were assessed for their level of expertise in the forensic psychology field. After providing the instructions, the experimenter randomly gave each participant one of the 12 questionnaires.

Specifically, naïve judges were instructed to read each transcript carefully and to rate its credibility on a 10-point scale (1=totally fabricated, 10=totally genuine).

Forensic experts were first trained on the use of RM criteria for lie detection through a video lesson, which explained the theoretical background of the RM framework, the operationalization of the eight criteria and how to compute the final RM score; moreover, it provided two practical examples of application of the RM criteria to transcripts. After the training session, the experimenter remained available to reply to any question the experts had related to the RM procedure. Then, they were randomly assigned one of the 12 questionnaires. For each transcription, they were instructed to carefully read the text, evaluate it based on the eight RM criteria, and compute the overall RM score as in formula (1). Based on the final RM score, they were then asked to rate the transcripts' credibility on a 10-point scale (1=totally fabricated, 10=totally genuine). Other measures were also collected for naïve and expert judges but were not included in the analysis (see Supplementary Material for a detailed description).

2.2 Results

Accuracy was first computed at the subject level (i.e., number of correct classifications/total number of transcripts) and then averaged among naïve judges and forensic experts. The random baseline was set using the zero rule (see the Supplementary Material for more information). Nonparametric analyses were conducted after we checked whether the distribution of data violated normality assumptions. Data were preprocessed in Python using the Google Colab interface, and statistical analyses (i.e., Wilcoxon signed-rank and Mann-Whitney U tests) were conducted in Rstudio.

The results showed that naïve judges and forensic experts obtained an average accuracy slightly above the chance level ($accuracy_{NJ}$ = 54.1% ± 20.1, $accuracy_{FE}$ = 59.4% ± 19.9) in identifying lie-tellers and truth tellers. However, a Wilcoxon signed-rank test revealed that the naïve judges' performance was not significantly higher than chance level (V = 3831.00, p = .299, r_{rb}= .05, 95% CI [- .15, .26]). A Wilcoxon signed-rank test also showed that forensic experts' performance was not significantly higher than chance level (V = 431.00, p = .061, r_{rb} = .294, 95% CI [- .07, .59]). Contrary to expectations, a Mann-Whitney U test showed also that the forensic experts' average accuracy was not significantly higher than that of naïve judges (U = 2417.5, p = .134, r_{rb} = .12, 95% CI [- .10, .32]), suggesting there is no enough evidence in favor of Hypothesis 1a.
In the Supplementary Material, we report the accuracy the naïve judges and forensic experts reached in each experimental condition (truth-tellers vs. lie-tellers).

3. Experiment 2: machine-learning models trained on expert vs computerized reality monitoring scores

3.1 Methods and Materials

3.1.1 Text preprocessing

Before we extracted linguistic features, two raters manually preprocessed each transcription by removing semantic repetitions ("*we left at noon… ~~yes at noon~~*") and false starts ("*~~What I remember uhm~~… we went to London.*"). Using a two-way mixed-effects model for single measures (Shrout and Fleiss, 1979) with the JASP software (JASP Team, 2024), the intraclass correlation coefficient ($ICC_{3,1}$) was found to be .99 (95% CI [.98, .99]) for repetitions and .98 (95% CI [.96, .99]) for false starts, indicating excellent reliability among raters.
NER and features extraction of the CL were computed after this preprocessing step. For the LIWC scoring, an additional preprocessing step was conducted before the computation and consisted of automatic lowercase and tokenization of text using SpaCy and manual adjustment of bigrams (Italian: "non so," bigram: "nonso," English: "I don't know") and trigrams (Italian: "al di fuori," trigram: "aldifuori," English: "outside"). Word stemming was already included in the LIWC dictionary.

3.1.2 Feature extraction for reality monitoring

In this experiment, the RM criteria were automatically computed for each transcription using the LIWC software in combination with the NER technique.
LIWC is the gold standard software for analyzing word usage to identify psychosocial processes (Pennebaker et al., 2001). It calculates the percentage of words in a text corresponding to more than 80 categories related to linguistic and psychosocial dimensions in a validated dictionary (a detailed description of the functioning of LIWC-15 software and categories is reported in Boyd et al., 2022). Using the Italian dictionary (software version LIWC-15), semantic features related to time, space, affect, sensory processes, and cognition were computed to reflect five of the eight RM criteria. The RM criteria of vividness, reconstructability, and realism are subjective scores that do not fit any LIWC categories (see Table 2) and are therefore often excluded for computation.
NER is an NLP technique to identify and extract information (named entities) from texts and classify them into predefined categories, such as people, locations, organizations, and times. Named entities were automatically extracted using SpaCy, a Python library for NLP, on preprocessed text using a transformer-based model for the Italian language, available in Huggingface (*ita_nerIta_trf*, [https://huggingface.co/bullmount/it_nerIta_trf]). Table 2S in the Supplementary Material lists all entities available in the *ita_nerIta_trf* model with their descriptions and examples. Figure 1S in the Supplementary Material depicts a common way to represent text with annotated named entities.

Named entities related to “DATE,” “TIME,” and “EVENT” and ‘“GPE”, “LOC,” and “FAC” were added to the LIWC features related to time (tempo) and space (spazio), respectively.
This procedure was adopted because the LIWC software was mainly focused on words with a meaning associated with time and space (e.g., adverbs such as *“then”* and *“now”* and verbs such as “*to go*”) without taking into account specific information about space and time (e.g., toponymies, such as “*Ibiza,*” and terms indicating time, such as “*Monday*” and “*11am*”) that were detected using the NER technique.
Table 2 provides a summary of how the RM features using LIWC and NER scores were computed.

3.1.3 Machine-Learning Models and Training

Logistic regression, support vector machine (SVM), decision tree, and random forest were employed for the computational analysis, conceptualizing the lie detection task as a binary-classification problem. The inclusion of four ML models was intended to ensure that the obtained results were not dependent on the specific model assumptions and were stable across classifiers. The aforementioned models’ performance was evaluated and discussed in terms of accuracy. A random baseline was set using the zero rule. In the Supplementary Material, we provide a brief description of each model and include additional metrics to add interpretations, such as the Area Under the Curve (AUC), precision, recall, and F1 score.

A nested cross-validation (NCV) framework was utilized to evaluate ML models’ performance. NCV is a robust method for model evaluation and hyperparameter tuning in ML, especially in scenarios in which unbiased estimation of model performance is required (Muller & Guido, 2016). This method incorporated two layers: an inner loop for hyperparameter optimization and an outer loop for model performance evaluation. The inner loop, dedicated to hyperparameter optimization, utilized Grid Search for a systematic exploration of hyperparameter space to identify the optimal hyperparameter combination for each model. This process was repeated across the 10 folds of the inner cross validation (CV), ensuring that the hyperparameter selection was based solely on the training subset and not influenced by the test data. The best hyperparameter combination identified in the inner loop was then used to train the model on the entire training set of the outer loop. The outer loop then assessed the generalizability of the model with a 10-fold CV.
To enhance the performance estimates’ robustness and reliability, the NCV process was repeated with three random seeds to mitigate the effects of random variations in data partitioning, thus providing a robust unbiased estimation of the model’s performance. Following this rigorous procedure, the assessment of the model was safeguarded against overfitting and accurately reflected the model’s capability to classify truthful and deceptive statements in our dataset.

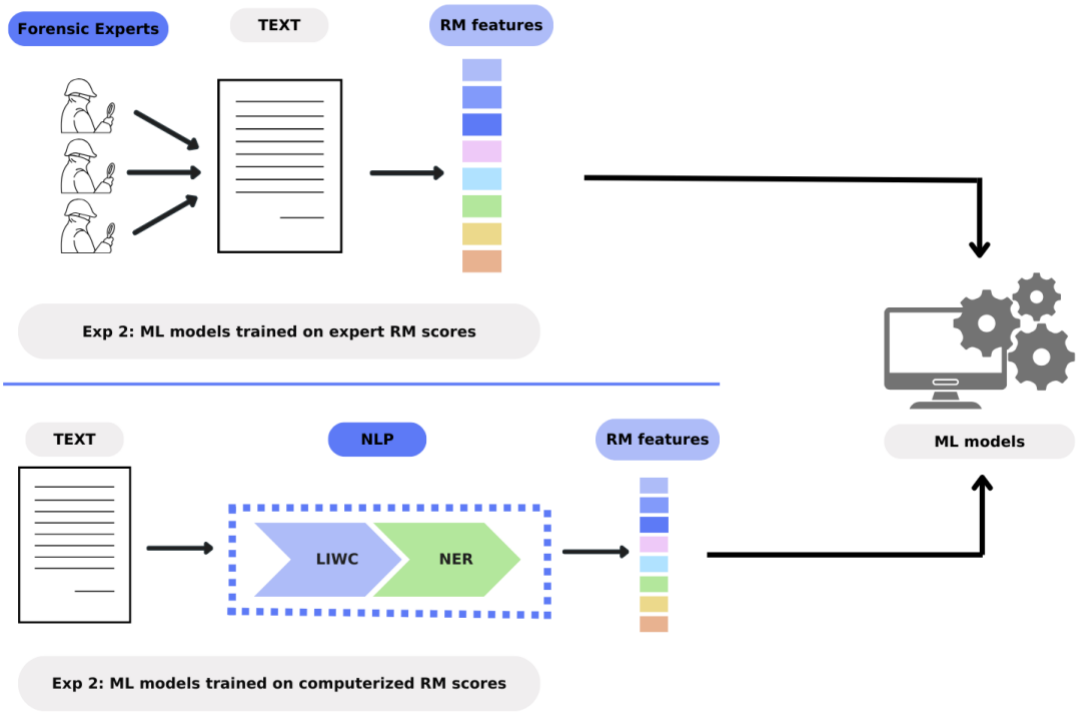
3.1.4 Procedure

Figure 1 depicts the steps adopted to conduct the computational analyses in this experiment. Specifically, ML models were trained on two sets of ratings. For the first set, the scores the three experts provided in Experiment 1 for the eight RM criteria were averaged for each story. This

resulted in a vector of eight scores per story, which was then used to train the ML models with an NCV procedure. For the second set, the RM features were extracted using NLP techniques following the procedure described in section 3.1.2. This resulted in a vector of five scores per story, which was employed to train the ML models with an NCV procedure.

Figure 1

Procedure employed in Experiment 2 to obtain two sets of features to train ML models.



Note. The first set of features was obtained by averaging the ratings three forensic experts provided for each RM criteria (upper part of the figure) on each text. A second set of features was obtained by leveraging NLP techniques (i.e., LIWC and NER) to extract linguistic features that mimic RM criteria on each text.

[a] ML= Machine Learning

[b] RM = Reality Monitoring

[c] NLP = Natural Language Processing

[d] LIWC = Linguistic Inquiry and Word Count

[e] NER = Named-Entity Recognition

3.2 Results

Table 3 provides the results of Experiment 2. Table 3S (in the Supplementary Material) reports the average performance and standard deviation in terms of accuracy, AUC, precision, recall, and F1

score obtained from the four ML models when they are trained on expert ratings of RM and computerized RM applied to full text.

When we used expert ratings of RM, the SVM and random forest models produced the highest average accuracy, 57.9% (±17.4) and 56.1% (±20.0), respectively. However, these performances were only slightly above chance level. Similarly, when we used computerized RM scores, the decision tree model exhibited the highest average accuracy, 57.1% (±15.8), but also this performance was just slightly above chance level. A Kruskal-Wallis test showed that the average accuracy of forensic experts from Experiment 1 was not significantly different from those of the best ML models trained with expert and computerized RM scores in Experiment 2 ($H(2) = 0.006$, $p = .997$, $\eta^2(H) = 0$, 95% CI [0, .05]). These findings support the second hypothesis (Hyp. 2b), namely that expert judges performed poorly in the lie detection task because the RM criteria were poorly informative for this type of dataset.

Table 3

ML models' performance is reported in terms of average accuracy in the 10-fold nested cross validation.

		Experiment 2		Experiment 3	Experiment 4
Dataset	ML models	RM - Expert Scores	RM - Computerized Scores	RM + CL	Data-driven approach
Free speech	Logistic regression	-	-	68.3 (±20.6)	68.9 (±21.1)
	SVM	-	-	64.1 (±19.7)	**69.0 (±18.3)**
	Decision Tree	-	-	60.3 (±18.5)	59.6 (±18.8)
	Random Forest	-	-	**69.4 (±16.5)**	68.2 (±17.9)
Unexpected questions	Logistic regression	-	-	56.9 (±19.1)	66.2 (±19.5)
	SVM	-	-	53.3 (±19.2)	64.7 (±17.3)
	Decision Tree	-	-	**60.8 (±14.6)**	57.5 (±18.6)
	Random Forest	-	-	60.6 (±18.5)	**67.4 (±18.8)**
Full text	Logistic regression	53.4 (±19.8)	40.2 (±18.0)	**67.2 (±19.1)**	73.3 (±18.6)
	SVM	**57.9 (±17.4)**	48.8 (±22.9)	62.3 (±17.6)	**77.3 (±17.2)**
	Decision Tree	49.9 (±19.7)	**57.1 (±15.8)**	57.2 (±16.6)	53.5 (±21.7)
	Random Forest	56.1 (±20.0)	52.8 (±20.4)	64.9 (±20.3)	75.1 (±17.5)

Note. Standard deviations are reported in brackets. The best accuracy achieved in each experiment for each part of the dataset analyzed is in bold.

[a] ML = Machine Learning

[b] RM = Reality Monitoring

[c] CL = Cognitive Load
[d] SVM = Support Vector Machine

4. Experiment 3: Theory-driven approach combining RM and CL

4.1 Methods and Materials

4.1.1 Feature Extraction for Cognitive Load

Previous research employed statistics regarding the text's length, readability, and complexity to extract linguistic features associated with CL in deception studies (Zhou et al., 2004; Pèrez-Rosas & Mihalcea, 2015; Solà-Sales et al., 2023; Sarzynska-Waver et al., 2023; Hauch et al., 2015). Statistics associated with CL were automatically computed on preprocessed text using the Python library TEXTSTAT and are reported in Table 4.

Table 4
List of the linguistic features associated with the cognitive load framework and their operational definition.

Features associated with cognitive load	Operational definition
num_sentences	Total number of sentences
word_count	Total number of words
num_unique_words	Total number of unique words
type-token ratio	Total number of unique words divided by the total number of words
num_syllables	Total number of syllables
avg_num_syllables_per_word	Average number of syllables per word
num_content_words	Total number of words that express lexical meaning
num_unique_content_words	Total number of unique content words
content-word diversity	Total number of unique content words divided by the total number of content words
fk_grade	The Flesch-Kincaid Grade Level Index expressing the grade level required to understand the text
fk_read	The Flesch Reading-Ease Level Index expressing the texts's readability

4.1.2 Procedure

Figure 2 depicts the procedure adopted for the computational analyses in Experiments 3 and 4. Specifically, the dataset was first split into three sections:
i) Free Speech, which contained the transcription of the free recall of the holiday;
ii) Unexpected Questions, which contained the responses to unexpected questions; and
iii) Full Text, intended as the combination of text from the Free Speech and Unexpected Questions sections.
Then, linguistic features associated with RM and CL were automatically extracted following the procedure defined in sections 3.1.2 and 4.1.1. Subsequently, the same ML models and NCV procedure employed in Experiment 2 were applied to each section of the dataset.

4.2 Results

Table 3 reports the results from this experiment (see also Table 4S in the Supplementary Materials). Considering the four ML models trained on linguistic features extracted from the Full Text dataset - using the RM and CL framework - we observed a general increase in the obtained predictive performance. In fact, after combining features from two theoretical frameworks, we reached an accuracy of 69.4% (±16.5), with an improvement of up to 9.3% over models trained solely on expert or computerized RM scores. The results show that the inclusion of linguistic features associated with the CL framework resulted in enhanced accuracy of ML models in detecting verbal deception compared to models trained solely on features from RM, confirming our hypothesis (Hyp. 3a).
Additionally, features from RM and CL were specifically extracted from statements in the Free Speech, Unexpected Questions, and Full Text sections to investigate their potential informative and predictive role. Findings comparing the performance obtained from ML models trained on each section showed that linguistic features from the Free Speech section significantly contributed to an increase in overall accuracy across the four models. Contrary to our expectations (Hyp. 3b), linguistic features from the Unexpected Questions section yielded lower accuracy rates, similar to those achieved by models trained exclusively on expert and computerized RM scores. Interestingly, when we leveraged the Full Text section for feature extraction, there was a slight decline in performance, with the highest average accuracy recorded at 67.2% (±19.1).

5. Experiment 4: Data-driven approach using NLP features

5.1 Methods and Materials

5.1.1 Feature Extraction and Selection

This experiment involved the extraction of a comprehensive set of 128 linguistic features, using a combination of NLP techniques on preprocessed texts. The Python library TEXTSTAT was employed to compute 11 basic textual features related to the text's length, readability, and complexity, as in Experiment 3. The LIWC software was employed to extract 85 psychological, linguistic, and affective dimensions from texts. The Python SpaCy library was employed to extract 15 named entities (with the NER technique) and 17 grammatical and syntactical parts of speech (with the POS-tagging technique) in the text.

Using the scikit-learn library in Python, the original set of 128 features was narrowed down to a more manageable and informative set of 20 features that best captured the nuances of the textual data, following these steps:

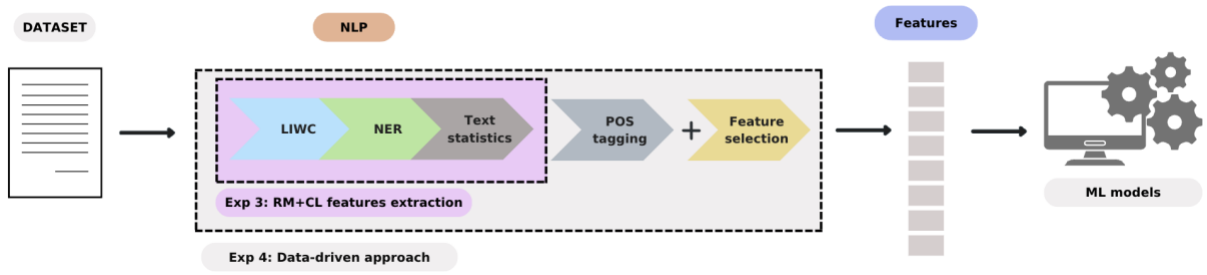
1. Removing POS features unrelated with the purpose of the task, such as the number of spaces (using spacebar; SPACE), punctuation usage (PUNCT), the number of symbols (SYM), and other noncanonical pos tags (X); those features were more related to the transcription process than to telling a truthful or a deceptive story and could represent a confounder if included in the analysis.
2. Removing duplicates, such as the numerical features in LIWC and POS tagging that were already detected with the NER technique and the LIWC "Non_flu" category, which was a duplicate of the POS tag "INTJ."
3. Removing LIWC linguistic features overlapping with the grammatical and syntactic features extracted with the POS-tagging technique given that the latter is more efficient and complete in extracting these features than the LIWC software.
4. Feature-engineering a new variable named "fillers" by summing filler words and non-fluencies, typical of hesitation and oral speech patterns, extracted with the LIWC software and the POS tagging. Specifically, the LIWC "riempiti" category was added to the POS tag "INTJ."
5. Removal of features that showed more than 60% of zero values across the dataset.
6. Selection of the best 20 features after testing for mutual information, which measures the linear and nonlinear dependency between random variables, ensuring that each selected feature contributed significantly to the predictive models (Ross, 2014). This selection process was performed using the function *sklearn.feature_selection.SelectKBest* (Pedregosa et al., 2011).

5.1.3 Procedure

Figure 2 depicts the procedure adopted for the computational analyses in Experiments 3 and 4. As in Experiment 3, the dataset was first divided into three sections (i.e., Free Speech vs. Unexpected Questions vs. Full Text). Then, following the feature-extraction and -selection process described in the previous section, four ML models were trained using an NCV procedure on the best 20 linguistic features from each section of the dataset. The ML models and the NCV procedure are described in Section 3.1.3.

Figure 2

Procedures employed in Experiments 3 and 4 to create a set of linguistic features to feed ML models.



Note. In Experiment 3, ML models were trained on linguistic features that mimic RM and CL using NLP techniques. In Experiment 4, a wider range of linguistic features was extracted from texts using NLP techniques; then, an automatic feature selection step was applied to obtain a final set of features used to train ML models.

[a] NLP = Natural Language Processing

[b] LIWC= Linguistic Inquiry and Word Count

[c] NER = Named Entity Recognition

[d] POS = Part-of-Speech

[e] RM = Reality Monitoring

[f] CL = Cognitive Load

5.2 Results

5.2.1 Data-driven approach

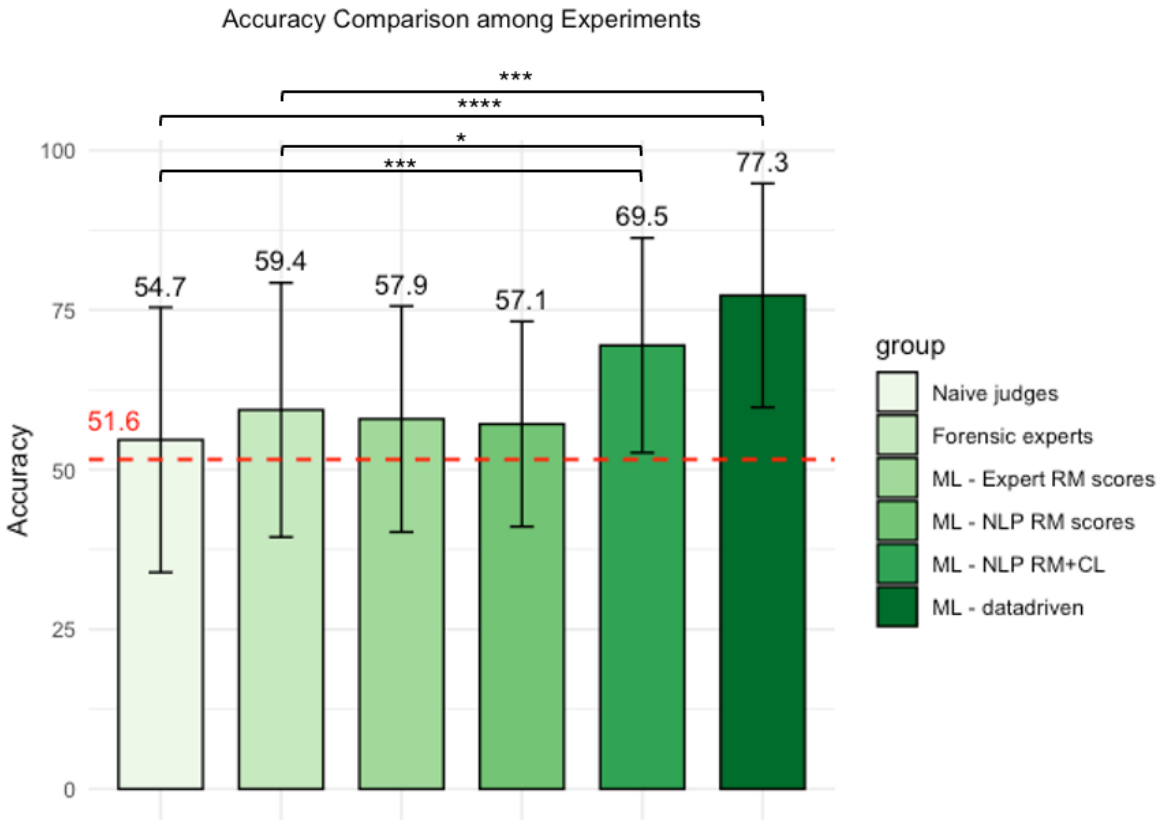
The data-driven approach demonstrated an overall improvement in performance compared to Experiment 3 (Table 3; see also Table 5S in Supplementary Material), providing evidence in support of the superior performance of data-driven approaches compared to theory-driven approaches (Hyp. 4a). Specifically, the Full Text section showed a significant leap in accuracy, particularly with the SVM, which reached the highest accuracy (77.3% ±17.2). The Unexpected Questions section, despite being the section with the lower performance, showed a noticeable improvement in accuracy, with the random forest model performing the best, at 67.4% (±18.8). In the Free Speech section, the SVM achieved the highest accuracy, 69.0% (±18.3), but it did not surpass the performance of the best model trained with RM and CL in the same section.

5.2.2 Comparing accuracy among Experiments

Figure 3 presents the accuracy achieved by human judges and the best ML models in each experiment. A Kruskal-Wallis test revealed a statistically significant difference in the average accuracy scores across experiments ($H(5) = 39.21$, $p < .01$, $\eta^2(H) = 0.13$, 95% CI [0.07, 0.24]). Dunn's post hoc tests with false-discovery rate correction were applied to assess differences between pairs of conditions. Notably, the average accuracy of the best ML model trained on RM and CL features was significantly higher than the average accuracy achieved by naïve judges ($z = 3.92$, $p < .01$) and forensic experts ($z = 2.39$, $p = 0.017$) in Experiment 1. Similarly, the average accuracy of the best ML model trained on features extracted with a data-driven approach was found to be significantly higher than the average accuracy reached by naïve judges ($z = 5.47$, $p < 0.001$) and forensic experts ($z = 3.67$, $p < 0.01$) in Experiment 1. Table 6S of the Supplementary Materials presents also the remaining post hoc comparisons. These findings support our last hypothesis, proving that theory-driven and data-driven approaches leveraging ML and NLP techniques are more effective in detecting verbal deception than human judges (Hyp. 4b).

Figure 3

Bar plot of the average accuracy (and standard deviation) obtained from the four experiments.



Note. Error bars represent standard deviations. The red dashed line represents the chance level (51.6%) as defined using the zero rule. Significant comparisons are reported only for human vs. machine approaches.

Naive judges' and *forensic experts'* accuracy are derived from Experiment 1; *ML-Expert RM scores*: best accuracy achieved in Experiment 2 using ML models trained on RM scores provided by forensic experts in Experiment 1. *ML-NLP RM scores*: best accuracy achieved in Experiment 2 using ML models trained on RM scores computed with NLP techniques; *ML-NLP RM+CL*: best accuracy achieved in Experiment 3 using ML; *ML-data-driven*: best accuracy achieved in Experiment 4 using ML.

[a] ML = machine learning

[b] RM = Reality Monitoring

[c] CL = Cognitive Load

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

6. General Discussion

This series of studies contributes to deception detection research by examining the theoretical frameworks of Reality Monitoring (RM) and Cognitive Load (CL) through computational methods, advancing our understanding of how these frameworks function in the analysis of deceptive language. In addition, through four experiments, we assessed the effectiveness of human (naïve vs experts) and ML-based (theory-driven vs data-driven) approaches in deception detection when applied to a dataset of interviews with unexpected questions.

In the first experiment, we tested the RM framework by comparing the performance of naive and expert judges, with the latter trained specifically in RM criteria. Neither naïve judges nor forensic experts surpassed the chance level ($accuracy_{NJ}$ = 54.1% ± 20.1, $accuracy_{FE}$ = 59.4% ± 19.9). Additionally, the average accuracy of forensic experts was not significantly higher than that of naïve judges. Although this result was expected for naïve judges and aligns with previous studies (Bond & DePaulo, 2006; Curci et al., 2019; DePaulo et al., 2003; Pérez-Rosas et al., 2015), it was contrary to expectations for forensic experts (Hyp. 1a). In facts, previous research proved the effectiveness of RM criteria in verbal-deception detection, reaching approximately 70% accuracy (Vrij et al., 2022; Gancedo et al., 2021; Amado et al., 2016; Vrij, 2008). Additionally, a meta-analysis of studies showed that the average accuracy that we can obtain after following different cues is 67% (Hartwig and Bond, 2014).

To address the reasons behind this poor performance of experts in Experiment 1, we introduced a second experiment that leverages computational techniques and ML models. ML models were trained on two sets of ratings, those given by expert judges using the RM criteria (i.e., expert ratings) in Experiment 1 and those obtained by computerized methods using NLP techniques for RM (i.e., computerized ratings), to determine whether naïve experts' poor performance was due to an inaccurate assessment of RM criteria, a lack of informativeness of those criteria for this dataset, or a decision-making problem in combining all the information. Our findings showed that the average accuracy of forensic experts was not significantly higher than those of the best ML

models trained on experts (accuracy = 57.9% ± 17.4) or computerized ratings of RM (accuracy = 57.1% ± 15.8), supporting the hypothesis that RM criteria might be poorly informative for the dataset under analysis, regardless of whether they were evaluated by forensic experts or derived from computational methods (Hyp. 2b).

The results of Experiments 1 and 2, collectively, challenged the presumed robustness of RM in deception detection and raises questions about its sensitivity across different datasets and contexts. Furthermore, these results challenge the efficacy of computational approaches built on theoretical frameworks that ultimately exhibit limited effectiveness, as demonstrated in this case with the RM. Considering these premises, we tested in a third experiment whether the combination of multiple theoretical frameworks, specifically the RM and CL frameworks, could enhance ML models' accuracy in detecting verbal deception (Hyp. 3a). The results from Experiment 3 demonstrated that the combination of features from two theoretical frameworks resulted in an accuracy of 69.4% (±16.5), with an improvement of 9.3% compared to models trained solely on expert or computerized RM scores. Furthermore, the average accuracy of the ML model trained on RM and CL features was significantly higher than the average accuracy that naïve and expert judges achieved in Experiment 1. This finding may be attributed to the fact that the dataset under analysis was specifically designed to increase CL in lie-tellers by posing unexpected questions (Monaro et al., 2020; Monaro et al., 2022) and to the inherently higher accuracy of the CL approach (Vrij et al., 2015). Alternatively, the simple inclusion of a higher number of relevant features might have led to this higher accuracy.

Because CL features were more prevalent in responses to unexpected questions, we hypothesized that ML models trained on features extracted from Unexpected Questions would yield higher accuracy than those trained on features extracted from Free Speech (Hyp. 3b). Contrary to this hypothesis, the results indicated that linguistic features derived from the Free Speech dataset significantly contributed to an increase in overall accuracy across all models. Linguistic features derived from the Unexpected Questions dataset yielded lower accuracy, similar to that achieved by models trained exclusively on expert and computerized RM scores. Notably, when we employed the Full Text dataset for feature extraction, performance slightly declined, with the highest average accuracy recorded at 67.2% (± 19.1). One possible interpretation of this result is that the extraction of linguistic markers useful for the detection of deceptive narratives is more effective with longer texts, as in the Free Speech and Full Text datasets.

Finally, when NLP techniques are used, various methodologies are available for extracting features from textual data, such as theory-driven, data-driven, and hybrid approaches (Van Der Zee et al., 2022). Although in forensic contexts hybrid models should be preferred to data-driven models given that data-driven models may be effective at predicting but ineffective at explaining, studies have proven a lower effectiveness of hybrid approaches compared to data-driven models (Van Der Zee et al., 2022). Therefore, in the fourth experiment, we investigated the performance of a data-

driven approach in this lie detection task and compared the results with those of previous experiments. Specifically, comparisons from Experiments 3 and 4 allow us to examine the effectiveness of theory-driven vs. data-driven approaches. NLP techniques were employed to extract a broad set of linguistic features, and a data-driven feature-selection strategy (Chandrill, 2022) was employed to identify a subset of highly informative features.
The results from Experiment 3 showed that training ML models on combined linguistic features from two deception frameworks (i.e., RM and CL) yielded higher but moderate accuracy ($best\_accuracy_{freespeech}$ = 69.4% ± 16.5). However, in Experiment 4, there was a significant leap, particularly with the SVM, which reached the highest accuracy ($best\_accuracy_{fulltext}$ = 77.3% ± 17.2). This performance was also significantly better than that of naïve judges and forensic experts (Experiment 1). These findings underline the efficacy of a data-driven approach in discerning patterns in comprehensive textual data compared to theory-driven approaches that combined linguistic features derived from the RM alone or in combination with the CL framework. We confirmed our hypothesis that a data-driven approach may be particularly relevant in contexts in which theory-based methods have demonstrated limited effectiveness (Hyp. 4a). Indeed, while previous studies have shown that RM typically achieves around 70% accuracy in distinguishing truth from deception (Vrij, 2008), it yielded lower accuracy in our study (from around 57% to 59%). In contrast, our data-driven NLP approach reached the expected 70% accuracy, suggesting that it could serve as a reliable alternative in cases where traditional and theory-based methods, like RM, fall short.

Most importantly, the results from Experiments 1 to 4 suggest that training ML models on features extracted using NLP techniques may represent a more advantageous approach in detecting deception from narratives, overcoming the modest accuracy achieved by naïve and expert judges because of their ability to handle complex patterns of language data (Hyp. 4b). Moreover, they could help identify which linguistic features are more informative to derive a final decision.

6.1 Limitations and Future perspectives

Although this study's results highlight significant advancements in the field of deception detection with the comparison of human judgments to computational predictions, several limitations must be acknowledged to properly contextualize the results and guide future research.

First, the reliance on a relatively small dataset significantly constrains the findings' generalizability. The dataset, comprising only 62 narratives and solely in Italian, limits our ability to confidently extend these results to broader and more heterogeneous contexts. Future studies replicating our experiments using larger and more varied datasets in different languages would enhance our findings' robustness and potentially reveal cultural and linguistic nuances in deception detection. Additionally, the dataset under analysis was designed to collect outright false statements. However, a more frequent and ecological form of deception is constituted by embedded lies (Caso et al., 2023; Verigin et al., 2019), where people interwire truth and lies

together. As a consequence, the detection rates found in our series of studies may be even lower when considering this type of deception.

Secondly, forensic experts assessed the RM criteria on a 7-point scale to judge the narratives' veracity. However, other approaches are available in the literature. For example, one approach involves evaluating the absence and presence of each criterion on a 3-point scale (*0=absent, 1=partially present, 2=totally present*), and another counts the frequency of details for at least five of the eight criteria. This study's results provide insights limited to the qualitative assessment of RM criteria on a 7-point scale and may not be generalizable to methodologies utilizing frequency of details. Future research could employ a different approach for RM assessment, for instance, by taking into account the frequency of details.

Third, by focusing exclusively on specific deception cues, such as those provided by the RM criteria, people may overlook other potentially informative cues. For instance, details' verifiability plays a crucial role in deception detection. The verifiability approach suggests that truth-tellers provide details that are more verifiable and a higher proportion of verifiable details (Nahari et al., 2014; Palena et al., 2021; Verschuere at al., 2021). Accordingly, a recent study showed that asking judges to assess narratives for their verifiability rather than their veracity yields higher accuracy, up to 70% (Verschuere et al., 2023). In our study, forensic experts may have underperformed relative to ML models because they employed an ineffective heuristic, as also evidenced by the results from Experiment 2, which demonstrated the RM's limited informativeness in assessing our dataset's veracity. Our research can therefore be extended by asking forensic experts to use different deception cues, such as assessing details' verifiability and using criteria-based content analysis (Steller, M., & Koehnken, 1989).

Lastly, we found that a data-driven approach yielded the highest accuracy when testing the models using nested cross-validation. However, it is essential to recognize the limitations of these results, especially considering the potential impact of error rates in forensic contexts. Although our model achieved a significant improvement, the approximate 30% error rate remains a concern, particularly given the serious implications of misclassification in legal settings where credibility assessments can influence case outcomes. These findings underline the need for continued research and refinement of NLP and machine learning methods to improve reliability in high-stakes applications. Indeed, there is substantial room to explore more sophisticated ML approaches in future studies. For example, techniques such as word embeddings offer a promising avenue for future research. Word embeddings provide a way to capture semantic relationships between words by representing them in a high-dimensional space (Lai et al., 2016), thereby uncovering subtle linguistic patterns associated with deceptive speech that traditional models do not capture. Moreover, using neural-network architectures, such as long short-term memory networks and transformers, would allow future research to process sequential data more effectively, potentially achieving higher accuracy in models trained on textual data. Finally, fine-tuning large language

models has also been proven to be effective in detecting deception in raw texts (Loconte et al., 2023).

However, a significant limitation of data-driven approaches is their lack of explainability, which is particularly relevant in forensic settings, in which understanding the rationale behind an algorithm's decision is as crucial as the decision itself. Although data-driven methods can efficiently identify patterns, make predictions, and sometimes explain which specific linguistic features contributed to those predictions, these outputs often are not easy to interpret. This opacity makes it challenging to align these findings with general theories of memory and deception, which is necessary for forensic credibility. Ensuring that computational techniques not only predict but also explain their predictions in terms that relate to established psychological theories will be essential for their acceptance and ethical application in legal contexts.

Overall, these future perspectives suggest a trajectory toward more integrated and sophisticated systems that leverage a combination of theoretical insights and cutting-edge ML techniques. By broadening the theoretical frameworks and enhancing the computational methods used in deception detection, researchers can provide more accurate, reliable, and explainable tools for forensic and other critical applications. This progression promises not only to advance the understanding of deception but also to enhance practical lie detection capabilities in real-world settings.

7. Conclusion

To conclude, experimental results from four experiments provided theoretical and practical considerations for advancing verbal deception detection research.

From a theoretical perspective, the exploration of multiple theoretical frameworks, such as RM and CL, through computational methods, has demonstrated the potential to enhance accuracy in identifying deceptive narratives and calls for the fusion of more diverse theoretical perspectives to offer more robust tools for deception detection, especially when one framework alone falls short.

From a practical perspective, the integration of computational methods in deception detection holds significant implications in forensic contexts when credibility assessment is required in criminal proceedings. Potentially, computational methods may aid forensic experts when they perform only slightly above chance level, even after being trained on well-established frameworks. However, the ethical implications of deploying computational methods in such sensitive settings are significant. Ethical considerations must include discussions on the transparency of the algorithms used (Zerilli et al., 2020; Von Eschenbach, 2021), the potential for overreliance on automated systems without adequate human oversight, and the need for ongoing evaluation of these systems' efficacy and fairness, especially when they influence judicial outcomes. Although our results were modest in Experiment 3 compared to those of Experiment 4, we highlight the importance of employing hybrid approaches that combine data-driven and theory-driven methodologies. Such approaches would provide a more explainable model, which is crucial in forensic contexts in which the reasoning behind decisions must be transparent and justifiable.

The results of this series of studies suggest the need for future studies aimed at integrating advanced computational techniques into the field of lie detection as well as providing transparent algorithms for the interpretation of the results in the forensic context.

Data availability statement

Data and scripts used to run the experiments are available at https://osf.io/usz26/

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

Supplemental material for this article is available online.

References

Alonso, M. A., Vilares, D., Gómez-Rodríguez, C., & Vilares, J. (2021). Sentiment analysis for fake news detection. Electronics, 10(11), 1348. https://doi.org/10.3390/electronics10111348

Amado, B. G., Arce, R., & Fariña, F. (2015). Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. The European Journal of Psychology Applied to Legal Context, 7(1), 3-12. https://doi.org/10.1016/j.ejpal.2014.11.002

Amado, B. G., Arce, R., Farina, F., & Vilarino, M. (2016). Criteria-Based Content Analysis (CBCA) reality criteria in adults: A meta-analytic review. International Journal of Clinical and Health Psychology, 16(2), 201-210. https://doi.org/10.1016/j.ijchp.2016.01.002

Bond Jr, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and social psychology Review, 10(3), 214-234. https://doi.org/10.1207/s15327957pspr1003_2

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). The development and psychometric properties of LIWC-22. Austin, TX: University of Texas at Austin, 1-47. https://doi.org/10.13140/RG.2.2.23890.43205

Caso L, Cavagnis L, Vrij A and Palena N (2023). Cues to deception: can complications, common knowledge details, and selfhandicapping strategies discriminate between truths, embedded lies and outright lies in an Italian-speaking sample? Front. Psychol. 14:1128194. doi: 10.3389/fpsyg.2023.1128194

Chandril G., Data Analysis with Machine Learning for Psychologists. Springer Cham. 2022. https://doi.org/10.1007/978-3-031-14634-3

Constâncio, A. S., Tsunoda, D. F., Silva, H. de F. N., Silveira, J. M. da, & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. PLOS ONE, 18(2), e0281323. https://doi.org/10.1371/journal.pone.0281323

Curci, Antonietta, Lanciano, Tiziana, Battista, Fabiana, Guaragno, Sabrina, & Ribatti, Raffaella Maria (2019). Accuracy,Confidence, and Experiential Criteria for Lie Detection Through a Videotaped Interview. Frontiers in Psychiatry, 9, 283. http://dx.doi.org/10.3389/fpsyt.2018.00748

Deeb, H., Vrij, A., Palena, N., Hypšová, P., Dib, G., Leal, S., & Mann, S. (2024). Honesty repeats itself: comparing manual and automated coding on the veracity cues total details and redundancy. *Applied Psycholinguistics*, 45(5), 934-962. https://doi.org/10.1017/S0142716424000298

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. Psychological bulletin, 129(1), 74. https://doi.org/10.1037/0033-2909.129.1.74

Ekman, P., & O'Sullivan, M. (1991). Who can catch a liar?. American psychologist, 46(9), 913. https://doi.org/10.1037/0003-066X.46.9.913

Elaad, E. (2009). Lie-detection biases among male police interrogators, prisoners, and laypersons. Psychological Reports, 105(3_suppl), 1047-1056. https://doi.org/10.2466/PR0.105.F.1047-1056

Faul, Franz, Erdfelder, Edgar, Lang, Albert-Georg, & Buchner, Axel (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39(2), 175–191. http://dx.doi.org/10.3758/BF03193146

Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M., & Arce, R. (2021). Reality monitoring: A meta-analytical review for forensic practice. European Journal of Psychology Applied to Legal Context, 13(2), 99-110. https://doi.org/10.5093/ejpalc2021a10

Hartwig, M., & Bond Jr, C. F. (2014). Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, *28*(5), 661-676. https://doi.org/10.1002/acp.3052

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. Personality and social psychology Review, 19(4), 307-342. https://doi.org/10.1177/1088868314556539

Hauch, V., Sporer, S. L., Masip, J., & Blandón-Gitlin, I. (2017). Can credibility criteria be assessed reliably? A meta-analysis of criteria-based content analysis. Psychological Assessment, 29(6), 819–834. https://doi.org/10.1037/pas0000426

JASP Team. (2024). JASP (Version 0.18.3) [Computer software]. Retrieved from https://jasp-stats.org/

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. Psychological review, 88(1), 67. https://doi.org/10.1037/0033-295X.88.1.67

Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. (2018a). 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

Kleinberg, B., Van Der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. (2018b). Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. Applied cognitive psychology, 32(3), 354-366.

Lai, S., Liu, K., He, S., & Zhao, J. (2016). How to generate a good word embedding. IEEE Intelligent Systems, 31(6), 5-14. https://doi.org/10.1109/MIS.2016.45

Lancaster, G. L., Vrij, A., Hope, L., & Waller, B. (2013). Sorting the liars from the truth tellers: The benefits of asking unanticipated questions on lie detection. *Applied Cognitive Psychology*, *27*(1), 107-114. https://doi.org/10.1002/acp.2879

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. Journal of Language and Social Psychology, 33(4), 378-392. https://doi.org/10.1177/0261927X14535916

Loconte, R., Russo, R., Capuozzo, P., Pietrini, P., & Sartori, G. (2023). Verbal lie detection using Large Language Models. Scientific Reports, 13(1), 22849. https://doi.org/10.1038/s41598-023-50214-0

Masip, J., Sporer, S. L., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: a review of the empirical evidence. Psychology, Crime & Law, 11(1), 99–122. https://doi.org/10.1080/10683160410001726356

Melis, G., Ursino, M., Scarpazza, C., Zangrossi, A., & Sartori, G. (2024). Detecting lies in investigative interviews through the analysis of response latencies and error rates to unexpected questions. *Scientific Reports*, *14*(1), 12268. https://doi.org/10.1038/s41598-024-63156-y

Mihalcea, R., & Strapparava, C. (2009, August). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 conference short papers* (pp. 309-312).

Monaro, M., Maldera, S., Scarpazza, C., Sartori, G., & Navarin, N. (2022). Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. Computers in Human Behavior, 127, 107063. https://doi.org/10.1016/j.chb.2021.107063

Monaro, M., Capuozzo, P., Ragucci, F., Maffei, A., Curci, A., Scarpazza, C., ... & Sartori, G. (2020). Using blink rate to detect deception: A study to validate an automatic blink detector and a new dataset of videos from liars and truth-tellers. In Human-Computer Interaction. Human Values and Quality of Life: Thematic Area, HCI 2020, Held as Part of the 22nd International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part III 22 (pp. 494-509). Springer International Publishing. https://doi.org/10.1007/978-3-030-49065-2_35

Monaro, M., Galante, C., Spolaor, R., Li, Q. Q., Gamberini, L., Conti, M., & Sartori, G. Covert lie detection using keyboard dynamics. Sci. Rep. 8, 1976 (2018). https://doi.org/10.1038/s41598-018-20462-6

Monaro, M., Gamberini, L., & Sartori, G. (2017). The detection of faked identity using unexpected questions and mouse dynamics. *PloS one*, *12*(5), e0177851. https://doi.org/10.1371/journal.pone.0177851

Müller, A. C., & Guido, S. (2016). Introduction to machine learning with Python: a guide for data scientists. Reilly Media, Inc

Nahari, G. (2016). When the long road is the shortcut: A comparison between two coding methods for content-based lie-detection tools. Psychology, Crime & Law, 22(10), 1000-1014. https://doi.org/10.1080/1068316X.2016.1207770

Nahari, G., Vrij, A., & Fisher, R. P. (2014). Exploiting liars' verbal strategies by examining the verifiability of details. Legal and Criminological Psychology, 19(2), 227-239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. Personality and social psychology bulletin, 29(5), 665-675. https://doi.org/10.1177/0146167203029005010

Oberlader, V. A., Naefgen, C., Koppehele-Gossel, J., Quinten, L., Banse, R., & Schmidt, A. F. (2016). Validity of content-based techniques to distinguish true and fabricated statements: A meta-analysis. Law and Human Behavior, 40(4), 440–457. https://doi.org/10.1037/lhb0000193

Oberlader, V. A., Quinten, L., Banse, R., Volbert, R., Schmidt, A. F., & Schönbrodt, F. D. (2021). Validity of content-based techniques for credibility assessment—How telling is an extended meta-analysis taking research bias into account?. Applied Cognitive Psychology, 35(2), 393-410. https://doi.org/10.1002/acp.3776

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Palena, N., Caso, L., Vrij, A., & Nahari, G. (2021). The verifiability approach: A meta-analysis. *Journal of Applied Research in Memory and Cognition*, *10*(1), 155-166. https://doi.org/10.1016/j.jarmac.2020.09.001

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pérez-Rosas, V., & Mihalcea, R. (2015, September). Experiments in open domain deception detection. In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1120-1125).

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. PloS one, 9(2), e87357. https://doi.org/10.1371/journal.pone.0087357

Salminen, J., Kandpal, C., Kamel, A. M., Jung, S. G., & Jansen, B. J. (2022). Creating and detecting fake reviews of online products. Journal of Retailing and Consumer Services, 64, 102771. https://doi.org/10.1016/j.jretconser.2021.102771

Sarzynska-Wawer, J., Pawlak, A., Szymanowska, J., Hanusz, K., & Wawer, A. (2023). Truth or lie: Exploring the language of deception. Plos one, 18(2), e0281179. https://doi.org/10.1371/journal.pone.0281179

Savy, R (2006). Specifiche per la trascrizione ortografica annotata dei testi raccolti. Italiano parlato. Analisi di un dialogo,1–37.

Schutte, M., Bogaard, G., Mac Giolla, E., Warmelink, L., & Kleinberg, B. (2021). Man versus Machine: Comparing manual with LIWC coding of perceptual and contextual details for verbal lie detection (preprint). https://doi.org/10.31234/osf.io/cth58

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. Psychological bulletin, 86(2), 420. https://doi.org/10.1037/0033-2909.86.2.420

Solà-Sales, S., Alzetta, C., Moret-Tatay, C., & Dell'Orletta, F. (2023). Analysing Deception in Witness Memory through Linguistic Styles in Spontaneous Language. Brain Sciences, 13(2), 317. https://doi.org/10.3390/brainsci13020317

Sporer, S. L. (2004). Reality monitoring and detection of deception. In P.-A. Granhag & L. Strömwall (Eds.), The detection of deception in forensic contexts (pp. 64–101). Cambridge University Press. https://doi.org/10.1017/CBO9780511490071.004

Sporer, S. L. (1997). The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 11(5), 373–397. https://doi.org/10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

Steller, M., & Koehnken, G. (1989). Criteria-Based Content Analysis. The Suggestibility of Children's Recollections. https://doi.org/https://doi.org/10.1037/t27704-000

Street, C. N., & Masip, J. (2015). The source of the truth bias: Heuristic processing?. Scandinavian Journal of Psychology, 56(3), 254-263. https://doi.org/10.1111/sjop.12204

Strömwall, L. A., Bengtsson, L., Leander, L., & Granhag, P. A. (2004). Assessing children's statements: the impact of a repeated experience on CBCA and RM ratings. Applied Cognitive Psychology, 18(6), 653–668. https://doi.org/10.1002/acp.1021

Tomas, F., Dodier, O., & Demarchi, S. (2022). Computational measures of deceptive language: prospects and issues. Frontiers in Communication, 7, 792378. https://doi.org/10.3389/fcomm.2022.792378

Van Der Zee, S., Poppe, R., Havrileck, A., & Baillon, A. (2022). A personal model of trumpery: linguistic deception detection in a real-world high-stakes setting. *Psychological science*, *33*(1), 3-17. https://doi.org/10.1177/09567976211015941

Verigin, B. L., Meijer, E. H., Bogaard, G., & Vrij, A. (2019). Lie prevalence, lie characteristics and strategies of self-reported good liars. PloS one, 14(12), e0225566. https://doi.org/10.1371/journal.pone.0225566

Verschuere, B., Lin, C. C., Huismann, S., Kleinberg, B., Willemse, M., Mei, E. C. J., ... & Meijer, E. (2023). The use-the-best heuristic facilitates deception detection. Nature human behaviour, 7(5), 718-728. https://doi.org/10.1038/s41562-023-01556-2

Verschuere, B., Bogaard, G., & Meijer, E. (2021). Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis. *Applied Cognitive Psychology*, *35*(2), 374-384. https://doi.org/10.1002/acp.3775

Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. (2022). Verbal lie detection: its past, present and future. Brain Sciences, 12(12), 1644. https://doi.org/10.3390/brainsci12121644

Vrij, A., Fisher, R. P., & Blank, H. (2017). A cognitive approach to lie detection: A meta-analysis. Legal and Criminological Psychology, 22(1), 1-21. https://doi.org/10.1111/lcrp.12088

Vrij, A., Leal, S., Granhag, P. A., Mann, S., Fisher, R. P., Hillman, J., & Sperry, K. (2009). Outsmarting the liars: The benefit of asking unanticipated questions. *Law and human behavior*, *33*, 159-166. https://doi.org/10.1007/s10979-008-9143-y

Vrij, A. (2008). Detecting lies and deceit: Pitfalls and opportunities. John Wiley & Sons.

Vrij, A., Fisher, R., Mann, S., & Leal, S. (2008). A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, *5*(1-2), 39-43. https://doi.org/10.1002/jip.82

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview styles. Law and Human Behavior, 31(5), 499–518. https://doi.org/10.1007/s10979-006-9066-4

Vrij, A. (2005). Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. Psychology, Public Policy, and Law, 11(1), 3–41. https://doi.org/10.1037/1076-8971.11.1.3

Vrij, A., Edward, K., Roberts, K. P., & Bull, R. (2000). Detecting deceit via analysis of verbal and nonverbal behavior. Journal of Nonverbal Behavior, 24(4), 239–263. https://doi.org/10.1023/A:1006610329284

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. Philosophy & Technology, 34(4), 1607-1622. https://doi.org/10.1007/s13347-018-0330-6

Walczyk, J. J., Igou, F. P., Dixon, A. P., & Tcholakian, T. (2013). Advancing lie detection by inducing cognitive load on liars: A review of relevant theories and techniques guided by lessons from polygraph-based approaches. *Frontiers in psychology*, *4*, 14. https://doi.org/10.3389/fpsyg.2013.00014

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). Transparency in algorithmic and human decision-making: is there a double standard?. Philosophy & Technology, 32, 661-683. https://doi.org/10.1007/s00146-020-00960-w

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. Group decision and negotiation, 13, 81-106. https://doi.org/10.1023/B:GRUP.0000011944.62889.6f