Verbal Lie Detection using Large Language Models

Riccardo Loconte (riccardo.loconte@imtlucca.it)
IMT School of Advanced Studies Lucca

Roberto Russo
University of Padua

Pasquale Capuozzo
University of Padua

Pietro Pietrini
IMT School of Advanced Studies Lucca

Giuseppe Sartori
University of Padua

Article

Keywords:

Posted Date: August 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3126100/v1

Abstract

Given that human accuracy in detecting deception has been proven to not go above the chance level, several automatized verbal lie detection techniques employing Machine Learning and Transformer models have been developed to reach higher levels of accuracy. This study is the first to explore the performance of a Large Language Model, FLAN-T5 (small and base sizes), in a lie-detection classification task in three English-language datasets encompassing personal opinions, autobiographical memories, and future intentions. After performing stylometric analysis to describe linguistic differences in the three datasets, we tested the small- and base-sized FLAN-T5 in three Scenarios using 10-fold cross-validation: one with train and test set coming from the same single dataset, one with train set coming from two datasets and the test set coming from the third remaining dataset, one with train and test set coming from all the three datasets. We reached state-of-the-art results in Scenarios 1 and 3, outperforming previous benchmarks. The results revealed also that model performance depended on model size, with larger models exhibiting higher performance.Furthermore, stylometric analysis was performed to carry out explainability analysis, finding that linguistic features associated with the Cognitive Load framework may influence the model's predictions.

Furthermore, stylometric analysis was performed to carry out explainability analysis, finding that linguistic features associated with the Cognitive Load framework may influence the model's predictions.

Introduction

Lie detection involves the process of determining the veracity of a given communication. When producing deceptive narratives, liars employ verbal strategies to create false beliefs in the interacting partners and are thus involved in a specific and temporary psychological and emotional state [1]. For this reason, the Undeutsch hypothesis suggests that deceptive narratives differ in form and content from truthful narratives [2]. This topic has always been under constant investigation and development in the field of cognitive psychology, given its significant and promising applications in the forensic and legal setting [3]. Its potential pivotal role is in determining the honesty of witnesses and potential suspects during investigations and legal proceedings, impacting both the investigative information-gathering process and the final decision-making level [4].

Although considerable research has focused on identifying verbal cues for deception and developing effective methods to differentiate between truthful and deceptive narratives, such verbal cues are, at best, subtle and typically result in both naive and expert individuals, both performing just above chance levels [5, 6]. A potential explanation coming from social psychology for this unsatisfactory human performance is the intrinsic human inclination to the *truth bias* [7], i.e., the cognitive heuristic of presumption of honesty, which makes people assume that an interaction partner is truthful unless they have reasons to believe otherwise [8, 9].

Given the human inability to satisfactorily detect deception from verbal cues, recently this issue has been tackled by employing computational techniques, such as stylometry. Stylometry refers to a set of methodologies and tools from computational linguistic and artificial intelligence that allow to conduct quantitative analysis of linguistic features within written texts to uncover distinctive patterns that can infer and characterize authorship or other stylistic attributes [10, 11, 12]. Albeit with some limitations, stylometry has been proven to be effective in the context of lie detection [13, 14]. The main advantage is the possibility to code and extract verbal cues independently from human judgment, hence reducing the problem of inter-coder agreement, as researchers using the same technique for the same data will extract the same indices [14].

Alongside this trend, several recent studies have explored computational analysis of language in different domains, such as fake news [15, 16], transcriptions of court cases [17, 18, 19], evaluations of deceptive product reviews [20, 21, 22], investigations into cyber-crimes [23], analysis of autobiographical information [24], and assessments of deceptive intentions regarding future events [25]. Taken together, most of those studies focused on the usage of Machine Learning and Deep Learning algorithms combined with Natural Language Processing (NLP) techniques to detect deception from verbal cues automatically (see Constacio et al., 2003 [26] for a systematic review of the computerized techniques employed in lie-detection studies). However, to the best of our knowledge, the procedure of fine-tuning a Large Language Model (LLM) on small *corpora* for a lie-detection task has never been explored in this sense.

LLMs are Transformer language models with at least hundreds of millions of parameters trained on a large collection of *corpora* (i.e., pre-training phase) [27]. Thanks to this pre-training phase, LLMs have proven to capture the intricate patterns and structures of language and develop a robust understanding of syntax, semantics, and pragmatics. In addition, once pre-trained, these models can be fine-tuned on specific tasks using smaller task-specific datasets, achieving state-of-the-art results [27]. Common tasks for LLMs fine-tuning include language translation, text classification (e.g., sentiment analysis), question-answering, text summarization, and code generation. Therefore, LLMs excels at a wide range of NLP tasks, as opposed to models uniquely trained for one specific task [27].

Given the extreme flexibility of LLMs, the main objective of this study was to investigate the effectiveness of fine-tuning an LLM in classifying the veracity of short narratives from raw texts. To this aim, we fine-tuned an open-source LLM named FLAN-T5, developed by Google researchers and freely available on HuggingFace, using three datasets encompassing personal opinions (the **Deceptive Opinions dataset** [28]), autobiographical experiences (the **Hippocorpus dataset** [29]) and future intentions (the **Intention dataset** [30]). Additionally, we examined the performance and generalization capabilities of the fine-tuned FLAN-T5 model when tested on deceptive narratives from different contexts than those from the training set or when trained and tested on a multi-context dataset. According to empirical evidence, classical machine learning models tend to experience a decline in performance when trained and tested on the aforementioned scenarios [31, 32, 33]. However, LLMs have acquired a comprehensive understanding of language patterns during the pre-training phase. This raises the question of whether LLMs can effectively learn and generalize an inner linguistic representation of deception across different contexts and domains.

Finally, after applying stylometry to truthful and deceptive statements in the three datasets to describe and understand their main differences, we applied stylometry to understand whether the linguistic style by which truthful or deceptive narratives are delivered is a feature that the model takes into account for its final prediction. Notably, we employed a theory-based stylometric approach by extracting linguistic features related to the psychological frameworks of Distancing [34], Cognitive Load [35], Reality Monitoring [36], and Verifiability Approach [37, 38]. In this way, we analyzed and discussed the issue of verbal lie detection considering multiple theoretical frameworks.

Methods

Datasets

Three datasets were employed for this study: the Deceptive Opinions dataset [28], from now on **Opinion Dataset**, the Hippocorpus dataset [29], from now on **Memory Dataset**, and the **Intention dataset** [30]. For each dataset, participants were required to provide genuine or fabricated statements in three different domains: personal opinions, autobiographical experiences, and future intentions, respectively. Notably, the specific topic within each domain was counterbalanced among liars and truth-tellers. A more detailed description of each dataset is available in the Supplementary Material as well as in the method section of each original article.

FLAN-T5

We adopted FLAN-T5, an LLM developed by Google researchers and freely available through HuggingFace Python's library Transformers (https://huggingface.co/docs/transformers/model_doc/flan-t5), because of its valuable trade-off between computational load and goodness of the learned representation. FLAN-T5 is the improved version of MT-5, a text-to-text general model capable of solving many NLP tasks (e.g., sentiment analysis, question answering, and machine translation), which has been improved by pre-training [39]. The peculiarity of this model is that every task they were trained on is transformed into a text-to-text task. For example, while performing sentiment analysis, the output prediction is the string used in the training set to label the positive or negative sentiment of each phrase rather than a binary integer output (e.g., 0 = positive; 1 = negative). Hence, their power stands in both the generalized representation of natural language learned during the pre-training phase and the possibility of easily adapting the model to a downstream task with little fine-tuning without adjusting its architecture.

Experimental set-up

Expressing opinions, recalling autobiographical memories, and stating intentions - truthfully or deceptively - involve distinct linguistic styles; hence, the model has to learn and differentiate linguistic patterns for each context to be deemed a reliable lie detector. Based on these considerations, we designed three Scenarios to test our research hypothesis (Fig. 1):

1. **Scenario 1:** The model was fine-tuned on a portion of a single dataset and tested on the remaining part. This procedure was repeated three times - one for each dataset - with a different copy of the same model each time (i.e., the same parameters before the fine-tuning process). This Scenario assesses the model's capacity to learn how to detect lies related to the same context;
2. **Scenario 2:** The model was fine-tuned on two out of the three datasets and tested on the remaining unseen dataset. As for the previous scenario, this procedure was iterated three times, employing separate instances of the same model, each time with a distinct combination of dataset pairings. This Scenario assesses how the model performs on samples from a new context to which it has never been exposed during the training phase;
3. **Scenario 3:** We first aggregated the three train and test sets from Scenario 1. Then we fine-tuned and tested the model on those aggregated sets. This Scenario assesses the capacity of the model to learn and generalize from samples of truthful and deceptive narratives from multiple contexts.

Together, Scenarios 2 and 3 provide evidence about the generalized capabilities of the fine-tuned FLAN-T5 model in a lie-detection task when tested on unseen data and on a multi-domain dataset.

Furthermore, we tested whether model performance may depend on model sizes. Therefore, we first fine-tuned the *small*-sized version of FLAN-T5 in every scenario, and then we repeated the same experiments in every scenario with the *base*-sized version. In both Scenarios 1 and 3, each experiment underwent a 10-fold cross-validation procedure, employing identical train-test splits within each scenario and for both model sizes. The average test accuracy and its corresponding standard deviation are presented as performance metrics. Conversely, in Scenario 2, each pairing combination underwent fine-tuning using the entire two paired datasets as training set, while the model's performance was assessed using the complete unseen dataset as test set.

Notably, the Opinion dataset was developed to have for each participant truthful and deceptive statements for a total of five opinions. Therefore, we treated each opinion as a separate sample. In order to avoid the model exhibiting inflated performance on the test set as a result of learning the participants' linguistic style, we adopted the following precautionary measure. Specifically, we ensured an exclusive division of participants between the training and test sets, such that any individual who had their opinions assigned to the training set did not have their opinions assigned to the test set, and vice versa.

Fine-tuning strategy

Fine-tuning of LLMs consists in adapting a pre-trained language model to a specific task by further training the model on task-specific data, thereby enhancing its ability to generate contextually relevant and coherent text in line with the desired task objectives [39]. We fine-tuned FLAN-T5 in its small and base size using the three datasets and following the experimental set-up described above. We approached the lie-detection task as a binary classification problem, given that all the three datasets comprised raw texts associated with a binary label, specifically instances classified as either truthful or deceptive.

To the best of our knowledge, no fine-tuning strategy is available in the literature for this novel downstream NLP task. Therefore, our strategy followed an adaptation of the Hugginface's guidelines on fine-tuning an LLM for translation. Specifically, we chose the same optimization strategy used to pre-train the original model and the same loss function.

Noteworthy, the classification task between deceptive and truthful statements has never been performed during the FLAN-T5 pre-training phase, neither is it included in any of the tasks that the model has been pre-trained on. Therefore, we performed preliminary experiments on different training-validation splits for each scenario. All experiments and runs of the three scenarios were conducted on Google Colaboratory Pro + using their NVIDIA A100 Tensor Core GPU. After several experiments, the following hyperparameter configuration, reported in Table 1, yielded the best performance in terms of accuracy.

Table 1

FLAN-T5 hyperparameters configuration for the small- and base-sized version. The initial learning rate for every scenario was 5e-4 for the small model and 5e-5 for the base model. This choice was motivated by preliminary experiment results, with the smaller model, but not the base model, generally performing better with higher learning rates. The weight decay coefficient was set to 0.01 in all models and Scenarios. The batch size was set to 2 for computational reasons, specifically to avoid running out of available memory, even though it is known that a larger batch size usually leads to better performance. Finally, the number of epochs was set to 3 after preliminary experiments showing the maximum test accuracy after the third epoch without overfitting.

Model	Hyperparameter	Value
FLAN-T5 small	Learning rate	5e-4
	Weight decay coefficient	0.01
	Batch size	2
	Number of Epochs	3
FLAN-T5 base	Learning rate	5e-5
	Weight decay coefficient	0.01
	Batch size	2
	Number of Epochs	3

DeCLaRatiVE stylometric analysis

This study employed stylometric analysis to achieve two primary objectives. First, we aimed to describe the linguistic features that distinguished the three datasets before initializing the fine-tuning process. Second, we conducted explainability analysis to gain insights into the role of linguistic style that differentiated truthful and deceptive statements in the model's classification process. For this purpose, a new framework that we referred to as **D**e**CL**a**R**ati**VE** stylometry was adopted, which involved the extraction of 26 linguistic features in conjunction with the psychological frameworks of **D**istancing [34], **C**ognitive **L**oad [35], **R**eality **M**onitoring [36], and **VE**rifiability Approach [37, 38]. A full list of the 26 linguistic features with a short description is shown in Table 2. This comprehensive approach enabled the analysis of verbal cues of deception from a multidimensional perspective, encompassing various theoretical frameworks, and was performed using a range of NLP techniques described in the Supplementary Information.

Table 2
List and short description of the 26 linguistic features pertaining to the DeCLaRatiVE Stylometry technique.

Label	Description
num_sentences	Total number of sentences
num_words	Total number of words
num_syllables	Total number of syllables
avg_syllabes_per_word	Average number of syllables per word
fk_grade	Index of the grade level required to understand the text
fk_read	Index of the readability of the text
Analytic	LIWC summary statistic analyzing the style of the text in term of analytical thinking (0–100)
Authentic	LIWC summary statistic analyzing the style of the text in term of authenticity (0–100)
Tone	Standardized difference (0-100) of ‘tone_pos’ - ‘tone_neg’
tone_pos	Percentage of words related to a positive sentiment (LIWC dictionary)
tone_neg	Percentage of words related to a negative sentiment (LIWC dictionary)
Cognition	Percentage of words related to semantic domains of cognitive processes (LIWC dictionary)
memory	Percentage of words related to semantic domains of memory/forgetting (LIWC dictionary)
focuspast	Percentage of verbs and adverbs related to the past (LIWC dictionary)
focuspresent	Percentage of verbs and adverbs related to the present (LIWC dictionary)
focusfuture	Percentage of verbs and adverbs related to the future (LIWC dictionary)
Self-reference	Sum of LIWC categories ‘i’ + ‘we’
Other-reference	Sum of LIWC categories ‘shehe’ + ‘they’ +‘you’
Perceptual details	Sum of LIWC categories ‘attention’ + ‘visual’ + ‘ auditory’+ ‘feeling’
Contextual Embedding	Sum of LIWC categories ‘space’ + ‘motion’ + ‘time’
Reality Monitoring	Sum of Perceptual details + Contextual Embedding + Affect - Cognition
Concreteness score	Mean of concreteness score of words
People	Unique named-entities related to people: e.g., ‘Mary’, ‘Paul’, ‘Adam’
Temporal details	Unique named-entities related to time: e.g., ‘Monday’, ‘2:30 PM’, ‘Christmas’

Label	Description
Spatial details	Unique named-entities related to space: e.g., ‘airport’, ‘Tokyo’, ‘Central park’
Quantity details	Unique named-entities related to quantities: e.g., ‘20%’, ‘5 $’, ‘first’, ‘ten’, ‘100 meters’

Results

Descriptive Linguistic Analysis

This section outlines the results of the descriptive linguistic analysis in terms of vocabulary uniqueness and **D**e**CL**a**R**ati**VE** stylometric analysis to compare the three datasets on linguistic features.

Vocabulary uniqueness was computed for each dataset using the Jaccard’s index on the truthful and deceptive vocabulary sets. Details of this computation process are provided in Supplementary Information. The obtained Jaccard’s index values revealed the extent of similarity between vocabulary sets of truthful and deceptive statements. In the Intention and Opinion datasets, the Jaccard’s index revealed a moderate level of similarity among truthful and deceptive sets with a value of 0.34 and 0.35, respectively. Conversely, for the Memory dataset, a value of 0.46 indicated a relatively higher degree of similarity.

The linguistic style was analyzed after applying the **D**e**CL**a**R**ati**VE** stylometry technique, by which we obtained a stylistic vector of 26 linguistic features for each text of the three datasets. Subsequently, for the Memory and Intention dataset, we computed a non-parametric permutation t-test (n = 10,000) for independent samples for the 26 linguistic features to outline significant differences among the truthful and deceptive texts. For the Opinion dataset, our analysis proceeded as follows. Firstly, we computed the **D**e**CL**a**R**ati**VE** stylometry technique for all the subjects’ opinions. This resulted in a 2500 (opinions) x 26 (linguistic features) matrix. Then, since each subject provided five opinions (half truthful and half deceptive), we averaged the stylistic vector separately for the truthful and deceptive sets of opinions. This allowed us to obtain for the same subject two different averaged stylistic vectors, one for the truthful opinions and one for the deceptive opinions. Importantly, this averaging process enabled us to obtain results which are independent from the topic (e.g., abortion or cannabis legalization) and the stance taken by the subject (e.g., in favor or against that particular topic). Finally, we validated the statistical significance of these differences by conducting a non-parametric paired sample permutation test (n = 10000). Results for each dataset were corrected for multiple comparisons with Holm-Bonferroni correction.

For the three datasets, Fig. 2 shows the differences in the number, the type, the magnitude of the effect-size, and the direction of the effect for the linguistic features that survived post-hoc corrections. Magnitude of effect-size is expressed by Common Language Effect Size (CLES), which is a measure of effect size that is meant to be more intuitive in its understanding by providing the probability that a specific linguistic feature, in a picked-at-random truthful statement, will have a higher score than in a picked-at-random deceptive one [40]. To make an example of these differences, the concreteness score of words (‘concr_score’) presented the larger effect size within the Intention dataset towards the truthful statements, while in the Opinion dataset, it showed a smaller effect size towards the deceptive statements.

Overall, the Intentions dataset displayed a lower number of significant differences in linguistic features among truthful and deceptive statements than both the Opinion and Memory datasets. Notably, the CLES scores of linguistic features associated to truthfulness (Opinion CLES range: 0.56–0.69; Memory CLES range: 0.53–0.63; Intention CLES range: 0.53–0.61) were consistently larger than the CLES scores of those associated to deception (Opinion CLES range: 0.32–0.46; Memory CLES range: 0.39–0.48; Intention CLES range: 0.45 – 0.40).

Performance on the Lie-Detection classification task

This section presents the performance, in terms of averaged accuracy (and standard deviation) of the 10-folds, on the test sets after the last epoch of the small and base model in all the Scenarios.

Scenario 1.

In Table 4 are depicted the test accuracies for the FLAN-T5 model, categorized by dataset and model size in Scenario 1. In each case, the base model on average outperformed the small model, with the Memory dataset showing the largest improvement of 4% and the Intention dataset showing just a 0.06% increase of average accuracy. These results indicate that the larger model size generally leads to improved performance across the three datasets, with higher accuracy observed in the base version.

Table 3

Test accuracy of FLAN-5 Models in Scenario 2 (three combination of train sets). The performance comparison is among the small and base version of the FLAN-T5 model in the three combination of train set: opinion + memory, opinion + intention, memory + intention.

Train set	Test set	Model Size	Test Accuracy
Opinion + Memory	Intention	Flan-T5 small	55.37
		Flan-T5 base	55.67
Opinion + Intention	Memory	Flan-T5 small	55.37
		Flan-T5 base	54.23
Memory + Intention	Opinion	Flan-T5 small	53.12
		Flan-T5 base	49.40

Table 4
Test acccuracy of the FLAN-T5 models in Scenarios 1 and 3 for the three datasets. Reported values are means ± standard deviation of the 10-folds. Best results per evaluation metric are in bold.

Model	Opinion	Memory	Intention
Flan-T5 small - Scenario 1	80.64 ± 0.02	76.87 ± 0.02	71.46 ± 0.03
Flan-T5 base - Scenario 1	82.6 ± 0.03	**80.61 ± 0.01**	71.52 ± 0.02
Flan-T5 small - Scenario 3	79 ± 0.02	75.67 ± 0.02	69.32 ± 0.037
Flan-T5 base - Scenario 3	**82.72** ± **0.024**	79.87 ± 0.016	**72.25** ± **0.03**

Scenario 2.

This scenario aimed to investigate our fine-tuned LLM’s generalization capability across different deception domains. As presented in Table 3, the test accuracy for the three experiments in this scenario significantly dropped to the chance level, showing that the model in no case was not able to learn a general rule to detect lies coming from different contexts.

Scenario 3.

In Scenario 3, we tested the accuracy of the FLAN-T5 small and base version on the aggregated Opinion, Memory, and Intention datasets. The small-sized FLAN-T5 achieved an average test accuracy of 75.45% (st. dev. ± 0.016), while the base-sized FLAN-T5 exhibited a higher average test accuracy of 79.31% (st. dev. ± 0.013). In other words, the base-sized model outperformed the small model by approximately four percentage points.

Results in Table 4 show the disaggregated performance on individual datasets between the small and base FLAN-T5 models in Scenario 3, with a comparison to their counterparts in Scenario 1. These comparisons show that FLAN-T5-small in Scenario 3 exhibited worse performance than in Scenario 1. Instead, in Scenario 3, the base model barely outperformed its counterparts of Scenario 1 relatively the Opinion and Intention datasets by less than 1% and slightly underperformed its counterpart of Scenario 1 on the Memory dataset.

Explainability Analysis

This section aims to gain a deeper understanding of the top-performing model identified in Scenario 3 (Flan-T5 base) through a **D**e**CL**a**R**ati**VE** stylometric analysis of statements correctly classified and misclassified by the model. The purpose of this analysis was to examine whether the linguistic style of the input statements exerted an influence on the resulting output of the model and to provide explanations for the wrong classification outputs.

To this aim, during each iteration from cross-validation, we paired the sentences and their actual labels belonging to the test set with the corresponding labels predicted by the current model. After the cross-validation ended, for each of the ten folds and for each of the 26 linguistic features of the sentences that composed the test set of that fold, we performed a non-parametric permutation t-test for independent samples (n = 10.000) to compare:

a. truthful vs. deceptive statements that were misclassified by the model, namely, truthful statements that were predicted as deceptive (False Negatives), with deceptive statements that were predicted as truthful (False Positives);
b. deceptive statements that were correctly classified as deceptive (True Negatives) vs. truthful statements that were misclassified as deceptive (False Negatives);
c. truthful statements that were correctly classified as truthful (True Positives) vs. deceptive statements that were misclassified as truthful (False Positives).
d. truthful vs. deceptive statements correctly classified by the model (True Positives vs. True Negatives).

We present the statistically significant features that survived post-hoc correction for multiple comparisons in each fold. Overall, for comparison a), b) and c) we observed no statistically significant differences ($p < 0.05$) in any linguistic features for most of the splits with the only exception of:

1. 'fk_read' in fold 1 ($t = 5.30$; $p = 0.02$) and 'Reality Monitoring' in fold 6 ($t = 4.74$; $p = 0.03$) for the a) comparison;
2. 'Contextual Embedding' in fold 7 ($t = -2.11$; $p = 0.01$) for the b) comparison;
3. 'Reality Monitoring' in folds 4 ($t = -3.39$; $p = 0.04$), 5 ($t = -3.39$; $p = 0.046$), 6 ($t = -3.39$; $p = 0.03$), and 9 ($t = -3.39$; $p = 0.042$) for the c) comparison.

Conversely, for the d) comparison, several significant features emerged in all the folds and survived corrections for multiple comparisons. Figure 3 depicts the effect sizes of linguistic features, sorted according to the number of times they were found to be significant among the 10-folds. The effect sizes were computed as the average of the effect sizes obtained from each fold. The top six features in Fig. 3 represented a cluster of linguistic features related to the Cognitive Load framework.

Discussion

In the present research, we investigated the efficacy of a Large Language Model, specifically FLAN-T5 in its small and base version, in learning and generalizing the intrinsic linguistic representation of deception across different contexts. To accomplish this, we employed three datasets encompassing genuine or fabricated statements regarding personal opinions, autobiographical experiences, and future intentions.

Descriptive linguistic analysis was performed to compare the three datasets on linguistic features. Vocabulary uniqueness, measured with the Jaccard's index, ranged from 0.34 to 0.46, suggesting that people tend to use a different vocabulary for truthful and deceptive statements, especially when expressing opinions and intentions. We also explored differences in the **D**e**CL**a**R**ati**VE** style, i.e., analyzing

26 linguistic features extracted from the psychological frameworks of **D**istancing, **C**ognitive **L**oad, **R**eality monitoring, and **VE**rifiability approach. The linguistic features that showed statistically significant differences between truthful and deceptive statements varied across datasets in terms of total number and type of features, the magnitude of the effect size, and the direction of the effect. Notably, the effect size scores of verbal cues of truthfulness were consistently larger than those of deception, confirming that the linguistic features associated with truthfulness are more reliable and robust than those associated with deception [5].

Overall, the descriptive Linguistic Analysis of the three datasets agreed well with existing studies in cognitive and memory-oriented approaches to verbal lie detection.

In line with the cognitive load framework, we observed that truthful opinions and narratives of autobiographical memories were characterized by greater complexity and verbosity, with opinions being stylistically more authentic and memories more analytical [35, 13].

In accordance with the Reality Monitoring (RM) framework [36], that states that truthful memory accounts tend to reflect the perceptual processes involved while experiencing the event, whereas fabricated accounts are constructed through cognitive operations, genuine memories exhibited higher scores in memory-related words - reflecting individual efforts to recollect the event- and in the number of words associated with spatial and temporal information (‘Contextual Embedding’), as well as an overall higher RM score. Conversely, we found deceptive memories showed higher scores in words related to cognitive processes (e.g., reasoning, insight, causation). Furthermore, along with the Verifiability Approach, truthful memories contained more verifiable details, as indicated by the greater number of named-entities about times and locations [22, 41]. The fewer named-entities in deceptive memories may suggest that deceivers may strategically omit potentially incriminating information even in low-stake scenarios, such as in this study. However, to gain credibility, they may compensate by fostering a sense of social connection by including self-references and mentions of other individuals, which may explain why we found a greater amount of named-entities of ‘People’ and self-references. Finally, truthful memories were overall characterized by words with higher scores of concreteness, supporting Kleinberg’s *truthful concreteness hypothesis* [42] based on RM framework and Verifiability Approach.

However, the opposite trends were found for truthful and deceptive opinions. Truthful opinions about abstract concepts were characterized by a lesser number of concrete words and a greater amount of cognitive words, as also previously shown [33], reflecting the reasoning processes that truth-tellers engage in evaluating the pros and cons of abstract and controversial concepts (e.g., abortion). Conversely, deceivers tended to produce opinions more grounded in reality, as shown by higher scores in the concreteness of words, contextual details, and reality monitoring.

Finally, in line with previous literature on distancing framework [34, 43] and deceptive opinions [33, 19], deceivers utilized more other-related word classes (‘Other-reference’) and fewer self-related words (‘Self-reference’), confirming that individuals tend to avoid personal involvement when expressing deceptive statements.

Our findings on linguistic indicators of truthful and deceptive intentions are consistent with previous research claiming that genuine intentions contain more ‘how-utterances’, i.e., indicators of careful planning and concrete descriptions of activities. In contrast, false intentions are characterized by ‘why-utterances’, i.e., explanations and reasons for why someone planned an activity or for why doing something in a certain way [41]. Indeed, we observed that true intentions were more likely to provide concrete and distinct information about the intended action, grounding their statements in real-world experiences and providing temporal and spatial references. Additionally, true intentions were characterized by a more analytical style and a greater presence of numerical entities, suggesting that individuals were more involved in a specific and detailed execution of the plan. In contrast, false intentions exhibited a higher amount of cognitive words and expressions and were temporally oriented toward the present and past, suggesting that liars were more focused on providing explanations for their planning, likely in order to be believed. Furthermore, we found evidence in line with the claim that liars may over-prepare their statements [41], as indicated by higher verbosity and a greater number of self-references and mentions of people. Taken together, deceptive individuals may attempt to appear more credible, providing excessive information and creating a sense of social connection by incorporating self-references and mentions of people.

In Scenario 1, we fine-tuned FLAN-T5 in its small and base versions to perform lie detection as a classification task. This fine-tuning process yielded promising results when applied to a single dataset (i.e., opinion vs. memory vs. intention), with the base version providing a higher accuracy. The model size influenced the performance, likely because a bigger model is able to learn a better representation of linguistic patterns of genuine and deceptive narratives.

However, there are no universal rules the model can learn to distinguish truthful from deceptive statements, enabling a generalization of the task across different contexts. This consideration was highlighted in Scenario 2, in which the Flan-T5 model performed at chance level when trained on two datasets and tested on the third one (e.g., train: opinion + memory; test: intention). Given that the three datasets differ significantly in terms of the content and the linguistic style by which truthful and deceptive narratives are delivered, the model appears to engage in domain-specific learning, tailoring its classification capabilities to the specific domain of deception.

In Scenario 3, we fine-tuned Flan-T5 with the three aggregated datasets (i.e., opinion + memory + intention). Our findings demonstrate that the base version of the model is capable of effectively classifying all the datasets without compromising performance on any individual dataset when compared to Scenario 1. However, the ability to accurately detect deception in a multi-context scenario depended on the model size. Specifically, the smaller model employed in Scenario 3 exhibited a slightly reduced performance compared to Scenario 1 when evaluating the model’s accuracy on each individual dataset. This disparity in performance may be attributed to the size of the FLAN-T5 small, which hinders the capacity to learn all the distinctive features across the three datasets simultaneously. Consequently, to classify deception across different contexts, the small model must relinquish certain specialized abilities that are beneficial for specific datasets. Conversely, FLAN-T5 base, with its larger size, possesses the capability to comprehend and integrate the features of the three distinct datasets, thereby maintaining consistent performance across all individual datasets.

Findings from Scenarios 2 and 3 suggest that LLMs, despite having acquired a comprehensive understanding of language patterns, still require exposure to prior examples to accurately classify deceptive texts within different domains. Overall, the results obtained from FLAN-T5 in its small and base versions surpassed the performance of Transformer models previously employed in the literature on the Opinion [28] and Intention datasets [44].

To improve the explainability of the performance collected, we investigated whether the linguistic style that characterizes truthful and deceptive narratives could have a role in the model's final predictions. Overall, truthful and deceptive statements in the misclassified sample did not differ significantly for any linguistic feature extracted with **D**e**CL**a**R**ati**VE** stylometry technique. The only exception was in fold 1 and fold 6, which showed significant differences in text's readability and reality monitoring score, respectively. No significant differences were detected in each fold in linguistic features between deceptive statements that were correctly classified as deceptive (True Negatives) and truthful statements that were misclassified as deceptive (False Negatives), with the exception of Contextual Embedding score in fold 7. Finally, truthful statements that were correctly classified as truthful (True Positives) and deceptive statements that were misclassified as truthful (False Postives) exhibited significant differences in the Reality Monitoring score in four out of the ten folds, suggesting that deceptive statements with higher RM scores may be misleading and drive the model to classify those statements as truthful.

Altogether, most of the analyzed folds showed a complete overlap in linguistic style, suggesting that the linguistic style characterizing truthful or deceptive narratives is a feature that the model may consider for its final prediction. Therefore, the model may exhibit poor classification performance when statements possess ambiguous features, such as deceptive statements being delivered in a similar style as truthful statements.

In contrast, correctly classified statements displayed a cluster of linguistic features associated with the Cognitive Load framework [35] in most of the 10-folds, specifically low-level features related to length, complexity, and analytical style of the texts, that may have enabled the distinction between truthful and deceptive statements. According to this framework, it is plausible explanation behind these findings is that liars would experience increased cognitive load while fabricating their fake responses by checking their congruency with other fabricated information to maintain credibility and consistency [45], therefore producing shorter and less complex sentences.

At the time of writing and to the best of our knowledge, this is the first study involving the use of an LLM for a lie-detection task. The main advantage of our approach consists of its applicability to raw text without the need for extensive training or handcrafted features. We highlighted the importance of a diversified dataset to achieve a generalized good performance. We also considered crucial the balance between the diversity of the dataset and the size of the LLM, suggesting that the more diverse the dataset is, the bigger the model required to achieve higher-level accuracy. Therefore, future works could explore the inclusion of new datasets, trying different LLMs (e.g., the most recent GPT-4), different sizes (e.g., FLAN-T5 XXL version), and different fine-tuning strategies to investigate the variance in performance within a lie-detection task. Furthermore, our fine-tuning approach completely erased the previous capabilities possessed by the model; therefore, future works should focus also on new fine-tuning strategies that do not compromise the model's original capabilities.

Despite the demonstrated success of our model, three significant limitations impact the ecological validity of our findings and their practical application in real-life scenarios.

The first notable limitation pertains to the narrow focus of our study, which concentrated solely on lie detection within three specific contexts: personal opinions, autobiographical memories, and future intentions. This restricted scope limits the possibility to accurately classify deceptive texts within different domains. A second limitation is that we exclusively considered datasets developed in experimental set-ups designed to collect genuine and completely fabricated narratives. However, individuals frequently employ embedded lies in real-life scenarios, in which substantial portions of their narratives are true, rather than fabricating an entirely fictitious story. Finally, the datasets employed in this study were collected in experimental low-stake scenarios, in which the participants had low incentives to lie and appear credible. Because of all the above issues, the application of our model in real-life contexts may be limited, and caution is advised when interpreting the results in such situations. These limitations underscore the need for future research to address these concerns and expand the applicability and generalizability of lie-detection models in real-life settings.

Declarations

Data availability.

For the Opinion dataset we obtained the full access after contacting the corresponding author. The Memory dataset is downloadable at the link: https://msropendata.com/datasets/0a83fb6f-a759-4a17-aaa2-fbac84577318. The Intention dataset is publicly available at the link: https://osf.io/45z7e/.

Code availability.

All the Colab Notebooks to perform linguistic analysis on the three datasets, fine-tune the model in the three Scenarios, and conduct explainability analysis is available at https://github.com/robecoder/VerbalLieDetectionWithLLM.git

Acknowledgements

We would like to thank Bruno Verschuere and Bennett Kleinberg for sharing with us the full version of their Intention dataset.

Author Contributions

G.S. conceptualized the research. R.L., R.R., P.C., and G.S. designed the research. P.C. shared the updated version of the Deceptive Opinion Dataset. R.L. performed the descriptive linguistic analysis and explainability analysis. R.R. developed and implemented the fine-tuning strategy. R.L. and R.R. wrote the paper. P.P. and G.S. supervised all aspects of whole the research and provided critical revisions.

Additional Information

Competing interests: The author(s) declare that no competing interests.

References

1. Walczyk, J. J., Harris, L. L., Duck, T. K., & Mulay, D. A social-cognitive framework for understanding serious lies: Activation-decision-construction-action theory. *New Ideas in Psychology*, **34,** 22–36. https://doi.org/10.1016/j.newideapsych.2014.03.001 (2014).
2. Amado, B. G., Arce, R., & Fariña, F. Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review. *The European Journal of Psychology Applied to Legal Context*, **7**, 3–12. https://doi.org/10.1016/j.ejpal.2014.11.002 (2015).
3. Vrij, A., Granhag, P. A., Ashkenazi, T., Ganis, G., Leal, S., & Fisher, R. P. Verbal lie detection: Its past, present and future. *Brain Sciences*, **12**, 1644. https://doi.org/10.3390/brainsci12121644 (2022).
4. Vrij, A., & Fisher, R. P. Which lie detection tools are ready for use in the criminal justice system? *Journal of Applied Research in Memory and Cognition*, **5,** 302–307. https://doi.org/10.1016/j.jarmac.2016.06.014 (2016).
5. DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. Cues to deception. *Psychological Bulletin*, **129,** 74–118. https://doi.org/10.1037/0033-2909.129.1.74 (2003).
6. Bond, C. F., Jr., & DePaulo, B. M. Accuracy of deception judgments. *Personality and Social Psychology Review*, **10,** 214–234. https://doi.org/10.1207/s15327957pspr1003_2 (2006).
7. Levine, T. R., Park, H. S., & McCornack, S. A. Accuracy in detecting truths and lies: Documenting the "veracity effect." *Communication Monographs*, **66,** 125–144. https://doi.org/10.1080/03637759909376468 (1999).
8. Levine, T. R. Truth-Default theory (TDT). *Journal of Language and Social Psychology*, **33,** 378–392. https://doi.org/10.1177/0261927x14535916 (2014).
9. Street, C. N. H., & Masip, J. The source of the truth bias: Heuristic processing? *Scandinavian Journal of Psychology*, **56,** 254–263. https://doi.org/10.1111/sjop.12204 (2015).
10. Chen, X., Hao, P., Chandramouli, R., and Subbalakshmi, K. P. "Authorship Similarity Detection from Email Messages," in *International Workshop On Machine Learning And Data Mining In Pattern Recognition*. Editor P. Perner (New York, NY: Springer), 375–386. https://doi.org/10.1007/978-3-642-23199-5_28 (2011).
11. Chen, H. Dark web: Exploring and mining the dark side of the web. In *2011 European Intelligence and Security Informatics Conference,* 1-2. IEEE. (2011, September).

12. Daelemans, W. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing*, 451–462. Springer Berlin Heidelberg. http://dx.doi.org/10.1007/978-3-642-37256-8_37 (2013).
13. Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and social psychology Review*, **19,** 307-342. https://doi.org/10.1177/1088868314556539 (2015).
14. Tomas, F., Dodier, O., & Demarchi, S. Computational measures of deceptive language: Prospects and issues. *Frontiers in Communication*, **7.** https://doi.org/10.3389/fcomm.2022.792378 (2022).
15. Conroy, N. K., Rubin, V. L., & Chen, Y. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology*, **52,** 1–4. https://doi.org/10.1002/pra2.2015.145052010082 (2015).
16. Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. Automatic detection of fake news. *arXiv preprint arXiv:1708.07104* (2017).
17. Fornaciari, T., & Poesio, M. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, **21,** 303–340. https://doi.org/10.1007/s10506-013-9140-4 (2013).
18. Yancheva, M., & Rudzicz, F. Automatic detection of deception in child-produced speech using syntactic complexity features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* **1,** 944-953, (2013, August).
19. Pérez-Rosas, V., & Mihalcea, R. Experiments in open domain deception detection. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. http://dx.doi.org/10.18653/v1/d15-1133 (2015).
20. Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*. (2011).
21. Fornaciari, T., & Poesio, M. Identifying fake Amazon reviews as learning from crowds. *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. http://dx.doi.org/10.3115/v1/e14-1030n (2014).
22. Kleinberg, B., Mozes, M., Arntz, A., & Verschuere, B. Using named entities for computer-automated verbal deception detection. *Journal of forensic sciences,* **63,** 714-723. https://doi.org/10.1111/1556-4029.13645 (2017).
23. Mbaziira, A. V., & Jones, J. H. Hybrid text-based deception models for native and Non-Native English cybercriminal networks. *Proceedings of the International Conference on Compute and Data Analysis*. http://dx.doi.org/10.1145/3093241.3093280 (2017, May 19).
24. Levitan, S. I., Maredia, A., & Hirschberg, J. Linguistic cues to deception and perceived deception in interview dialogues. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* **1.** http://dx.doi.org/10.18653/v1/n18-1176 (2018).
25. Kleinberg, B., Nahari, G., Arntz, A., & Verschuere, B. An investigation on the detectability of deceptive intent about flying through verbal deception detection. *Collabra: Psychology*, **3.**

https://doi.org/10.1525/collabra.80 (2017).

26. Constâncio, A. S., Tsunoda, D. F., Silva, H. de F. N., Silveira, J. M. da, & Carvalho, D. R. Deception detection with machine learning: A systematic review and statistical analysis. *PLOS ONE*, **18,** e0281323. https://doi.org/10.1371/journal.pone.0281323 (2023).
27. Zhao, W. X., *et al.* A survey of large language models. *arXiv preprint arXiv:2303.18223*. (2023).
28. Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 1423-1430, (2020, May).
29. Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 1970-1978, http://dx.doi.org/10.18653/v1/2020.acl-main.178 (2020, July).
30. Kleinberg, B., & Verschuere, B. How humans impair automated deception detection performance. *Acta Psychologica*, **213**, https://doi.org/10.1016/j.actpsy.2020.103250 (2021).
31. Hernández-Castañeda, Á., Calvo, H., Gelbukh, A., & Flores, J. J. G. Cross-domain deception detection using support vector networks. *Soft Computing*, **21,** 585–595. https://doi.org/10.1007/s00500-016-2409-2 (2016).
32. Pérez-Rosas, V., & Mihalcea, R. Cross-cultural deception detection. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* **2**. http://dx.doi.org/10.3115/v1/p14-2072 (2014).
33. 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* 309-312. http://dx.doi.org/10.3115/1667583.1667679 (2009, August).
34. Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, **29,** 665–675. https://doi.org/10.1177/0146167203029005010 (2003).
35. Vrij, A., Fisher, R., Mann, S., & Leal, S. A cognitive load approach to lie detection. *Journal of Investigative Psychology and Offender Profiling*, **5,** 39–43. https://doi.org/10.1002/jip.82 (2008).
36. Johnson, M. K., & Raye, C. L. Reality monitoring. *Psychological Review*, **88,** 67–85. https://doi.org/10.1037/0033-295x.88.1.67 (1981).
37. Nahari, G., Vrij, A., & Fisher, R. P. Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, **19,** 227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x (2012).
38. Vrij, A., & Nahari, G. The verifiability approach. In *Evidence-Based Investigative Interviewing,*116–133. Routledge. http://dx.doi.org/10.4324/9781315160276-7 (2019).
39. Chung, H. W., *et al.* Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*. (2022).

40. McGraw, K. O., & Wong, S. P. A common language effect size statistic. Psychological bulletin, **111,** 361. https://doi.org/10.1037/0033-2909.111.2.361 (1992).

41. Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A., & Verschuere, B. Automated verbal credibility assessment of intentions: The model statement technique and predictive modeling. *Applied Cognitive Psychology*, **32,** 354–366. https://doi.org/10.1002/acp.3407 (2018).

42. Kleinberg, B., van der Vegt, I., & Arntz, A. Detecting deceptive communication through linguistic concreteness. Center for Open Science. http://dx.doi.org/10.31234/osf.io/p3qjh (2019).

43. Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. Discourse Processes, **45,** 1–23. https://doi.org/10.1080/01638530701739181_(2007).

44. Ilias, L., Soldner, F., & Kleinberg, B. Explainable Verbal Deception Detection using Transformers. *arXiv preprint arXiv:2210.03080*. (2022).

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

Figures

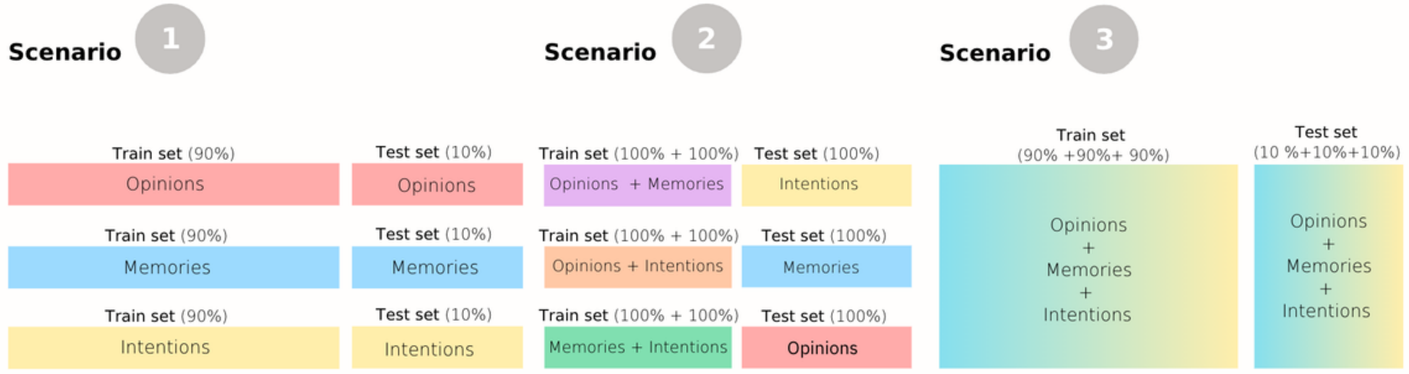

Figure 1

Visual illustration of the train-test splits in the three Scenarios. In brackets the percentages of the train-test splits.

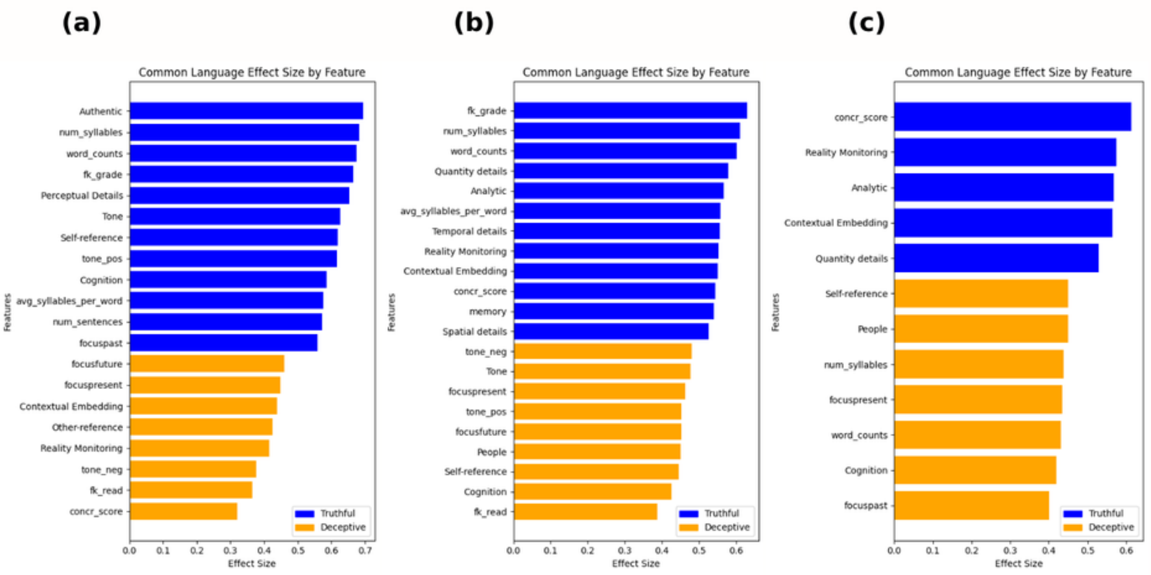

Figure 2

Common Language Effect Size of linguistic features that survived post-hoc corrections, sorted in descending order (from the larger CLES to the smaller). Linguistic features that were on average higher in truthful texts are shown in blue, while those that were on average higher in deceptive texts are shown in orange.
From left to right: significant linguistic features for Opinions (a), for Memories (b) and for Intentions (c).

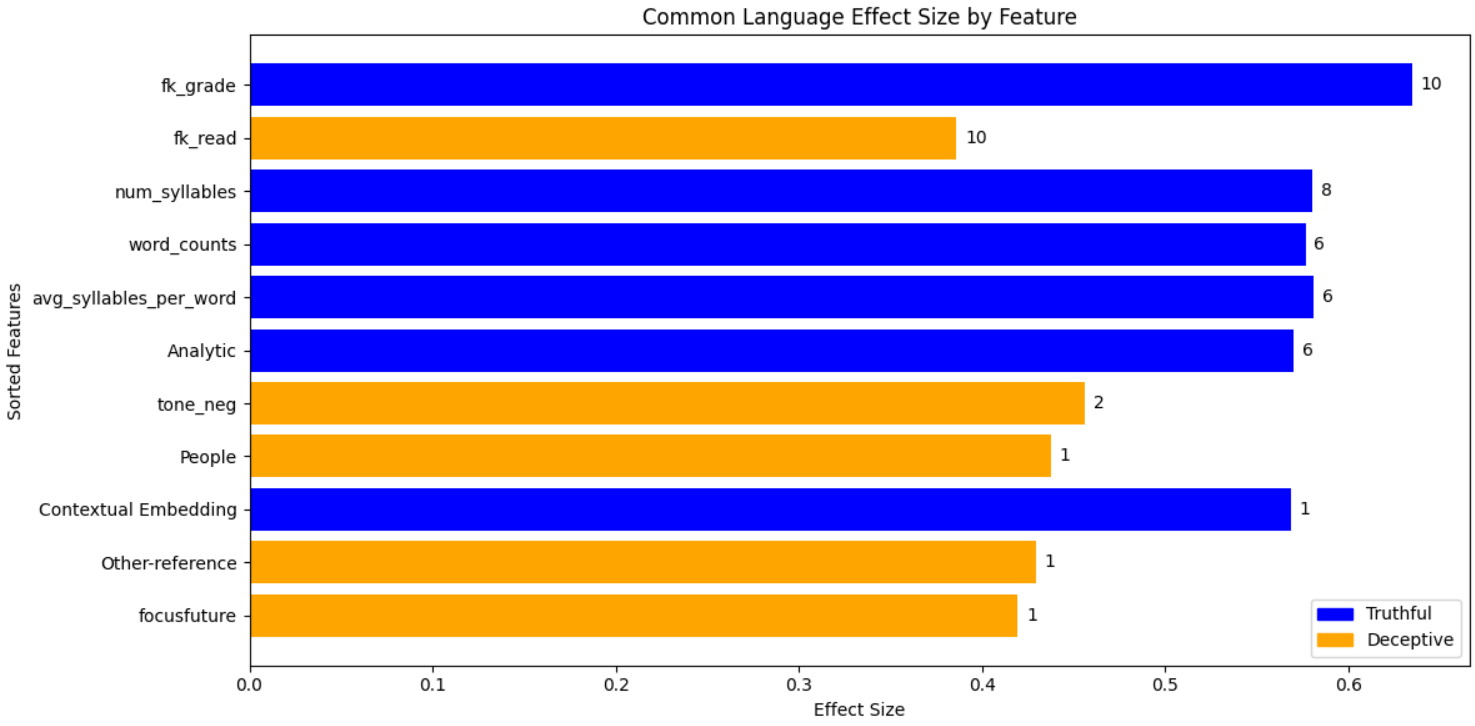

Figure 3

Linguistic features in Truthful and Deceptive statements that were accurately classified by FLAN-T5 base in Scenario 3. The bar plot shows the averaged Common Language Effect Size among the 10-folds of linguistic features that survived post-hoc corrections. Linguistic features are sorted in descending order, according to the number times they were found to be significant among the 10-folds (displayed at the side of each bar). Linguistic features that are higher on average in truthful texts are shown in blue, while those that are higher on average in deceptive texts are shown in orange.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.pdf