Domain-Independent Deception: Taxonomy, Datasets and Linguistic Analysis

Rakesh Verma, Nachum Dershowitz, and Xuting Liu

November 15, 2021

Abstract

Internet-based economies and societies are drowning in deceptive attacks such as fake news, phishing, and others, which we call domains of deception. Machine Learning (ML) and Natural Language Processing (NLP) researchers have been attempting to ameliorate this precarious situation by designing domain-specific detectors. Only a few recent works have considered domain-independent deception. We collect these disparate threads of research and investigate domain-independent deception along four dimensions: definition, taxonomy, datasets and linguistic analysis. First, we give a new *computational* definition of deception and a new taxonomy for it. Second, we examine the debate on linguistic cues for deception. Third, we build/collect datasets of deceptive attacks and perform rigorous linguistic analysis of these datasets. We argue that ML/NLP research and surveys on domain-specific and domain-independent deception have the potential for social good, but to redeem their promises much more care and attention are required than is prevalent. Finally, we integrate these datasets and create a diverse, larger dataset, which we propose to share with the community. We find that several linguistic features fail across different domain(s), but a few do survive.

> The Pretences of the ancient and the modern Zamzummim to that Ray of Divinity, were, and are, Deceptions.
>
> John Dove, *An Essay on Inspiration* (1756)

1 Introduction

Trust in Internet-dependent societies and economies is eroding rapidly due to the proliferation of deceptive attacks, such as fake news, phishing and disinformation. For example, the situation has deteriorated so much in the US and UK that a significant fraction of the US population believes that the 2020 US election was stolen, and, in the UK, Brexit has been credited to a deliberate disinformation campaign.

Social-media platforms have come under severe scrutiny regarding how they police content [19,30]. Facebook and Google are partnering with independent fact-checking organizations that typically employ manual fact checkers. Natural-language processing (NLP) and machine learning (ML) researchers have joined the effort to combat this dangerous situation by designing fake news [46, 62], phishing [53] and other kinds of domain-specific detectors.

We think that applying multiple detectors for different kinds of deception may not be optimal. If they are composed sequentially, then more time is needed. If they are composed in parallel, more hardware is needed. Moreover, there may be new situations involving lies in which labeled data may not be available. Hence, we would like to build domain-independent deception detectors.

Unfortunately, this research is currently hampered by the lack of: (A) good definition/taxonomy, (B) high quality datasets, and (C) systematic approaches to domain-independent deception detection. Thus, results are neither generalizable nor reliable, leading to much confusion.

In what follows, we elaborate on these challenges and make the following contributions:

1. We propose a new computational definition of deception[1] (Section 1.1) and a new taxonomy for it (Section 2).
2. We examine the debate on linguistic deception detection, identify works that show up the challenges that must be overcome to develop domain-independent deception detectors, and examine them critically.
3. We develop a quality dataset for deception that consists of a diverse range of deceptive attacks.
4. We provide a rigorous linguistic analysis of four datasets covering different domains and discover some common traits of deceptive attacks.

We hope that this paper, besides scrutinizing the claims on general linguistic signals for deception, will also aid those planning to conduct systematic reviews of NLP literature and the literature of related subfields of computer science. A Google Scholar search with phrase queries of the form "guidelines for systematic literature reviews in X," where X $\in$ { natural language processing, NLP, machine learning } returned nothing.[2]

The rest of this paper is organized as follows. In the rest of this section, we first give a new, computational definition of deception and then a new taxonomy for it. Section 2 presents the related work on domain-independent deception. In Section 3 we examine the debate on linguistic cues for deception detection. Section 4 presents our arguments for markers of domain-independent deception and Section 5 presents the datasets and their linguistic analysis. Section 6 presents the conclusions and some directions for future work.

1.1 Definition

We first examine a general definition of deception [15] intended to capture a wide variety of deceptive situations and attacks.

Definition 1.1 *Deception is an intentional act of manipulation to gain compliance. Thus, it has at least one source, one target and one goal. The source is intentionally manipulating the target into beliefs, or actions, or both, intended to achieving the goal(s).*

We refer the reader to [15] for a discussion on several other definitions of deception and their pros and cons.

Since we are interested in automatic verifiability, we modify this definition of deception and propose one that is computationally feasible. As intentions are notoriously hard to establish, we replace it with an exposure of manipulation/goal(s) clause. Our revised definition is the following:

Definition 1.2 (Deception) *Deception is an act of manipulation designed to gain compliance such that were the manipulation or the goal(s) of compliance exposed, the chances of compliance would decline significantly. Thus, it has at least one source, one target and one goal. The source is manipulating the target into beliefs, or action, or both, intended to achieve the goal(s).*

One might require that the goal(s) be harmful to an individual or an organization. However, this would necessitate either a computational definition of harm, or a comprehensive list of potential harms, which could be checked computationally, and is therefore a less desirable alternative. Of course, there remains the task of showing a significant decline in the compliance probability. We argue that this would be usually evident from the exposure of manipulation and/or the goal(s). If the manipulation is non-trivial or the goal(s) malicious, then we expect a significant decline in the probability of success of the deceptive attack upon the exposure of the deceptive method(s) or goal(s).

[1] Whenever we use the term "deception" without qualifying adjectives, we mean domain-independent deception. Hence, we will use the term "lies" instead of "deception" whenever authors have not been careful about the goals of the deception.

[2] When we dropped the "in X" part, we did get some guidelines for reviews of software engineering literature [27], agile software and the like.

1.2 Taxonomy

Before constructing our own taxonomy based on the new definition of deception, we searched for previous ones. There have been a few attempts at constructing taxonomies for fake news or other forms of deception, e.g., phishing. Molina et al. [29] give a taxonomy of *fake news* with four dimensions, viz., (a) message and linguistic, (b) sources and intentions, (c) structural and (d) network. In Kapantai et al. [26] a valiant effort is made to come up with a unified taxonomy of *disinformation*. They conducted a systematic search for papers that have proposed taxonomies for disinformation. Based on their analysis of previous taxonomies, they proposed a taxonomy with three dimensions: (a) facticity, (b) motivation and (c) verifiability.

Yet no one, to our knowledge, has given a comprehensive taxonomy for deception in the real world.[3] We put forward such a taxonomy, which has the following ten dimensions:

1. *Source*: This includes: human (individual or group), bot, etc., or mixed, i.e., combinations such as human assisted by a bot.

2. *Target*: This includes: human (individual or group), automatic detector, or both. For example, spam targets automatic detectors and phishing targets the human but needs to fool automatic detectors also.

3. *Stratagem(s):* This includes: falsification, distortion, taking words out of context, persuasion techniques, and combinations thereof. These are all collectively referred to as implicit or explicit misrepresentations. Persuasion techniques include: authority, scarcity, social proof, reward claims, etc. (see [4]).

4. *Goal(s):* This includes a wide range of objectives, e.g., stealing money or identity information, malware installation, manipulation of vote, planting fear, sowing confusion, initiating chaos, gaining an unfair edge in a competition (e.g., swaying opinions and preferences on products), persuading people to take harmful actions, etc. The goal may also include winning board games or satisfying participation, e.g., in a laboratory experiment participants may asked to lie.

5. *Motivation:* This is the rationale for the goal(s).

6. *Dissemination:* This dimension includes aspects such as whether the targets are a few specific individuals or general classes of people. It also refers to how the deception is conveyed to the target(s).

7. *Modality:* This dimension refers to the presentation of the deceptive content. It includes: face-to-face communication (speaker+speech), audio only, textual (a.k.a. verbal) deception (e.g., SMS/email), visual only (e.g., images or videos), and combinations thereof. For example, audio-visual means it has both speech and visual components, but lacks face-to-face communication in which gestures could be used to facilitate deception.

8. *Facticity:* Can we establish whether it is factual or not? For example, currently we are unable to establish the truth or falsity of utterances such as, "There are multiple universes in existence right now."

9. *Verifiability:* Assuming facticity, how easy or difficult it is to verify whether it is legitimate or deceptive? In this paper, we are interested in machine or automatic verification. If a simple machine learning algorithm can detect it with high (some threshold, say over 95%) recall and precision, we will deem it easy.

10. *Timeliness*: This includes whether the deception occurs in an interactive or non-interactive manner. A real-time interview or debate is an interactive scenario, whereas an Amazon Mechanical Turker (AMT) typing a deceptive opinion or essay is a non-interactive one.

In this paper, we use the term "domain" to refer to the goal of deception. Therefore, when we use the term domain-independence we mean that the goals of deception can be quite different.

[3] Artificial, laboratory situations in which participants are compelled to lie for collecting datasets or some other purpose can also be handled without much modification.

2 Related Work on domain-independent deception

The related work on deception detection can be categorized into: Datasets, Detection and Literature Reviews. Of the last category, we focus on reviews of the linguistic deception detection literature here. The DBLP query, "domain decepti" (since authors can use "deception" or "deceptive"), on 9 November 2021, gave 16 matches of which nine were deemed relevant.

Remark. Unfortunately, previous researchers have generally left the term "domain" undefined. Hence, terms such as "cross-domain deception" in previous work could mean that the topics of essays or reviews are varied but the goal(s) could stay pretty much the same.

2.1 Datasets

Several datasets have been collected for studying lies. However, researchers have not carefully delineated the scope by considering the goals of the deception. For example, Zhou et al. [61] paired students and asked one student in each pair to deceive the other using messages. There are multiple datasets for fake news detection, opinion spam (a.k.a. fake reviews) detection and for phishing [1]. Next, we discuss datasets, where the term "domain" is used in the topic sense.[4]

In [34], researchers collected demographic data and 14 short essays (7 truthful and 7 false) on open-ended topics by 512 Amazon Mechanical Turkers. We refer to this as the *Open-domain* dataset. They tried to predict demographic information and facticity. In [35], researchers collected short essays on three topics: abortion, best friend, and death penalty by people from four different cultural backgrounds. The definition of domains in these two papers, viz., topics, is finer-grained than ours, viz., wherein the goal(s) of deception are varied not just the topical content.

In [24], researchers analyzed three datasets: a two class, balanced-ratio 236 Amazon reviews dataset, a hotel opinion spam dataset consisting of 400 fabricated opinions from Amazon Mechanical Turkers and 400 reviews from TripAdvisor (likely to be truthful), and 200 essays from [35]. In [57], researchers studied a masking technique on two datasets: a hotel, restaurant and doctor opinion spam dataset and the dataset from [35]. The idea is to mask the content that is not relevant in deception detection. In [5] in-domain experiments were done with positive and negative hotel opinion spam dataset and cross-domain experiments were conducted with the hotel, restaurant and doctor opinion spam dataset. In [6], truthful and deceptive opinions on five topics are collected in multiple languages. In these papers also we see that the topics of deception are varied rather than the goal(s).

In [17], researchers refer to different social networks as domains, e.g., Twitter versus Reddit.

To our knowledge, the following works have developed domain-independent deception datasets in our sense of domain-independence, wherein the goal(s) of deception can be quite different: [40, 44, 55, 57, 58].

In [40], researchers used two datasets: the American English subset of [35] consisting of a balanced-ratio 600 essays and transcriptions of 121 trial videos (60 truthful and 61 deceptive), which we call Real-life_Trial below.

In [55], researchers used three datasets: positive and negative hotel reviews, essays on emotionally-charged topics, and personal interview questions.

In [57], multiple fake news datasets, a Covid19 dataset and some micro-blogging datasets are collected and analyzed.

In [44], researchers collected fake news, twitter rumor and spam datasets.[5] They applied their models trained on these datasets to a new Covid 19 dataset.

In [58], researchers collected seven datasets (Diplomacy, Mafiascum, Open-domain, LIAR, Box of Lies, MU3D and Real-life_Trial) and analyzed them using LIWC categories. In [58], researchers do not claim domain independence or cross-domain analysis. However, their datasets do involve different goals, e.g.,

[4] Note that the topics can vary in a heterogeneous application, e.g., fake news detection, since some items could be on sport and some on politics or religion. Moreover, the goals could be different too. Hence, we will not use the term "domain" to refer to applications such as fake news.

[5] Spam is essentially advertising and deception is employed to fool automatic detectors rather than the human recipient of the spam. We focus more on human targets in this paper.

LIAR includes political lies with the goal of winning elections, whereas the lies in Real-life_Trial have other goals, and Diplomacy/Mafiascum are about winning online games.

It is clear from the above discussion that we still lack large, comprehensive datasets for deception that have a wide variety of deceptive goals. We will elaborate on this in Section 5.

2.2 Detection

Deception detection in general is a useful and challenging open problem. There have been many attempts at specific applications, such as phishing and fake news. Indeed, there are more than 1,100 DBLP results, including at least 10 surveys and reviews, on phishing, an attack that involves deception and persuasion (query: phish). Similarly, there are 419 papers on scams, 81 on opinion spam, 74 on fake reviews, and 536 on fake news.

The only works on domain-independent deception detection are already covered in the Datasets section. However, there are some reviews, surveys and meta-analysis of linguistic cues for deception detection.

2.3 Reviews on Linguistic Deception Detection

Recently, Gröndahl and Asokan [20] conducted a survey of the literature on deception. They defined implicit and explicit deception,[6] focused on automatic deception detection using input texts, and then proceeded to review 17 papers on *linguistic* deception detection techniques. These papers covered two forms of deception: (a) dyadic pairs in the laboratory, where one person sends a short essay or message to another (some truthful and some lies), and (b) fake reviews (a.k.a. opinion spam). Based on their analysis of the literature on laboratory deception experiments and the literature on opinion spam, they concluded that *there is no linguistic or stylistic trace that works for deception in general.*

Similarly, the authors of [55] assert that extensive psychology research [12, 56] shows that "generalized linguistic cue to deception is unlikely to exist." As we explain later, this statement and the conclusion from the deception survey suffer from a common fault, namely, confirmation bias. There are also other issues, which we will discuss.

In subsequent sections, we collectively refer to the deception survey of [20] and the triad of papers [12, 55, 56] as the *Critiques*. We argue that, at best, their analyses and conclusion may be a bit too hasty. We elaborate on several aspects that need further and deeper investigation/analysis with specific examples from the reviewed literature. Although we focus on those specific critiques here, many of the issues we raise are more generally applicable to any systematic review of scientific literature.

3 The Debate on Linguistic Cues for Domain-independent Deception Detection

We begin with some general guidelines for systematic reviews and then focus on the debate on linguistic cues for domain-independent deception detection.

3.1 Guidelines for Systematic Reviews

One may observe a recent explosion in systematic reviews on all kinds of problems in natural language processing (NLP) and machine learning. When a systematic review is carefully done, it can be very beneficial in organizing the literature, for both students and professionals, and in highlighting the key results and gaps in the literature. However, in the computer science literature in general, and in NLP, there is a dearth of good guidelines and procedures for such systematic reviews.

According to [48], "A systematic review is a defined and methodical way of identifying, assessing, and analyzing published primary studies in order to investigate a specific research question." Such a review

[6] Explicit deception is when the deceiver explicitly mentions the false proposition in the deceptive communication.

can reveal the structure and patterns of existing research, and identify gaps that can be filled by future research [27,48]. Furthermore, per [48], systematic reviews are formally planned and methodically executed.

Evaluating Reviews

A good systematic review is independently replicable, and so has additional scientific value over that of a literature survey. In collecting, evaluating, and documenting all available evidence on a specific research question, a systematic review may provide a greater level of validity in its findings than might be possible in any individual study reviewed. However, systematic reviews obviously require much more effort than ordinary literature surveys.

The following features differentiate a systematic review from a conventional one (Kitchenham, 2004):

- Definition and documentation of a systematic review protocol *in advance*[7] of conducting the review, to specify the research questions and the procedures to be used to perform the review.
- Definition and documentation of a search strategy as part of the protocol, to find *as much of the relevant literature as possible.*
- Up front description of the explicit inclusion and exclusion criteria as part of the protocol, to be used to assess each potential study.
- Description of quality assessment mechanisms as part of the protocol, to evaluate each study.
- Description of review and cross-checking processes as part of the protocol, and involving multiple independent researchers, to control researcher bias.

3.2 The Debate on Linguistic Cues

With respect to the deception survey of [20], we observe that they do specify the research questions and hypotheses and involve two researchers (presumably mentor and mentee). However, we also note that: (1) no review protocol is presented, (2) no search strategy is defined, (3) no inclusion/exclusion criteria are explicated, and (4) no quality assessment mechanism is specified. Thus, it is probably better to view their paper as a conventional literature review. However, as it is published in an influential journal, it is likely to leave a lasting impression on deception researchers. Hence, we believe it is worth the time and effort to examine its strengths and weaknesses more closely.

The statements of [56] and [12] stem from a meta-analysis conducted by [9]. On the positive side, a meta-analysis can be very useful and adds statistical analysis methods to a systematic review. However, this meta-analysis is now quite dated and suffers from confirmation bias among other issues as we outline below.

Next, we consider the issues and challenges that can arise with systematic reviews in general and then those that are specific to the papers under consideration.

3.3 Issues and Challenges

We enumerate several challenges with reviews and surveys, whether they are systematic or conventional, emphasizing those that are common to the Critiques.

3.3.1 Publication Bias

Not having a clear, explicit search strategy for literature and the lack of clearly defined inclusion and exclusion criteria can lead to a study that displays biases regarding the publications that are covered. We note that the deception survey [20] suffers from this issue. Although their goal was to survey automatic linguistic

[7] Emphasis added.

Table 1: The nine most prolific authors in the textual deception detection literature studied in [20].

Author	Papers
Judee K. Burgoon	4
Jeffrey T. Hancock	4
Jay F. Nunamaker Jr.	3
Lina Zhou	3
Doug P. Twitchell	3
Myle Ott	3
Claire Cardie	3
Yejin Choi	2
Tiantian Qin	2

deception detection literature, they missed not just many relevant papers, including [10, 14, 28, 43, 60], but also the meta-analysis of [23].[8]

Moreover, this meta-analysis examined 79 cues from 44 different studies on automatic linguistic deception detection. They state: "the meta-analyses demonstrated that, relative to truth-tellers, liars experienced greater cognitive load, expressed more negative emotions, distanced themselves more from events, expressed fewer sensory–perceptual words, and referred less often to cognitive processes. However, liars were not more uncertain than truth-tellers. These effects were moderated by event type, involvement, emotional valence, intensity of interaction, motivation, and other moderators. Although the overall effect size was small, theory driven predictions for certain cues received support."

However, there is another, more serious issue. Reviewing only papers that are published means ignoring papers that remain unpublished for one reason or another. For example, positive studies are more likely to be published than negative studies, papers in English have a higher chance of being published as also papers authored/coauthored by researchers who are already highly reputed, etc. Moreover, longer works such as theses and dissertations are also missed in the emphasis on published literature.

To study if this bias exists in the Critiques, we did a systematic search of the Proquest Global Database [37]. We identified 117 dissertations and theses with the keywords "deception" *and* "detection" in the title. Three of these also have the word "linguistic" in the title and all three are relevant [13, 23, 25].[9] Replacing linguistic with "natural language processing" (or "textual") and keeping "deception" yielded two more relevant dissertations [8, 36]. Finally, "verbal" with "deception" yielded four more relevant results, viz., [3, 33, 39, 51], out of nine total. The last three searches looked for both keywords in the title only.

None of the above dissertations are cited in the Critiques, although a paper by the author of [25] does appear in [12] and the author of [3] appears as a coauthor on some papers cited in [56]. The meta-analyses of [9], with more than 80 studies covered, and of [23], 44 studies covered, are much more systematic and comprehensive. For example, the queries used are listed in [23] and the inclusion and exclusion criteria are spelled out. Publication bias is mentioned in the meta-analysis of [23] and an effort is made to select independent studies. However, the meta-analysis of [9] also did not check for publication bias.

There is another kind of publication bias, when a review focuses too much on papers from a clique of interconnected researchers, or papers that analyze the same dataset multiple times. We observe this kind of bias in the deception survey [20]. To see this, we listed all the authors of the 17 papers cited in Section 2 ("Deception Detection Via Text Analysis") of their paper. There are 56 authors in total, but only 38 unique names. We report all authors with more than one publication in Table 1. Next, we analyze papers with common authors in the graph of Figure 1. It shows several cliques, two of them as large as a K_4 (complete graph on 4 vertices) and one K_3 that is not a subgraph of the K_4's.

[8] Since the [20] paper was revised in 2019, we used 2018 as the cutoff for listing missing literature.

[9] There are 34 with "deception" and "detection" in the title and "linguistic" anywhere in the document.

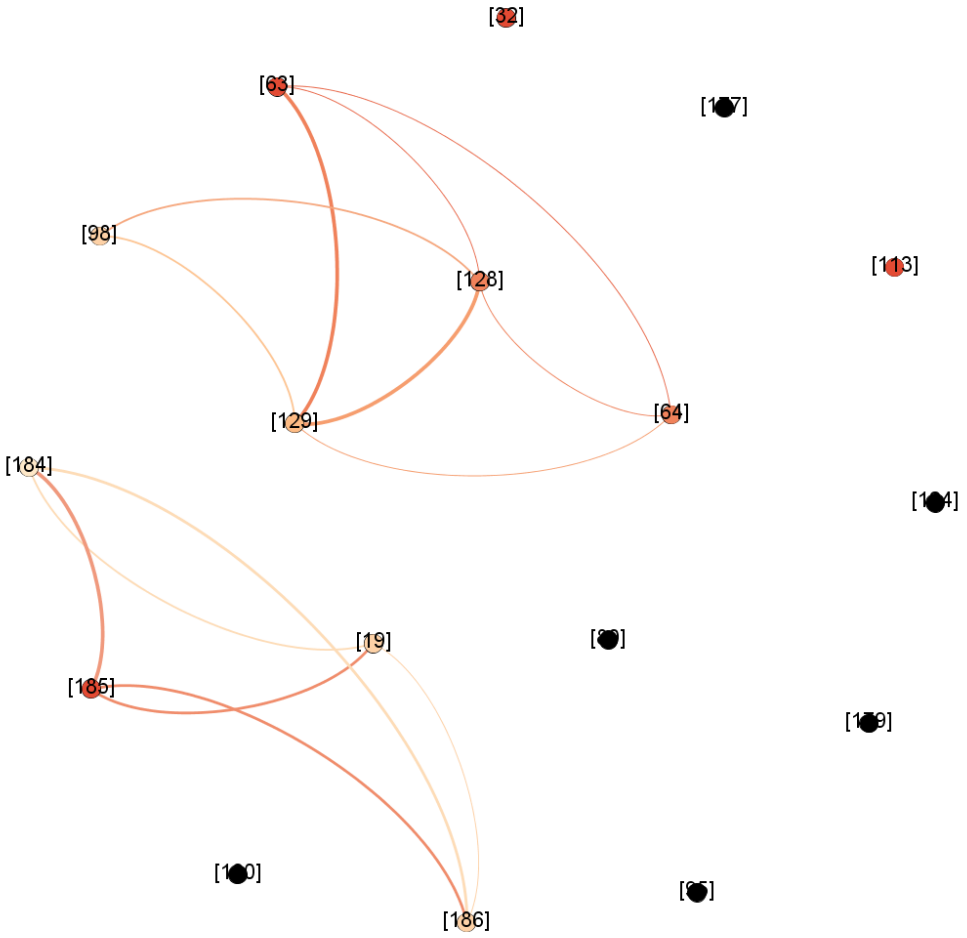

Figure 1: Graph showing the 17 papers as vertices. There is an edge between two papers (vertices) provided they have a common author. The thickness of the edge and color reflect the number of common authors. Several cliques (K_3 and K_4) are visible.

3.3.2 Confirmation Bias

Next, we discuss perhaps the most serious issue with the Critiques, which is confirmation bias. What do we mean by this? Here is the crux of the matter. None of the papers examined in the meta-analysis conducted by [9] and the deception survey by [20] built a general dataset for different deception goals (e.g., as in phishing, fake news *and* crime reports). None. If a group of researchers study a particular form of deception and build a dataset to study it, the chance that they would stumble upon general linguistic cues for deception is likely to be small, since that was not even their objective anyway! Hence, a review of these papers is also unlikely to find any general linguistic cues for deception.

3.3.3 Quality of Studies and Datasets

Another issue is the quality of the studies and the datasets collected. Quality of studies includes many different factors such as: (i) the design of the experiments, (ii) the sizes and the heterogeneity of the populations, (iii) whether the statistical tests used are appropriate for the datasets analyzed, whether tests of statistical significance were applied and correctly reported so that effect sizes can be obtained, (iv) whether something like the Bonferroni-Holm [2] correction was used for the multiple comparisons issue, and (v) their replicability.

Another issue that must be considered is whether the participants, typically undergraduates, in laboratory experiments are as motivated as real-world attackers in carrying out the deception [22].

3.3.4 Datedness

We note that the meta-analysis of [9] was conducted in 2003. The meta-analysis of [23] is more recent, but covers papers up to February 2012. The latest review of meta-analyses [49] on deception detection lists more than 50 meta-analyses. Of course, not all are relevant to linguistic deception detection, but this points to the large volume of work in the field and is indirect evidence for the inadequacy of the literature cited in the Critiques.

3.3.5 New Developments in NLP

Computer science, machine learning, and NLP have come a long way since 2012. There has been an explosion of progress in machine learning and natural language processing that we examine next. Recent breakthroughs in NLP such as attention, transformers, and new language models, such as BERT, have all occurred in the last 3–4 years. Even were the previous critiques valid, with the latest advances, there is a need for deeper investigation.

Next, we examine the positive case in favor of existence of general linguistic cues of deception.

4 Arguments for Domain-independent Deception Markers

In contrast to the assertion in the Critiques, there are several arguments in favor of general linguistic markers for deception.

4.1 Prior Analyses

First, the meta-analyses of [9] and [23] did find markers of deception in the studies they examined. Although the effect sizes were low to moderate, bear in mind that they conducted these meta-analyses on papers that studied specific forms or situations of deception and did not build any general domain independent datasets.

Second, the following papers all point to evidence for cross-domain deception detection: [40, 44, 55, 57, 58]. As mentioned above, these researchers have created domain-independent datasets and developed features and techniques for deception detection across domains.

Table 2: Comparing the latest surveys, reviews and meta-analysis on automatic deception detection. QL/DB – Whether the queries/databases searched are listed. Period – The time period of searches listed in the paper. Papers – Number of papers surveyed. Ling? – Is there support for linguistic features?

Reference	**QL**	**DB**	**Period**	**Papers**	**Ling?**
GA19 [20]	No	No	–	18	No
H16 [23]	Yes	Yes	2011–12	44	Yes
E19 [11]	Yes	Yes	2017–19	47	Partial

There is much more that could be done. For example, broad studies of deception should also include deceptive attacks such as phishing and social engineering attacks.

The meta-analysis of [23] searched four databases: PsycInfo, Social Science Citation Index, Dissertation Abstracts and Google Scholar for articles between 1945–February 2012 with "all permutations and combination of one keyword from three different clusters: (i) verb, language and linguistic; (ii) computer, artificial, software and automatic; (iii) lie, deceit, decept*."

The systematic review of [11] searched Google Scholar for articles between 2017–2019 using 10 queries listed in their paper. Their queries are a *proper* subset of the query

(fake ∨ false) news (identify ∨ detect) on (social media ∨ twitter)

which we repeated on Scholar on 11 November 2021, with a claim of 1,020,000 results.[10] Their queries produced a total claim of 157 potentially relevant results. We summarize the pertinent characteristics of the three latest reviews/surveys/meta-analysis of deception [11, 20, 23] in Table 2.

4.2 Our Analysis

Since the meta-analysis of [23] ended in February 2012, we searched the following databases: Google Scholar, PsycInfo and Dissertations and Abstracts Global, for the period 2013–21, with the boolean query

(verbal ∨ language ∨ linguistic ∨ text ∨ lexical) ∧ (computer ∨ artificial ∨ software ∨ automatic ∨ autonomous ∨ automated ∨ identify ∨ computational ∨ machine ∨ detect ∨ tool) ∧ (lie ∨ false ∨ fake ∨ deceit ∨ deception ∨ deceptive)

This query was formed by appropriately combining the queries from [11, 23], adding keywords after scanning the initial results, and after querying WordNet 3.1 with "deceit", "identify", and "lexical", and considering their synonyms for inclusion. Adding "recognition" to the middle clause reduced the set of results by more than 100K, a flaw of Google Search. Hence, other synonyms such as "discover", "recognize", and "recognizing" for "identify" and "fraud" for "deceit" were tried in separate queries, but their results on Scholar seemed irrelevant.

Scholar claimed over 1,150,000 results but only displayed the top 1,000 in relevance. A scan through this list identified 880 as potentially relevant matches. PsycInfo gave us 456 matches and the Dissertations database yielded 134 matches.

A new query was tried on Scholar without limiting the time period:

(verbal ∨ language ∨ linguistic ∨ text ∨ lexical) ∧ (computer ∨ artificial ∨ software ∨ automatic ∨ autonomous ∨ automated ∨ identify ∨ computational ∨ machine ∨ detect ∨ tool ∨ recognize ∨ recognition ∨ recognizing) ∧ (rumor ∨ hoax ∨ misinformation ∨ disinformation)

Scholar claimed 350,000 results and a scan of the top 1000 gave us 186 potentially relevant matches.

The results from Scholar were searched for feature selection and feature ranking papers. The 175 resulting papers included surveys, dataset and research papers. All surveys in the last five years were analyzed for insights on features. More than one survey mentioned n-grams of POS tags and semantic features [21, 45, 62] as examples of generalizable features. However, this analysis also revealed the lack of feature rankings for large, diverse and general datasets of deception.

[10] Google counts are loose upper bounds of actual matches.

5 Analysis of Datasets for Linguistic Cues

Next, we analyze datasets for domain-independent linguistic cues thereby tackling two challenges: (1) the ground truth problem for deception detection, and (2) evidence of linguistic cues for deception across domains.

A *ground truth* is something that is known to be correct, but this information is difficult to obtain, so we need models that do not rely on having too much ground truth data. Our approach is to focus on using linguistic information from the text. For the second challenge, we try to find universal linguistic markers for deception by looking for features that behave similarly across domains. We hope that an ML model built with these features could generalize across domains [18].

5.1 Sources for Linguistic Cues

Zhou [61] proposed 27 linguistic features for detecting deception and reported the values on her dataset.

Capuozzo et al. [6] collected a new dataset called DecOp using crowdsourcing. They also trained an SVM model that achieved 0.62–0.90 accuracy with different settings.

Siagian et al. [47] studied the effectiveness of using n-gram and function words (FW) as features. They found that word and character n-grams are effective features. For function words, although ML model with function words as features alone performed poorly, using function words with n-grams increased the robustness of the model.

Verma et al. [52] combined multiple fake news datasets into a large dataset called WELFake. They also designed an ML model with word embeddings and a voting classifier that achieved 96.56% F1 score, outperforming CNN and BERT.

These models, however, only work on their respective datasets, whereas we are interested in models that work across domains, even those domains that the model has never been trained on.

5.2 Datasets

We analyzed four labeled datasets, plus one set of calculated features, because the original dataset is unavailable. Zhou's dataset and DecOp contain deceptive opinions, WELFake dataset contains fake news, and the IWSPA-AP dataset contains phishing/legitimate emails. Thus, the goals in each dataset are quite different. Phishing email attackers wish to install malware or steal identity/money whereas deceptive opinion/review authors wish to sway opinions on services or products and fake news can have a variety of different goals such as swaying elections, dividing people, or causing chaos.

1. Zhou's dataset: This dataset is unavailable. Zhou et al. tested their features in their paper [61]. We use the results from this paper, which describes the features that are effective discriminators of truthful and deceptive texts in their corpus.

2. The DecOp Dataset: This dataset is from [6]. It contains truthful/deceptive opinions on various topics such as abortion and cannabis legalization.

3. The WELFake Dataset: This dataset is from [52]. It draws upon multiple true/fake news datasets.

4. The IWSPA-AP Dataset: The IWSPA Anti-Phishing competition dataset of emails [54].

5. The Amazon Reviews Dataset: This comprises real and fake Amazon reviews from a GitHub repository [42].

Dataset statistics are shown in Table 3.

5.3 Features: Linguistic Cues

These features are extracted from the texts of the sources. Features 1 to 27 are from [61], and features 28 to 30 are from [52]. Features with an asterisk, *, are selected after testing on the four available datasets.

Table 3: Statistics of the four available datasets covering different domains.

Dataset	Size	Truthful / Deceptive	Category
DecOp	1250	625 / 625	Deceptive Opinions
WELFake	72,134	35,028 / 37,106	Fake News
IWSPA-AP	5,723	5093 / 630	Phishing Emails
Amazon	20,974	10,481 / 10,493	Fake Reviews

	Zhou	Amazon Review	DecOp	IWSPA-AP	WELFake
feat1: word	+	-	-	-	+
feat2: verb	+	-	-	-	+
feat4: sentence	+	-	-	-	+
feat6: avg sen len	-	- *	-	-	-
feat7: avg word len	+ *	-	- *	-	-
feat9: pausality	-	-	-	-	-
feat10: modifiers	+	-	-	-	+
feat11: modal vebs	+	+*	+	+	-
feat12: certainty	-*	+	+	+	-
feat17: self ref	-	+	-	-	-
feat22: redundancy	-	+	-	-	+
feat29: SMOG readibility index	No Data	+	+	+	+
feat30: automatic readibility ind	No Data	-	+	-	-

Figure 2: Behavior of starred features. +: In deceptive texts, the feature has a higher value. −: In deceptive texts, the feature has a lower value. ∗: The difference is not statistically significant. Green: For the given feature, the majority +/− signs are colored green. Bold: The 10 features in bold are those still qualified after adjusting their p-values per the Holm-Bonferroni method.

List of 30 Cues Let document D denote a data instance, W(D) - the number of words in D, S(D) - the number of sentences in D and let Num- denote "number of". Using these notations the linguistic cues are presented in Table 4.

5.4 Results: selected features

Features with '*' in Table 4 are selected because, out of five datasets, they show consistent and significant difference between truthful and deceptive texts in at least three. A difference is significant if and only if its p-value, after being corrected using the Holm-Bonferroni method, is smaller than the threshold 0.01. Since there is some debate on the multiple comparisons issue (e.g., see [16]), we report statistically significant features both with and without the correction.

Next, we dig deeper into Figure 2, which shows their behaviors on the different datasets. A '+' means that the feature has a higher value in deceptive texts, and a '−' means the opposite. A '∗' next to the sign means that the difference is not statistically significant. The majority '+/−' sign of a feature is colored green, indicating its probable domain-independent behavior.

5.5 Discussion

Why do the features perform differently in different datasets? Why do Zhou et al.'s results tend to coordinate with WELFake's results, differing from the other three datasets? Looking at the different experimental

designs of Zhou and Capuozzo might explain the significant way in which the responses they collected were affected:

1. *Enthusiasm.* The plausible reason proposed in Zhou's paper for why deceptive texts have a higher quantity is that the task "solving the Dessert Survive Problem with your partner while deceiving him/her" is a more interesting task than just "solving the problem with your partner," so the deceivers are more engaged and put in effort to better deceive their partners. On the other hand, deceivers in DecOp experiment are not specifically deceiving anyone, instead they are just asked to write down an opinion contrary to their belief. So, there is not any sense of achievement of successfully deceiving someone as is the case of Zhou's experiment. Rather, deceivers in DecOp experiment might even be reluctant to write down these deceptive texts.

2. *Time.* Our finding that deceptive texts fall short on quantity is consistent with the prediction of Interpersonal Deception Theory (IDT), which states that deceivers "may exhibit reticence by using fewer words and sentences or less talk time than truth-tellers" [61]. In Zhou's experiment, "unlike interviews, ... deceivers in this investigation had ample opportunity to create and revise their messages so as to make them as persuasive as possible" [61]. So, deceivers in Zhou's experiment yielded more words probably because they have sufficient time (the experiment runs for 3 days), while DecOp's data are collected from AMT where participants were asked to write down their truthful/deceptive opinions to a prompt. The experiment in DecOp is similar to an interview setting, which is the setting of IDT. With time being not so efficient, deceivers in DecOp might have more difficulty in finding support for an opinion they do not believe in.

5.6 Features: Function Word n-gram

We are looking for n-grams of function words (FW) that appear significantly more or less in deceptive texts compared to truthful texts.

5.6.1 Method

Get Function Words Function words are words with no lexical meaning, such as 'when', 'at', 'the', in contrast to content words. There are two ways to identify them:

1. Use part of speech (POS) tagging and select words with specific POS. This website [32] shows us that words with certain POS are function words.

There are two popular POS tagging systems, Universal Dependencies (UPOS) [31] and Penn Treebank POS tags (XPOS) [41]. The tags that indicate function words, in both systems, are listed below:

UPOS: ["DET", "SCONJ", "CCONJ", "PRON", "AUX", "PUNCT"]

XPOS: ["IN", "MD", "WDT", "WP", "WP$", "WRB"]

2. Another approach is using a compiled list of function words. Linguists have made comprehensive lists of function words and we selected a list [7] used by Zeng in his phishing detection benchmark system, PhishBench 2.0 [59].

After testing, these two approaches produced similar results, and the second one is much faster without running neural network-based POS tagger. So, our experiment is conducted with the second method.

Select FW n-grams The steps of the method are depicted in Figure 3.

1. Split the dataset into truthful and deceptive texts.

2. Filter function words, discard all content words.

3. Extract the n-grams, with n ranging from 1 to 8. Larger n-grams are rare and are often not good discriminators between truthful and deceptive texts, our results show.

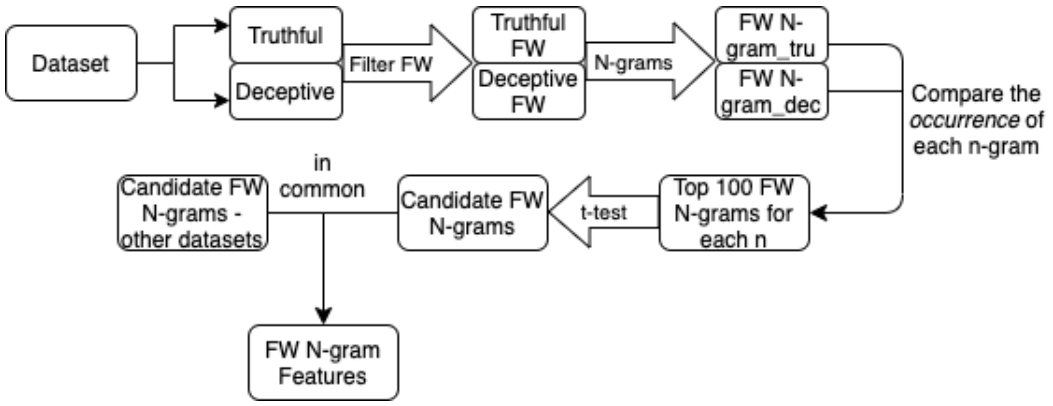

Figure 3: The selection of function-word n-grams.

4. Compute and compare the occurrence of each n-gram between the two groups. Select the top 100 FW n-grams for each n in terms of the difference of occurrence Occ:

$$Occ(n\text{-gram}) = \sum_{s:\text{a single response}}^{\text{all responses}} \frac{N(s)}{W(s) - n}$$

where $N(s)$ is the number of times the n-gram appears in data instance s and $W(s)$ denotes the number of words in s.

5. Run a t-test on these 100 FW n-grams for each n. Select those with a significant difference ($p < 0.01$). We also run a second version of the t-test where the p-values are adjusted per the Holm-Bonferroni method.

6. Select the common FW n-grams in the results from four datasets as the features.

5.6.2 Results

The results are given in Table 5.

1. There are only four n-grams that show significant difference between truthful/deceptive groups in all four datasets, viz. 'at', 'but', 'do', and 'that'. All four unigrams are less frequent in deceptive texts.

2. We relax our constraint a bit and select also n-grams that show significant difference between truthful and deceptive texts in at least three out of four datasets. These total 22, including 'me', 'they', 'do', 'is a', etc.

3. In the version where we use the adjusted p-values per the Holm-Bonferroni method to offset the effect of multiple comparisons, 12 n-grams are qualified.

5.6.3 Discussion

An individual dataset has thousands of FW n-grams, of which between 100 to 400 of them are significant discriminators of deceptive texts. Only 22 terms, 21 of them unigrams ($n = 1$), show a common behavior across at least three out of the four datasets, and only 4 across all four. With p-values adjusted by the Holm-Bonferroni method, only 12 are qualified. See Table 5. We also notice that the lower the n, the higher chance that the n-gram will show a significant difference between truthful and deceptive groups.

6 Conclusion

We have argued against hasty conclusions regarding linguistic cues for deception detection and especially their generalizability. We believe that the Critiques contained in [12,55,56] may present a valid point, namely that all linguistic cues might not generalize across the broad class of attacks. But sweeping statements such as those in the Critiques should be made with caution, since evidence suggests they are likely to be at least partly off the mark and may also discourage future research on the challenging topic.

We have given sound desiderata for systematic review and meta-analysis, which we hope will help computer science researchers, and machine learning/NLP researchers, conduct high quality analyses of the literature. Our cross-dataset analysis of five quite different deceptive datasets shows that there are linguistic features that can be used to build classifiers for more general deception datasets. We leave this direction for the future.

With all the new developments in machine learning and NLP, we believe that the existing research on linguistic deception detection is poised to take off soon and could result in significant advances.

References

[1] Ayman El Aassal, Luis Moraes, Shahryar Baki, Avisha Das, and Rakesh Verma. Anti-phishing pilot at ACM IWSPA 2018: Evaluating performance with new metrics for unbalanced datasets. In *Proc. of IWSPA-AP: Anti-Phishing Shared Task Pilot at the 4th ACM IWSPA*, pages 2–10, 2018.

[2] Mikel Aickin and Helen Gensler. Adjusting for multiple testing when reporting research results: The Bonferroni vs. Holm method. *American Journal of Public Health*, 86(5):726–728, 1996.

[3] L. Akehurst. *Deception and its Detection in Children and Adults via Verbal and Nonverbal Cues*. PhD thesis, Psychology Department, University of Portsmouth, 1997.

[4] Shahryar Baki and Rakesh Verma. Sixteen years of phishing user studies: What have we learned?, 2021. arXiv 2109.04661.

[5] Leticia C Cagnina and Paolo Rosso. Detecting deceptive opinions: intra and cross-domain classification using an efficient representation. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 25(Suppl. 2):151–174, 2017.

[6] Pasquale Capuozzo, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1423–1430, Marseille, France, May 2020. European Language Resources Association.

[7] Vivian Cook. `http://www.viviancook.uk/Words/StructureWordsList.htm`.

[8] Jeanna E. Cooper. *Using Natural Language Processing to Identify the Rhetoric of Deception in Business and Competitive Intelligence Email Communications*. PhD thesis, Robert Morris University, 2008.

[9] Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003.

[10] Nicholas D. Duran, Scott A. Crossley, Charles E. Hall, Philip M. McCarthy, and Danielle S. McNamara. Expanding a catalogue of deceptive linguistic features with NLP technologies. In H. Chad Lane and Hans W. Guesgen, editors, *Proceedings of the Twenty-Second International Florida Artificial Intelligence Research Society Conference, May 19-21, 2009, Sanibel Island, Florida, USA*. AAAI Press, 2009.

[11] Mohamed K. Elhadad, Kin Fun Li, and Fayez Gebali. Fake news detection on social media: A systematic survey. In *2019 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, pages 1–8. IEEE, 2019.

[12] Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. *Automatic Detection of Verbal Deception*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2015.

[13] Christie M. Fuller. *High-Stakes, Real-World Deception: An Examination of the Process of Deception and Deception Detection Using Linguistic-Based Cues*. PhD thesis, Oklahoma State University, 2008.

[14] Christie M. Fuller, David P. Biros, and Rick L. Wilson. Decision support for determining veracity via linguistic-based cues. *Decision Support Systems*, 46(3):695–703, 2009. Wireless in the Healthcare.

[15] Dariusz Galasinski. *The language of deception: A discourse analytical study*. Sage Publications, 2000.

[16] Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of research on educational effectiveness*, 5(2):189–211, 2012.

[17] Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova. Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. *arXiv preprint arXiv:2104.11761*, 2021.

[18] Stephanie Gokhman, Jeff Hancock, Poornima Prabhu, Myle Ott, and Claire Cardie. In search of a gold standard in studies of deception. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 23–30, Avignon, France, April 2012. Association for Computational Linguistics.

[19] Marcy Gordon. Tech firms struggle to police content while avoiding bias, 2019. `https://apnews.com/article/483752848e144227b53e3c38de5f5e7c`.

[20] Tommi Gröndahl and N Asokan. Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys (CSUR)*, 52(3):1–36, 2019.

[21] Bin Guo, Yasan Ding, Lina Yao, Yunji Liang, and Zhiwen Yu. The future of false information detection on social media: New perspectives and trends. *ACM Computing Surveys (CSUR)*, 53(4):1–36, 2020.

[22] Maria Hartwig and Charles F. Bond. Lie detection from multiple cues: A meta-analysis. *Applied Cognitive Psychology*, 28(5):661–676, 2014.

[23] Valerie Hauch. *Meta-analyses on the Detection of Deception with Linguistic and Verbal Content Cues*. PhD thesis, Justus-Liebig-Universität Gießen, 2016.

[24] Ángel Hernández-Castañeda, Hiram Calvo, Alexander F. Gelbukh, and Jorge J. García Flores. Cross-domain deception detection using support vector networks. *Soft Comput.*, 21(3):585–595, 2017.

[25] Sean L. Humpherys. *A System of Deception and Fraud Detection Using Reliable Linguistic Cues Including Hedging, Disfluencies, and Repeated Phrases*. PhD thesis, University of Arizona, 2010.

[26] Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5):1301–1326, 2021.

[27] Barbara Kitchenham and Stuart Charters. Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01, Keele University and Durham University, 2007.

[28] Sarah Ita Levitan, Angel Maredia, and Julia Hirschberg. Linguistic cues to deception and perceived deception in interview dialogues. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1941–1950, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[29] Maria D. Molina, S. Shyam Sundar, Thai Le, and Dongwon Lee. "Fake news" is not simply false information: A concept explication and taxonomy of online content. *American Behavioral Scientist*, 65(2):180–212, 2021.

[30] Casey Newton. What's good, bad, and missing in the Facebook whistleblower's testimony: What Frances Haugen gets right — and wrong. *The Verge*, 2021. Oct. 6, 2021.

[31] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[32] Richard Nordquist. What are function words in English grammar?, 2020. `https://www.thoughtco.com/function-word-grammar-1690876`.

[33] Valerie Perez. *Detecting Deception: Identifying Differences in Liars' and Truth Tellers' Verbal Strategies*. PhD thesis, Florida International University, 2010.

[34] Veronica Perez-Rosas. *Exploration of Visual, Acoustic, and Physiological Modalities to Complement Linguistic Representations for Sentiment Analysis*. PhD thesis, University of North Texas, 2014.

[35] Verónica Pérez-Rosas and Rada Mihalcea. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445, 2014.

[36] Isabel Picornell. *Cues to Deception in a Textual Narrative Context: Lying in Written Witness Statements*. PhD thesis, Aston University, Birmingham, UK, 2013.

[37] Proquest. Proquest dissertations & theses global, 2020. Searched October 3, 2020.

[38] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020.

[39] Ronald E. Riggio. *Verbal and Nonverbal Cues of Deception*. PhD thesis, University of California, Riverside, 1981.

[40] Rodrigo Rill-García, Luis Villaseñor Pineda, Verónica Reyes-Meza, and Hugo Jair Escalante. From text to speech: A multimodal cross-domain approach for deception detection. In *Pattern Recognition and Information Forensics - ICPR 2018 International Workshops, CVAUI, IWCF, and MIPPSNA, Beijing, China, August 20-24, 2018, Revised Selected Papers*, pages 164–177, 2018.

[41] Beatrice Santorini. Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania, 1990.

[42] Aayush Saxena. Deception-detection-on-Amazon-reviews-dataset, 2018. `https://github.com/aayush210789/Deception-Detection-on-Amazon-reviews-dataset`.

[43] Karen Schelleman-Offermans and Harald Merckelbach. Fantasy proneness as a confounder of verbal lie detection tools. *Journal of Investigative Psychology and Offender Profiling*, 7(3):247–260, 2010.

[44] Sadat Shahriar, Arjun Mukherjee, and Omprakash Gnawali. A domain-independent holistic approach to deception detection. In *Proceedings of RANLP*, 2021.

[45] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(3):1–42, 2019.

[46] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, September 2017.

[47] Al Hafiz Akbar Maulana Siagian and Masayoshi Aritsugi. Robustness of word and character n-gram combinations in detecting deceptive and truthful opinions. *J. Data and Information Quality*, 12(1), January 2020.

[48] Mark Staples and Mahmood Niazi. Experiences using systematic review guidelines. *Journal of Systems and Software*, 80(9):1425–1437, 2007.

[49] R Weylin Sternglanz, Wendy L. Morris, Marley Morrow, and Joshua Braverman. A review of meta-analyses about deception detection. In *The Palgrave Handbook of Deceptive Communication*, pages 303–326. Springer, 2019.

[50] Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[51] Anna Vartapetiance. *Computational Approaches for Verbal Deception Detection*. PhD thesis, Department of Computing, University of Surrey, 2015.

[52] Pawan Kumar Verma, Prateek Agrawal, Ivone Amorim, and Radu Prodan. Welfake: Word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4):881–893, 2021.

[53] Rakesh M. Verma and David J. Marchette. *Cybersecurity Analytics*. Chapman and Hall/CRC, 2019.

[54] Rakesh M. Verma, Victor Zeng, and Houtan Faridi. Data quality for security challenges: Case studies of phishing, malware and intrusion detection datasets. In Lorenzo Cavallaro, Johannes Kinder, XiaoFeng Wang, and Jonathan Katz, editors, *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11–15, 2019*, pages 2605–2607. ACM, 2019.

[55] Nikolai Vogler and Lisa Pearl. Using linguistically defined specific details to detect deception across domains. *Nat. Lang. Eng.*, 26(3):349–373, 2020.

[56] Aldert Vrij. *Detecting Lies and Deceit: Pitfalls and Opportunities*. John Wiley & Sons, 2008.

[57] Constantinos-Giovanni Xarhoulacos, Argiro Anagnostopoulou, George Stergiopoulos, and Dimitris Gritzalis. Misinformation vs. situational awareness: The art of deception and the need for cross-domain detection. *Sensors*, 21(16):5496, 2021.

[58] Min-Hsuan Yeh and Lun-Wei Ku. Lying through one's teeth: A study on verbal leakage cues. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4504–4510, 2021.

[59] Victor Zeng, Xin Zhou, Shahryar Baki, and Rakesh M. Verma. Phishbench 2.0: A versatile and extendable benchmarking framework for phishing. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, CCS '20, page 2077–2079, New York, NY, USA, 2020. Association for Computing Machinery.

[60] Fanghua Zheng. A corpus-based multidimensional analysis of linguistic features of truth and deception. In Stephen Politzer-Ahles, Yu-Yin Hsu, Chu-Ren Huang, and Yao Yao, editors, *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation, PACLIC 2018, Hong Kong, December 1-3, 2018*. Association for Computational Linguistics, 2018.

[61] Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker, and Doug Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13(1):81–106, January 2004.

[62] Xinyi Zhou and Reza Zafarani. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Computing Surveys (CSUR)*, 53(5):1–40, 2020.

Table 4: List of linguistic cues. Features 1–27 are from [61], and features 28–30 are from [52]. Features with an asterisk, $*$, are selected after testing on the four available datasets. Features in bold are still qualified after adjusting their p-values per the Holm-Bonferroni method.

1. **words***	W(D). NLTK's word tokenizer was used to identify words.	16. generalizing terms [skipped]	Missing computational description. The definition is: refers to a person (or object) as a class of persons or objects that includes the person (or object).
2. **verbs***	Num-verbs(D). NLTK's word tokenizer was used to identify verbs.	17. **self reference***	Num-first person singular pronouns(D) (i.e., Num-$\{I, me\}$/W(D).
3. noun phrase	Num-noun phrases(D). The noun chunk function in Spacy was used to identify noun phrases.	18. group reference	Num-first person plural pronoun(D) (i.e., Num-$\{we, us\}$)/W(D).
4. **sentence***	S(D). NLTK's sentence tokenizer was used to identify sentences.	19. emotiveness	(Num-adj.(D) + Num-adv.(D)) / (Num-nouns(D) + Num-verbs(D)).
5. average number of clauses	The average number of clauses per sentence. Stanza [38] was used to do POS tagging. Numclauses := Num-verb predicates (word.upos = "VERB") - Num-root (word.deprel = "root") - Num-conjugations (word.deprel = "conj").	20. lexical diversity	Num-distinct words / W(D).
6. **average sentence length***	W(D) / S(D)	21. content word diversity	Num-unique content words(D)/ Num-content words(D). Content words are words with lexical meanings, as opposite to function words. Methods to identify content/function words are discussed in the FW n-gram section.
7. **average word length***	Num-characters(D) / W(D). Characters include digits, punctuation, and spaces.	22. redundancy*	Num-function words(D) / S(D).
8. average length of noun phrase (NP)	Num-words in noun phrases(D) / Num-noun phrase(D). Noun phrases are identified in the same way as in (3).	23. typographical error ratio [skipped]	This feature is skipped because exploratory analysis showed that typographical error ratio is zero most texts from both categories. The popularity of auto-correct feature on browsers and text editing software has probably diminished the effectiveness of this feature.
9. **pausality***	Num-punctuation marks(D) / S(D)	24. spatio-temporal information	Num-('space' + 'time') / W(D), where 'space' and 'time' refer to the Num-words with the tag 'space' and 'time' in LIWC2015 dictionary.
10. **modifier***	Num-adjectives and adverbs(D). Stanza was used to do POS tagging. A word is an adjective or adverb iff word.upos = "ADJ" or word.upos = "ADV".	25. perceptual information	Num-'percep' / W(D). 'percep' is defined similarly, as in Feature (24).
11. modal verb*	Num-modal verbs(D) / W(D). Stanza was used to do POS tagging. A word is a modal verb iff word.upos = "AUX" and word.xpos = "MD".	26. positive affect	Num-'posemo' / W(D). 'posemo' is defined similarly, as in Feature (24).
12. certainty*	Num-words that have the tag 'certain' in LIWC2015 dictionary(D) / W(D). LIWC is a dictionary that associates words with various tags [50]. We used a commercial program developed by Pennebaker Conglomerates, Inc.	27. negative affect	Num-'negemo' / W(D). 'negemo' is defined similarly in Feature (24).
13. other reference	Num-third person pronoun(D) / W(D). Used Stanza to do POS tagging. A word is a third person pronoun iff word.xpos = "PRP" and word.feats = "Person=3".	28. Gunning fog grade readability index	An index to quantify the readability of a text by estimating the years of education required to understand the text. We used Textstat to calculate this index.
14. passive voice	Num-passive voice verb(D) / W(D). Used Stanza to do POS tagging. A word is a passive voice verb iff word.deprel = "aux:pass".	29. **SMOG readability index***	Another index trying to estimate the years of education required to understand the text. We used Textstat to calculate this.
15. objectification [skipped]	Missing computational description [61]. It is defined as an expression given (as an abstract notion, feeling, or ideal) in a form that can be experienced by others and externalizes one's attitude.	30. **automatic readability index***	Similar to (28) and (29), we also used Textstat for this.

Table 5: Table of n-grams with statistics and Gini coefficients for each dataset, sorted by their harmonic mean p-values (HMP). The t-statistic for each dataset quantifies the difference of *Occ* in deceptive and truthful texts. A positive statistic means the feature's *Occ* value is higher in deceptive texts. A '–' in a statistic column means that the p-value of this n-gram in the given dataset is higher than the threshold (0.01), so we have excluded its statistic. The 12 n-grams in bold are those still qualified after adjusting their p-values per the Holm-Bonferroni method.

	n-gram	HMP	statistics_amazon	statistics_decop	statistics_iwspa	statistics_welfake	gini_amazon	gini_decop	gini_iwspa	gini_welfake
0	**me**	1.03×10^{-19}	8.144	−4.859	−9.904	−8.211	0.653	0.559	0.792	0.535
1	**they**	1.04×10^{-234}	−4.095	–	−19.015	−33.074	0.901	0.762	0.932	0.745
2	do	1.27×10^{-110}	−3.594	−3.278	−3.253	−22.724	0.788	0.689	0.729	0.643
3	**only**	1.29×10^{-122}	−5.238	–	−5.047	−23.903	0.926	0.924	0.933	0.819
4	**will**	1.35×10^{-36}	–	7.306	8.312	13.164	0.907	0.910	0.870	0.746
5	many	1.44×10^{-15}	–	−3.297	−8.731	−8.674	0.968	0.930	0.960	0.836
6	**that**	1.68×10^{-53}	−2.659	−6.684	−11.302	−15.869	0.784	0.658	0.868	0.584
7	then	3.16×10^{-153}	−3.650	–	−3.438	−26.716	0.965	0.949	0.953	0.859
8	up	4.84×10^{-51}	−3.188	–	−3.092	−15.491	0.849	0.880	0.824	0.637
9	**but**	4.88×10^{-18}	−8.443	−7.955	−9.515	−2.943	0.782	0.883	0.887	0.699
10	their	5.03×10^{-87}	−2.886	–	−7.468	−20.156	0.970	0.825	0.940	0.744
11	is a	5.28×10^{-170}	5.392	4.617	3.148	−28.157	0.864	0.804	0.891	0.735
12	**there**	5.33×10^{-44}	–	−5.235	−8.706	−14.403	0.949	0.893	0.925	0.802
13	both	5.62×10^{-50}	−5.274	−3.063	−15.374	12.680	0.977	0.982	0.962	0.884
14	**I**	7.10×10^{-302}	11.651	−17.688	−36.593	−37.583	0.563	0.571	0.780	0.625
15	out	7.98×10^{-191}	−3.567	–	−7.779	−29.814	0.815	0.848	0.862	0.621
16	**them**	8.01×10^{-81}	−6.173	3.831	−7.453	−19.447	0.918	0.872	0.937	0.810
17	**at**	8.05×10^{-72}	−2.674	−4.533	−17.515	−18.343	0.586	0.475	0.729	0.508
18	about	9.49×10^{-104}	−2.888	–	−5.497	−21.996	0.925	0.947	0.910	0.729
19	**and**	9.52×10^{-19}	4.495	5.049	−4.613	9.570	0.530	0.507	0.808	0.544
20	**if**	9.95×10^{-68}	−4.349	–	−6.590	−17.804	0.830	0.734	0.836	0.672
21	than	9.96×10^{-13}	−4.008	−2.803	−8.057	–	0.913	0.876	0.927	0.774