Domain-Independent Deception: A New Taxonomy and Linguistic Analysis*

Rakesh Verma
Department of Computer Science, University of Houston

Nachum Dershowitz
School of Computer Science, Tel Aviv University

Victor Zeng
University of Houston and InstaBase

Dainis Boumber
University of Houston

Xuting Liu†
University of California – Berkeley

September 6, 2023

Abstract

Internet-based economies and societies are drowning in deceptive attacks. These attacks take many forms, such as fake news, phishing, and job scams, which we call "domains of deception." Machine-learning and natural-language-processing 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. First, we provide a new computational definition of deception and break down deception into a new taxonomy. Then, we analyze the debate on linguistic cues for deception and supply guidelines for systematic reviews. Finally, we investigate common linguistic features and give evidence for knowledge transfer across different forms of deception.

Keywords: Automatic/computational deception detection, cross domain, domain independent, email/message scams, fake news, meta-analysis, opinion spam, phishing, social engineering attacks, systematic review, text analysis.

1 Introduction

History is replete with famous lies and deceptions. Examples include: P. T. Barnum, Nicolo Machiavelli, Sun Tzu, Operation Mincemeat, and the Trojan Horse Levine [2014]. A chronology of deception is included in Levine [2014]. More recently, the proliferation of deceptive attacks such as fake news, phishing, and disinformation is rapidly eroding trust in Internet-dependent societies. The situation has deteriorated so much that 45% of the US population believes the 2020 US election was stolen.[1]

Social-media platforms have come under severe scrutiny regarding how they police content. Facebook and Google are partnering with independent fact-checking organizations that typically employ manual fact-checkers.

*This is a thoroughly revised version of a 2022 arXiv draft containing substantial new material.
†Work performed at the University of Houston and U. C. Berkeley.
[1] `https://www.surveymonkey.com/curiosity/axios-january-6-revisited`.

Natural-language processing (NLP) and machine learning (ML) researchers have joined the fight by designing fake news, phishing, and other kinds of domain-specific detectors.

Building single-domain detectors may be sub-optimal. Composing them sequentially requires more time, and composing them in parallel requires more hardware. Moreover, building single-domain detectors means one can only react to new forms of deception after they emerge.

Our goal here is to spur research on *domain-independent* deception. Unfortunately, research in this area is currently hampered by the lack of computational definitions and taxonomy, high-quality datasets, and systematic approaches to domain-independent deception detection. Thus, results are neither generalizable nor reliable, leading to much confusion.

Accordingly, we make the following contributions:

- We propose a new computational definition and a new comprehensive taxonomy of deception. (We use the unqualified term "deception" for the domain-independent case. When the goals of the deception are unclear, we refer to "lies.")
- We examine the debate on linguistic deception detection, identify works that demonstrate the challenges that must be overcome to develop domain-independent deception detectors and examine them critically.
- We conduct linguistic analysis of several detection datasets for general cues and find several statistically significant ones.
- We conduct deep learning experiments of deception sets and study correlations in performance for pairs of datasets.

We hope that this article, besides scrutinizing the claims on general linguistic signals for deception, will aid those planning to conduct systematic reviews. Google Scholar searches with phrase queries of the two forms: (a) "guidelines for systematic literature reviews in X" and (b) "systematic review guidelines for X," where $X \in$ {machine learning, ML, natural language processing, NLP, nlp} returned nothing.

This article is organized as follows: Section 2 presents a new definition of deception. Section 3 introduces our new taxonomy. Section 4 summarizes related work. Section 5 presents guidelines for systematic reviews and Sections 6, 7 the linguistic cues debate. Sections 8 and 9 describe our experiments, results, and analysis of domain-independent markers for deception. Finally, Section 10 presents some conclusions and directions for the future. The appendices provide the list of features tested and some preliminary significance testing of cues on four public deception datasets.

2 Definition

We first examine a general definition of deception, taken from Galasinski [2000], intended to capture a wide variety of deceptive situations and attacks.

Definition 1 (Preliminary) ***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 achieve the goals.*

Since we are interested in automatic verifiability, we would like to modify this definition of deception and propose one that is computationally feasible. Because intentions are notoriously hard to establish, we will use the effect of exposing the manipulation/goals instead.

Our revised definition is the following:

Definition 2 (Deception) ***Deception*** *is an act of manipulation designed to gain compliance such that, exposing the manipulation or the goal(s) of compliance significantly decreases the chance of compliance. 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 goals.*

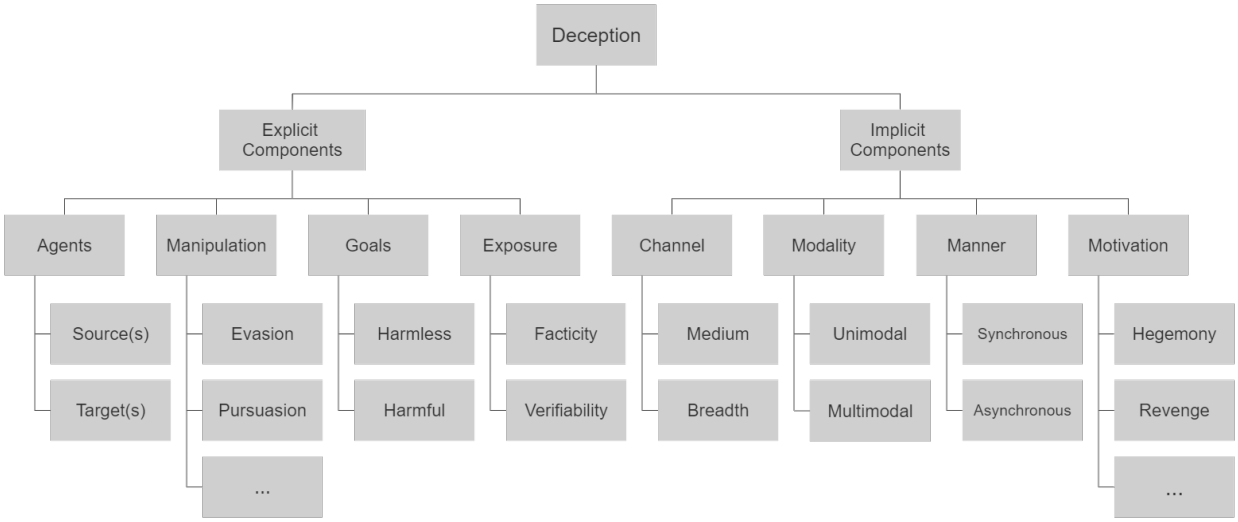


Figure 1: The proposed deception taxonomy – the full manipulation (or stratagem) and motivation subtrees are not shown.

One might argue that the goals of deception should be harmful to an individual or 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.

To formalize our definition, we borrow from the language of Markov decision processes. Let A be an action taken by an actor, and let C be a desired compliance state. We use $K(A, T)$ to denote the action A plus the full and truthful explanation of the actor's *relevant* private information to target T. We formalize (computational) deception using conditional probabilities as follows:

Definition 3 (Computational Deception – Formalized) *An action A deceives target T if $P(C \mid K(A,T)) < P(C \mid A)$.*

Moreover, we can quantify the degree to which A is deceptive by the amount θ, where $0 \leq \theta \leq 1$.

Definition 4 (Computational Deception – Quantified) *An action A θ-deceives T if $P(C \mid K(A,T)) \leq P(C \mid A) - \theta$.*

In practice, practitioners can apply this by exposing the manipulation and/or goals and measuring the change in compliance rates. For example, a Florida woman recently sued Kraft alleging that the "ready in $3\frac{1}{2}$ minutes" on the label of their microwavable Velveeta Shells & Cheese is deceptive. To determine whether the claim is in fact deceptive, a researcher could present the product by itself to one group of random consumers and the product with an explanation that the $3\frac{1}{2}$ minutes does not include the time to add water to another group. If there is a statistically significant decrease in purchases (which is the desired compliance) for the group with the explanation, then the claim is deceptive.

There is some work on finding out how good humans are at detecting certain kinds of deceptive attacks. For the detection capabilities of automatic detectors on specific domains of deception, one can look at surveys on fake news detection Sharma et al. [2019], Zhou et al. [2019, 2020] and phishing detection Das et al. [2020].

3 Taxonomy and Examples

There have been a few attempts at constructing taxonomies for fake news, phishing, or other forms of deception.

Molina et al. [2021] give a taxonomy of *fake news* with four dimensions: message and linguistic, sources and intentions, structural, and network. Kapantai et al. [2021] conducted a systematic search for papers proposing taxonomies for disinformation and synthesized a taxonomy with three dimensions: factuality, motivation, and verifiability.

No one, to our knowledge, has given a comprehensive taxonomy of real-world deception.

3.1 The New Taxonomy

We put forward a multi-dimensional taxonomy. Under our definition, deception explicitly involves four elements: (1) agents: the sources, and the targets, (2) stratagems for manipulation, (3) goals, and (4) threat/mechanisms of exposure. These explicit elements can be further broken down as follows:

1) *Agents.* Rowe [2006] calls this category "participant," and he further elaborates this into: (a) agent, who initiates the action, (b) beneficiary, who benefits, (c) object, what the action is done to, and (d) recipient, who receives the action. Rowe also includes experiencer ("who senses the action") and instrument ("what helps accomplish the action") components in this category, but we include them in the Channel category below.

1a) Sources. This includes human (individual or group), bot, etc., or mixed, in other words, combinations such as a human assisted by a bot. The Sources category includes initiators and beneficiaries.

1b) Targets. This includes humans (individual or group), automatic detectors, or both. For example, spam targets automatic detectors, and phishing targets humans, but needs to fool automatic detectors also. The Targets category includes the objects and the recipients.

2) *Stratagems.* The stratagem subtree in the taxonomy includes two sub-taxonomies for persuasion and action, which we discuss below. We believe that persuasion is fundamental to deception since its goal is to change the reasoning of the target(s), with the deception's end goal of compliance. The action taxonomy is adapted from Rowe [2006]. It includes space, time, causality, quality, essence, and speech-act theory, which specifies the external and internal preconditions for the action. The persuasion taxonomy combines Cialdini [2006] and Da San Martino et al. [2023].

3) *Goals.*

3a) Harmless: satire, parody, satisfying participation, as in a laboratory experiment where participants may be asked to lie, etc.

3b) Harmful. This includes a wide range of objectives, such as stealing money or identity information, malware installation, manipulation of votes, 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, winning competitions/games, etc. We avoid the terms defensive and offensive since they are dependent on the perspective of the participants/agents.

4) *Exposure.*

4a) 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."

4b) Verifiability. Assuming facticity, how easy or difficult it is to verify whether it is legitimate or deceptive? Here, we are interested in machine or automatic verification. If a simple machine-learning algorithm can detect it with high recall and precision, we will deem it easy.

In addition, there are four implicit concepts in the definition: (1) motivations behind the goals, (2) communication channels or media, (3) modality of deception, and (4) manner or timeliness of the exchange.

1) *Motivation.* This is the rationale for the goals. The agents involved and their characteristics reveal the underlying motivations, which could be political hegemony (nation-states), religious domination, revenge (disgruntled employee), ideological gains, money, control, power, etc.

2) *Channel.* This dimension includes two aspects:

2a) Breadth: Whether the targets are a few specific individuals or detector types or broad classes of people/categories of detectors.

2b) Media. How the deceptive capsule is conveyed to the target. Media also includes the experiencer and instrument components of Rowe [2006].

3) *Modality.* This dimension refers to the presentation of deceptive content. It includes:

3a) Unimodal. This includes only one type of modality such as (a) Gestural: body language is used to deceive, (b) Audio (a.k.a. verbal), (c) Textual (e.g., SMS/email), and (d) Visual (e.g., images or videos).

3b) Multimodal: combinations of different modalities.
For example, audio-visual has both speech and visual components but lacks face-to-face communication in which gestures could facilitate deception.

4) *Manner/Timeliness.*

4a) Interactive/Synchronous. A real-time interview or debate is an interactive scenario.

4b) Non-interactive/Asynchronous. An Amazon Mechanical Turker typing a deceptive opinion or essay is a non-interactive one. An asynchronous interaction can have multiple stages or steps some (but not all) of which may be synchronous.

3.1.1 Stratagems

Rowe's approach Rowe [2006] is based on linguistics. He states, "Each action has associated concepts that help particularize it, and these are conveyed in language by modifiers, prepositional phrases, participial phrases, relative clauses, infinitives, and other constructs." These associated concepts are called 'semantic cases' Fillmore [1968] in analogy to the syntactic cases that occur in some languages for nouns. Rowe claims that "every deception action can be categorized by an associated semantic case or set of cases." However, there is no canonical list of semantic cases in linguistics. Rowe prefers the detailed list from (Copeck et al, 1992), which he supplements with two important relationships from artificial intelligence, the upward type-supertype and upward part-whole links, and two speech-act conditions from (Austin, 1975), to get 32 cases altogether. However, since we include his participant category in the Agents and Channel categories, we have only 26 subcategories in the Stratagems category.

1. Space, which consists of: (a) direction, of the action, (b) location-at, where something occurred, (c) location-from, where something started, (d) location-to, where something finished, (e) location-through, where some action passed through, and (f) orientation, in some space.

2. Time, which is subdivided into: (a) frequency, of occurrence of repeated action, (b) time-at, time at which something occurred, (c) time-from, the time at which something started, (d) time-to, the time at which something ended, and (e) time-through, the time through which something occurred.

3. Causality, which consists of: (a) cause, (b) contradiction, what this action opposes if anything, (c) effect, and (d) purpose.

4. Quality, which is sub-divided into: (a) accompaniment, an additional object associated with the action, (b) content, what is contained by the action object, (c) manner, the way in which the action is done, (d) material, the atomic units out of which the action is composed, (e) measure, the measurement associated with the action, (f) order, with respect to other actions, and (g) value, the data transmitted by the action (the software sense of the term).

Table 1: The Persuasion Taxonomy, adapted from Da San Martino et al. [2023], is a sub-taxonomy in the Deception Taxonomy.

Category	Description
Justification	An argument made of two parts: a statement and a justification
Simplification	A statement is made that excessively simplifies a problem, usually regarding the cause, the consequence or the existence of choices
Distraction	A statement is made that changes the focus away from the main topic or argument
Call	The text is not an argument but an encouragement to act or think in a particular way
Manipulative Wording/Images	Specific language/imagery is used or a statement is made that is not an argument, and which contains words/phrases that are either non-neutral, confusing, exaggerating, etc., to impact the reader, for instance emotionally
Attack on Reputation	An argument whose object is not the topic of the conversation, but the personality of a participant, his experience and deeds, typically to question and/or undermine his credibility

5. Essence, which consists of: (a) supertype, a generalization of the action type, and (b) whole, of which the action is a part.

6. Speech-act theory, which is sub-divided into: (a) an external precondition on the action, and (b) an internal precondition, on the ability of the agent to perform the action.

3.1.2 Persuasion

We summarize the persuasion taxonomy in Table 1. For this taxonomy, we adapt the SemEval 2023 Persuasion Task's categories Da San Martino et al. [2023], and Cialdini's [2006] persuasion principles, which are essentially persuasion techniques or strategies. The persuasion strategies taxonomy of Guerini et al. [2007] is orthogonal to this taxonomy since their definition of persuasion is broader than ours, but we do include their specific strategies under Techniques. The techniques used for each category are as follows (30 in total):

- Justification: Appeal to popularity, Appeal to authority/expert, Appeal to values (or Commitment Cialdini [2006]), Appeal to fear/prejudice, Reciprocity Cialdini [2006] (or Goal Balance Guerini et al. [2007]), Scarcity Cialdini [2006], Reward, Appeal to relevant empirical evidence, Relevant Statistics, and Relevant Examples.

- Simplification: Causal oversimplification, False dilemma or no choice, and Consequential oversimplification.

- Distraction: Straw man, Red herring (includes irrelevant empirical evidence, statistics or examples), Whataboutism, Flag Waving, and Liking Cialdini [2006].

- Call: Slogans, Social Proof Cialdini [2006], Appeal to time, and Conversation killer.

- Manipulative Wording/Images: Loaded language/images, Repetition, Exaggeration or minimization, and Obfuscation – vagueness or confusion.

- Attack on reputation: Name calling or labeling, Doubt, Guilt by association, Appeal to hypocrisy, Questioning the reputation.

To the best of our knowledge, we are the first to use the following dimensions in a taxonomy of deception: target, persuasion, goal, dissemination, and timeliness. We add these to give a comprehensive view of

deception, to aid in domain-independent deception detection, and to clarify and classify deception in all its different manifestations. Such a comprehensive taxonomy will provide a solid foundation on which to build automatic and semi-automatic detection methods and training programs for the targets of deception.

3.2 Examples

To demonstrate the applicability of this taxonomy, we give three examples. More discussion of stratagems and examples of cyber deception can be found in Rowe [2006].

Phishing is when attackers pretend to be from reputable companies to trick victims into revealing personal information. The agents are the attackers as initiators and the targets are the Internet/email users. The harmful goals include information or malware installation. Establishing the facticity is difficult if the attacker is determined. The medium is the Internet/email. The breadth is high for phishing and narrower for spearphishing. The modality is text for phishing and audio for vishing. Images may also be used in phishing emails. The manner is non-interactive for phishing and interactive for vishing. Deliberate falsification and persuasion techniques such as authority, social proof, and reward or loss claims are employed in the stratagem.

Fake news is manufactured and misleading information presented as news. Here the harmful goals include swaying opinion, sowing unrest, and division, etc. The sources could be individuals, organizations, or nation-states. The breadth could vary depending on how deep-pocketed and determined the source(s) is (are). The modality could be text, audio, images, or video. The manner is asynchronous. Fake news could employ a range of techniques in the action component of the stratagem: from deliberate falsification to evasion and the persuasion component could include authority, social proof, etc.

Fake reviews are reviews designed to give consumers a false impression of a product or business. The harmful goal is to convince consumers to buy their product or avoid a competitor's. The sources could be humans, bots, or their combinations. The targets are potential customers as well as the platform's fake review detector. The breadth is thus a broad range of people. While most fake reviews use only texts, deliberate attacks could be multi-modal, adding visuals and/or audio. Falsification and social proof are the main stratagems. Facticity and verifiability could vary depending on the stratagems used. The manner is asynchronous.

4 Related Work

Deception has a vast social science literature. Hence, we focus on the most closely related work on computational deception, which can be categorized into: taxonomies, datasets, detection, and literature reviews. Of the latter, we focus here on reviews of linguistic deception detection. The DBLP query "domain decepti" on 30 August 2023 gave 30 matches of which 18 were deemed relevant.

Remark Unfortunately, previous researchers have generally left the term "domain" undefined. In Glenski et al. [2020], different social networks, such as Twitter and Reddit, are referred to as domains. Hence, terms such as "cross-domain deception" in previous work could mean that the topics of essays or reviews are varied whereas the goals could stay pretty much the same.

4.1 Taxonomies

Whaley and Aykroyd [2007] gave a taxonomy of perception in which deception was defined succinctly as "other-induced misperception." The full definition given in Whaley and Aykroyd [2007] is "any attempt – by words or actions – intended to distort another person's or group's perception of reality." In Bell and Whaley [2017] two groups were introduced as essential for deception: simulation (overt, showing the false) and dissimulation (covert, hiding what is real). They introduced three simulation techniques: mimicking, inventing, and decoying, and three dissimulation techniques: masking, repackaging, and dazzling.

Dunnigan and Nofi [2001] gave a taxonomy of deception in the military context. This included: concealment, camouflage, disinformation, lies, displays, ruses, demonstrations, feints, and insight.

The most comprehensive previous taxonomy of deception, to our knowledge, is proposed in Rowe [2006]. It is inspired by linguistic case theory and includes 32 cases which are grouped into seven categories: space (six cases), time (five cases), participant (six cases), causality (four cases), quality (seven cases), essence (two cases), speech-act theory (two cases). Analyzing this taxonomy, we find that, except for the participant category, all the other categories fit neatly into the stratagems class for deception in our taxonomy.

More recently, a few researchers have proposed more specialized taxonomies for what they call defensive deception Oluoha et al. [2021], Pawlick et al. [2019], Pawlick and Zhu [2021]. Some folksy and psychological taxonomies are given in Druckman and Bjork [1992].

4.2 Datasets

Several datasets have been collected for studying lies. However, researchers have not carefully delineated the scope by considering the goals of the deception. There is also another potentially more serious issue: some datasets are constructed by asking participants to lie in a laboratory setting, where there are no consequences and no incentive to lie. We will refer to them as *Lab Datasets*. Others are constructed by collecting samples of real attacks. We call them *Real-World Datasets*. Finally, there are some datasets in which data from laboratory settings are combined with real-world attack samples. We call them *Mixed Datasets*.

Lab Datasets include Zhou et al. [2004], wherein students were paired and one student in each pair was asked to deceive the other using messages. In Perez-Rosas [2014], researchers collected demographic data and 14 short essays (7 truthful and 7 false) on open-ended topics from 512 Amazon Mechanical Turkers (AMT). They tried to predict demographic information and facticity. We refer to this as the *Open-Lies* dataset. In Pérez-Rosas and Mihalcea [2014], researchers collected short essays on three topics: abortion, best friend, and the death penalty by people from four different cultural backgrounds. In Capuozzo et al. [2020], truthful and deceptive opinions on five topics are collected in two languages (English and Italian). See Ludwig et al. [2016] for more such efforts.

Next, we consider real-world datasets, where the goals may be information, disruption, financial or psychological. Here we have several datasets for fake news detection Raponi et al. [2022],[2] opinion spam (aka fake reviews) detection Ren and Ji [2019], for phishing Verma et al. [2019], and a company's reward program Ludwig et al. [2016].

Some researchers have mixed data obtained from laboratory settings with non-laboratory data, such as reviews obtained from forums. For example, in Hernández-Castañeda et al. [2017], researchers analyzed three datasets: a two-class, balanced-ratio dataset of 236 Amazon reviews, a hotel opinion spam dataset consisting of 400 fabricated opinions from AMT plus 400 reviews from TripAdvisor (likely to be truthful), and 200 essays from Pérez-Rosas and Mihalcea [2014]. In Xarhoulacos et al. [2021], researchers studied a masking technique on two datasets: a hotel, restaurant, and doctor opinion spam dataset and the dataset from Pérez-Rosas and Mihalcea [2014]. In Cagnina and Rosso [2017], in-domain experiments were done with a positive and negative hotel opinion spam dataset, and cross-domain experiments were conducted with the hotel, restaurant, and doctor opinion spam dataset.

A few works have developed domain-independent deception datasets in our sense, wherein the goals of deception can be quite different. In Rill-García et al. [2018], researchers used two datasets: the American English subset consisting of a balanced-ratio 600 essays and transcriptions of 121 trial videos (60 truthful and 61 deceptive), which we call Real-Life_Trial. In Vogler and Pearl [2020], three datasets were used: positive and negative hotel reviews, essays on emotionally-charged topics, and personal interview questions. In Xarhoulacos et al. [2021], multiple fake news datasets, a COVID-19 dataset, and some micro-blogging datasets were collected and analyzed. In Shahriar et al. [2021], researchers collected fake news, Twitter rumors, and spam datasets. (Spam is essentially advertising. Deception is employed to fool automatic detectors rather than the human recipient of the spam. We focus on human targets.) They applied their

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

models trained on these datasets to a new COVID-19 dataset. In Yeh and Ku [2021], seven datasets were collected (Diplomacy, Mafiascum, Open-Domain, LIAR, Box of Lies, MU3D, and Real-Life_Trial) and analyzed using LIWC categories, without claiming domain independence or cross-domain analysis. However, their datasets do involve different goals. LIAR, for instance, 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. In Feng et al. [2012], four datasets were collected: trip-advisor gold, a balanced hotel reviews dataset of 800 reviews introduced in Ott et al. [2011], trip-advisor heuristic, another balanced reviews dataset of 800 reviews collected by the authors, a third 800 review Yelp dataset of uncertain ground-truth collected by the authors, and the 296 essays on three topics dataset of Mihalcea and Strapparava [2009]. They show that features based on CFG parse trees along with unigrams performed the best on these datasets.

Thus, we still lack large, comprehensive datasets for deception that have a wide variety of deceptive goals.

4.3 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. On phishing alone (query: phish), there are more than 1,700 DBLP results, including over 15 surveys and reviews. Similarly, there are over 900 papers on scams (query: scam, not all of them are relevant, since many occurrences are part of acronyms such as SCAMP), over 100 on opinion spam, close to 100 on fake reviews, and over 1,800 on fake news.[3]

A soft domain transfer method is proposed in Shahriar et al. [2022]. They found that partial training on tweets helped in phishing and fake news detection. In Panda [2022], Panda and Levitan [2023], the authors study deception detection across languages and modalities. Other works on domain-independent deception detection have been discussed above under Datasets.

4.4 Reviews on Linguistic Markers

Recently, Gröndahl and Asokan [2019] conducted a survey of the literature on deception. They defined implicit and explicit deception, focused on automatic deception detection using input texts, and then proceeded to review 17 papers on *linguistic* deception detection techniques. (Explicit deception is when the deceiver explicitly mentions the false proposition in the deceptive communication.) 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 Vogler and Pearl [2020] assert that extensive psychology research shows that "a generalized linguistic cue to deception is unlikely to exist."

We collectively refer to Gröndahl and Asokan [2019], Fitzpatrick et al. [2015], and Vogler and Pearl [2020], Vrij [2008] as the *Critiques*. We argue that, at best, their analyses and conclusion may be a bit too hasty and elaborate on several aspects that need 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.

5 Guidelines for Systematic Reviews

According to Staples and Niazi [2007], "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." Unlike an ad-hoc literature review, systematic reviews are formally planned and methodically executed. Such a review can reveal the structure and patterns of existing research, highlight key results, and identify gaps for future research.

[3]All these DBLP search results are as of 31 August 2023.

Unsurprisingly, there has been an explosion of systematic reviews on all kinds of problems in natural language processing and machine learning. However, there is a dearth of good guidelines and procedures for them in NLP. In this section, we lay out guidelines for a good systematic review and identify several common pitfalls a team can stumble into.

5.1 What Makes a Good Systematic Review?

A good systematic review is independently replicable and thus 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 require much more effort than ordinary literature surveys.

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

(1) A predefined and documented protocol specifying the research question and procedures to be used in performing the review.

(2) A defined and documented search strategy designed to find *as much of the relevant literature as possible.*

(3) Explicitly predefined criteria for determining whether to include or exclude a candidate study.

(4) Description of quality assessment mechanisms to evaluate each study.

(5) Description of review and cross-checking processes involving multiple independent researchers, to control researcher bias.

5.2 Common Pitfalls

We identify several challenges faced by reviews and surveys, systematic or conventional.

Inadequate search strategies Not having a clear, explicit search strategy for literature or clearly defined inclusion and exclusion criteria can lead to bias in the selection of papers.

Confirmation nias The search strategy should be designed to avoid favoring one hypothesis over another.

Publication bias Even a rigorous and thorough search of the published literature might not give a full picture of the state of the field. Factors besides the quality of the work can influence whether a paper gets published. For example, studies with positive results, papers with well-written English, and papers authored by highly reputed researchers are more likely to get published. Longer works such as theses are also missed in the emphasis on published literature.

Clique bias Cliques of interconnected researchers and papers analyzing the same dataset may share biases. When performing a systematic review, researchers must ensure that they are not lending too much weight to one cluster of connected works.

Quality of studies and datasets Within the literature, there is a significant range in the quality of the studies. Quality assessment criteria should consider: (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 correction was used for the multiple comparisons issue, and (v) their replicability.

6 A Critique of the Critiques

We now take our guidelines and apply them to the critiques.

The deception survey of Gröndahl and Asokan [2019] has some of the features of a good systematic review: they specify the research questions and hypotheses and involve two researchers (presumably mentor and mentee). However, they lack a formal review protocol, search strategy, or explicit inclusion/exclusion criteria, and no quality assessment mechanism is specified. However, because it is published in an influential journal, it is likely to leave a lasting impression on deception researchers, so it is worth the time and effort to examine its strengths and weaknesses.

To check the completeness of their search, we searched the literature for relevant papers published before 2019. The deception survey Gröndahl and Asokan [2019] suffers from an incomplete search. Although their goal was to survey automatic linguistic deception detection literature, they missed many relevant papers including the meta-analysis by Hauch [2016].

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."

To check for publication bias, we performed a systematic search of the ProQuest Global Database. We identified 118 dissertations and theses with the keywords "deception" *and* "detection" in the title. Three of these also had the word "linguistic" in the title and all three were relevant. Replacing "linguistic" with "natural language processing" (or "textual") and keeping "deception" yielded two more relevant dissertations. Finally, "verbal" with "deception" yielded four more relevant results, out of nine total. None of the above dissertations are cited in the Critiques.

To check for clique bias in the deception survey Gröndahl and Asokan [2019], we listed all the authors of the 17 papers cited in Section 2 ("Deception Detection Via Text Analysis") of their paper and generated a graph of connected papers in Figure 2. We consider two papers connected if they share a common author. We find that nine authors account for 15 (88%) of the papers, and the connected papers graph contains several cliques, two as large as K_4 and one additional K_3.

None of the papers examined in the meta-analysis conducted by DePaulo et al. [2003] and the deception survey in Gröndahl and Asokan [2019] built a general dataset for different deception goals (e.g., as in phishing, fake news, *and* crime reports). If 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.

Datedness The meta-analysis of DePaulo et al. [2003] was conducted in 2003. The meta-analysis of Hauch [2016] is more recent, but still only covers papers up to February 2012. The latest review of meta-analyses Sternglanz et al. [2019] 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 contemporary inadequacy of the literature cited in the Critiques.

7 Domain-Independent Markers

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

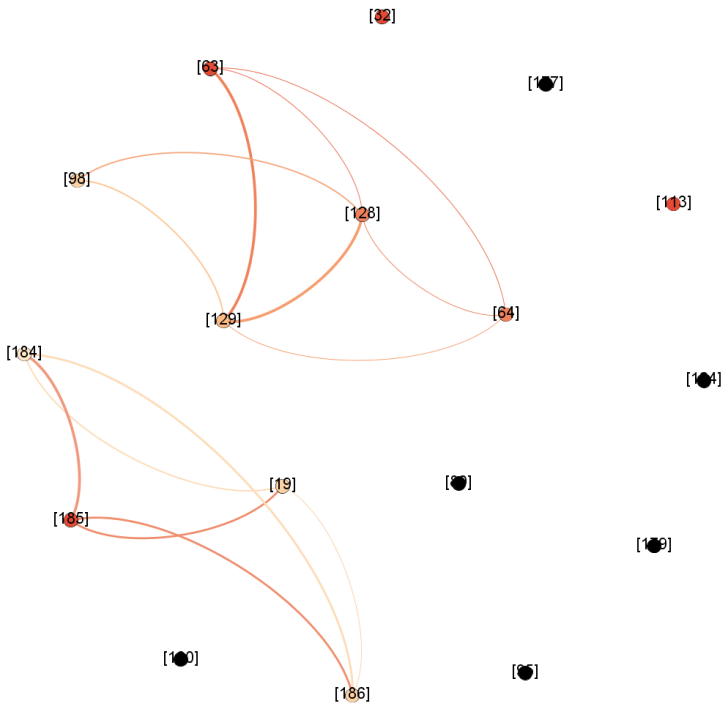


Figure 2: Graph showing the 17 papers as vertices. There is an edge between papers that have a common author. The thickness and color reflect the number of common authors. K_3 and K_4 cliques are visible.

Table 2: The latest surveys, reviews, and meta-analysis on automatic deception detection. Are the queries (**QL**) or databases **DB** listed? **Period** of the searches. The number of **papers** surveyed. Is there support for **ling**uistic features?

Reference	QL	DB	Period	Papers	Ling?
GA19 Gröndahl and Asokan [2019]	No	No	–	18	No
H16 Hauch [2016]	Yes	Yes	2011–12	44	Yes
E19 Elhadad et al. [2019]	Yes	Yes	2017–19	47	Partial

7.1 Prior Analyses

The meta-analyses DePaulo et al. [2003] and Hauch [2016] did find small markers of deception in the studies they examined despite analyzing studies of specific forms or situations of deception, not general domain-independent datasets.

Likewise, the following papers all point to evidence for cross-domain deception detection: Rill-García et al. [2018], Shahriar et al. [2021], Vogler and Pearl [2020], Xarhoulacos et al. [2021], Yeh and Ku [2021]. These researchers created domain-independent datasets and developed features and techniques for deception detection across domains.

The meta-analysis of Hauch [2016] searched four databases: PsycInfo, Social Science Citation Index, Dissertation Abstracts, and Google Scholar for articles between 1945 and 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 Elhadad et al. [2019] searched Google Scholar for articles between 2017–2019 using 10 queries listed in their paper. Their queries are a *proper* subset of the Boolean query `(fake ∨ false) news (identify ∨ detect) on (social media ∨ Twitter)`, which we repeated on Scholar on 11 November 2022, with a claim of over a million results. (Google counts are loose upper bounds.) Scholar displayed only the top 1000 results. The queries produced a total of fewer than 200 potentially relevant results. We summarize the pertinent characteristics of three recent reviews/surveys/meta-analyses in Table 2.

7.2 Our Analysis

Since the meta-analysis of Hauch [2016] ended in February 2012, we searched Google Scholar, PsycInfo and Dissertations, and Abstracts Global, for the period 2013–2022 with the query:

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

We formed this query by appropriately combining the queries from Hauch [2016], Elhadad et al. [2019], adding keywords after scanning the initial results, and adding relevant synonyms from querying WordNet 3.1 with `deceit`, `identify`, and `lexical`. Adding `recognition` to the middle clause reduced the set of results by more than 100K, a flaw of Google Search. (We tried other synonyms, but the results seem irrelevant.) Scholar claimed over a million results but only displayed the top thousand. A scan through them identified approximately 880 as potentially relevant. PsycInfo gave us around 450 matches and the Dissertations database yielded approximately 140 matches. The Scholar query:

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

over all time periods claimed 350K results; the top 1000 gave around 190 potentially relevant ones. The results were examined for feature selection and feature ranking papers. More than one recent survey mentioned n-grams of part of speech tags and semantic features as examples of generalizable features. However, this analysis also revealed a lack of feature rankings for large, diverse, general datasets.

7.3 New Developments in NLP

Moreover, computer science, machine learning, and NLP have come a long way in the intervening years. Recent breakthroughs such as attention, transformers, and pre-trained language models like BERT, have revolutionized NLP. Even if some of the previous criticisms were valid, we must reexamine the conclusions of the Critiques considering these new advances.

8 Linguistic Cues/Analysis

Because of the problems enumerated above, we collect and analyze datasets for domain-independent linguistic cues to tackle: (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 Gokhman et al. [2012].

8.1 Datasets

We summarize our deception domains and scenarios below. We focus on real-world datasets.

In the *product review* domain, we use the Amazon reviews dataset mentioned above Garcia [2019].

In the *job scam* domain, we identify fraudulent job listings. Our dataset contains the bodies of 13,735 legitimate and 608 fraudulent job listings.

In the *phishing* domain, we distinguish between legitimate emails and phishing emails. Our dataset contains the bodies of 9,202 legitimate and 6,134 phishing samples. The IWSPA-AP dataset analyzed above is a subset of this dataset.

In the *political statement* domain, we determine the truthfulness of claims made by US political speakers. Our dataset contains 7,167 truthful and 5,669 deceptive statements evaluated by PolitiFact.

In the *fake news* scenario, we distinguish between legitimate and fake news. Here we use the WELFake dataset Verma et al. [2021].

We analyzed each dataset for any artifacts of data collection and cleaned them to remove such artifacts. The cleaning procedures include two parts: text removal and text cleaning. We then sanitize the texts using the methods discussed in Zeng et al. [2022]. We remove meta-data in emails and source leaks in news and replace HTML break tags with new lines. Additionally, the authors of Zeng et al. [2022] found that the provided labels in WELFake Verma et al. [2021] are flipped, so we flip its labels as a final cleaning step. We are making the combined, cleaned dataset available on Zenodo at `https://zenodo.org/record/6512468#.ZBVRUhTMLQM`.

8.2 Sources for Linguistic Cues

Function words (FW) are words that express a grammatical relationship between words in a sentence. Unlike content words, function words such as 'when,' 'at,' and 'the' are independent of specific domains. Function words and n-grams are useful for many text classification tasks, including author gender classification, authorship attribution Argamon and Levitan [2005], and deception detection Siagian and Aritsugi [2020]. To gain an insight into the transfer of knowledge between domains, we utilized three types of explainable features: function words, part-of-speech (POS) tags of function words, and engineered linguistic features. POS tags were used to determine whether a word was a function or a content word; the content words were then removed. The last experiment utilized 151 engineered linguistic features ($13 + 55 + 86 - 3$ duplicates removed by the colinearity check below).

The engineered features are drawn from three sources. Linguistic Inquiry and Word Count Boyd et al. [2022], a popular source of features in the NLP literature, was the source of 86 features. The authorship

Table 3: Function word features: N – number of common features; F – fake news, J – job scams; P – phishing; Pr – product reviews, Ps – political statements; CC – cumulative count including common features inherited from supersets.

Subset	N	Common Features	CC
All	6	and, in, is, of, on, the	6
F, J, P, Pr	2	this, you	8
F, J, Pr, Ps	1	are	7
J, P, Pr, Ps	2	for, to	8
F, J, P	1	at	9
F, Pr, Ps	3	it, that, would	10
J, P, Ps	2	from, our	10
J, Pr, Ps	2	as, with	11
P, Pr, Ps	1	not	9
F, P	2	all, had	11
F, Ps	1	he	11
J, Pr	2	be, or	13
J, Ps	1	we	12
P, Pr	1	me	10
Pr, Ps	2	they, was	17

Table 4: Function word part-of-speech features: N – number of common features; F – fake news; J – job scams; P – phishing; Pr – product reviews; Ps – political statements; CC – cumulative count including common features inherited from supersets.

Subset	N	Common Features	CC
All	10	CC, CD, DT, IN, MD, PRP, RB, TO, VBP, VBZ	10
F, J, P, Ps	5	RP, VB, WDT, WP, WRB	15
F, P, Pr, Ps	1	VBD	11
F, J, P	2	POS, UH	17
F, P, Ps	2	EX, VBN	18
F, J	1	VBG	18
J, P	1	ADD	18

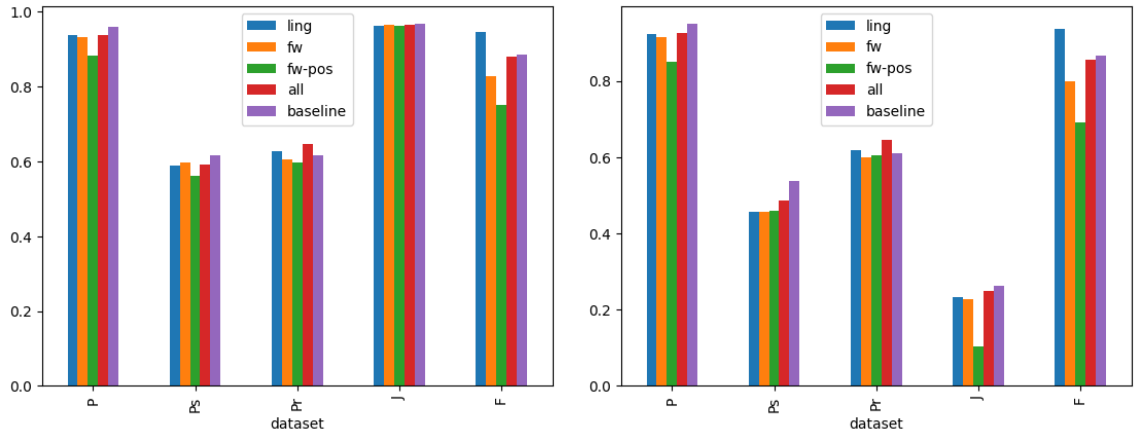
attribution paper Fabien et al. [2020] was the source of 55 features. Thirteen features were collected from two papers, one on deception Zhou et al. [2004] and the other on fake news Verma et al. [2021], after significance testing using t-tests with and without the Bonferroni-Holm correction of p-values.

The initial significance testing of 27 linguistic features from the two papers Zhou et al. [2004], Verma et al. [2021] on four public datasets is described in Appendix A. Appendix B describes an analysis of function word n-grams on the same datasets as in Appendix A. A complete source-wise list of the 55 features from Fabien et al. [2020] and 86 features from Boyd et al. [2022] is in Appendix C. Function words as features for deception have been studied before, in Siagian and Aritsugi [2020], for example. We also experimented with the part-of-speech tags of function words.

8.3 Results of Feature Analysis

We used the Stanza Qi et al. [2020] POS tagger and OntoNotes Release 5.0/Penn Treebank Marcus et al. [1993] tagset in all experiments involving POS tags. This tagset builds on top of the original Penn Treebank, and adds seven new tags:

```
ADD - Email, AFX - Affix, HYPH - Hyphen, NFP - Superfluous punctuation, UH -
```



(a) Accuracies (b) F1-scores

Figure 3: Random Forest performance for the five feature types: linguistic (ling), function words (fw), pos tags of function words (fw-pos), combination of the three (all), and unigram tfidf (baseline); F – fake news; J – job scams; P – phishing; Pr – product reviews; Ps – political statements.

Table 5: Engineered Linguistic Features: N – number of common features; F – fake news; J – job scams; P – phishing; Pr – product reviews; Ps – political statements; CC – cumulative count including common features inherited from supersets.

Subset	N	Common Features	CC
All	1	per_cap	1
J, P, Pr, Ps	5	Dic, f_b, f_g, per_digit, richness	6
F, J, P	4	cert, f_e_2, function, sen_len	5
F, J, Pr	1	period	2
F, P, Pr	1	paus	2
P, Pr, Ps	3	conj, f_f, modi	9
F, J	2	apostro, comm	8
F, P	5	f_e_0, f_e_1, f_e_3, f_e_7, sens	11
F, Pr	10	adverb, allPunc, analytic, f_e_8, focuspast, ipron, len_text, OtherP, pronoun, WPS	13
J, P	5	f_c, f_o, f_v, f_w, socrefs	15
P, Pr	7	avg_len, f_d, f_i, f_s, f_t, f_y, self_ref	17
P, Ps	2	f_1, f_p	11
J, Pr	3	allure, article, lifestyle	10
Pr, Ps	1	quantity	10

`Interjection, SP - Space, and XX - Unknown.`

Due to the parser's limitations, several samples of text that had a length more than one million characters had to be discarded. We did not remove stop words or further alter the data in any manner. Function words and their respective POS tags were separately vectorized as word unigrams using the tf-idf scheme. The raw texts were processed and vectorized identically and used as a baseline. The motivation behind it was to (i) understand whether it is possible to achieve similar results while using only a few non-domain-specific features that are highly indicative of deception, and (ii) investigate the impact of content words on deception through the contrast between the baseline and function words.

For each dataset, and for each set of features, we applied three techniques to select the most relevant features. First, a random forest algorithm Breiman [2001] was used, which allowed us to rank features by their importance. The least important ones were removed under the condition that the out-of-bag accuracy on the validation set either increased or remained the same after removing the features. Next, we applied scipy's Virtanen et al. [2020] single linkage hierarchical clustering Gower and Ross [1969] with Spearman's correlation Spearman [1904] as the measure of feature colinearity. Features exhibiting a high degree of colinearity were removed with their redundancy validated in the same manner as with the first technique. Finally, taking the remaining features, we applied Hyperopt's Bergstra et al. [2013] feature selection and the eXtreme Gradient Boosting algorithm Chen and Guestrin [2016] with SHAP Lundberg and Lee [2017] as a metric of each feature's contribution to the overall model performance. Ultimately, the aforementioned approach produced a subset of the features for each of the five datasets. A total of 81 linguistic, 28 function word POS, and 61 function word features were selected; 50/81, 22/28, and 29/61 were shared with at least one other dataset.

For our analysis of the potential for knowledge transfer any feature unique to a dataset was removed, leaving only those significant for at least two datasets and therefore being of interest for understanding of transfer between domains. The relationships of function words and function words' POS tags across datasets are depicted in Tables 3 and 4, while linguistic feature transfer is summarized in Table 5. The cumulative count (CC) in these tables serves as a measure of how many features a group of domains has in common. Several trends can be noticed from these three tables. For example, all five datasets share $6 + 10 + 1 = 17$ common features, and the fake news, job scams, and phishing datasets have a total of $9 + 17 + 5 = 31$ features in common. Also, the subset {F, J} has $9 + 18 + 8 = 35$ common features, and {J, P, Pr, Ps} has $8 + 10 + 8 = 26$ common features. Job scams and phishing together have 43 common features. Similarly, we can see that deceptive attacks can be differentiated using features such as 'to', personal pronouns, singular present verb forms, modals, and adverbs (compare with the quote from Rowe Rowe [2006] in Section 3.1.1). The richness, possessive ending and interjection features are significant for fake news, job scams and phishing. Fake news and product reviews have many significant LIWC features.

Datasets that share a significant number of features are good candidates for domain adaptation; however, the performance of a model using a potentially limited set of features shared across tasks must remain robust. To this end, we combined previously selected linguistic, function words, and function word POS features that were shared by two or more datasets. This resulted in a final set of 91 features. Upon further applying feature selection, the number of significant features of all three types shared among datasets has been reduced to 45.

To evaluate the features' performance, we used a random forest classifier with five-fold cross-validation. The model hyperparameters were set to 50 trees with the leaf nodes of 5 samples, and 50% of the features were considered on each split. Gini impurity was used as a criterion of split quality.

The accuracy and F1-scores of the model using each of the feature sets across the five datasets are shown in Figure 3. It is important to note that Job Scams' data appears to be heavily imbalanced and the models' performance on it is not an ideal indicator of feature quality. Generally, the combined set of shared features is nearly on par with the baseline, with linguistic, function word, and function word POS following in the order given. Notable exceptions are Product Reviews where linguistic and combined features beat the others, including the baseline, and Fake News with linguistic features outperforming the rest by a significant margin. We hypothesize that the relative length and richness of news articles may be in part responsible for this phenomenon.

9 Deep Learning Based Experiments

To investigate the possible existence of other deception signals, we turn to deep learning. If universal deception signals exist, then a deep-learning model can learn to detect them. To determine whether this happens, we perform two experiments on the same five cleaned datasets of the previous section. First, we evaluate the performance of models trained on multiple domains. Then we train models on one domain and evaluate their performance on other domains.

9.1 Model

Our model architecture consists of a base pre-trained transformer model, a dropout layer, and a linear layer. As standard in NLP, we prepend a [CLS] token to the text, pass the text through the base model, and perform classification on the last-layer embedding of the [CLS] token.

9.2 Multi-domain Experiment

If deep-learning models trained on multiple domains pick up on universal deception signals, then we should expect performance on *individual* domains to be positively correlated amongst each other. Conversely, if they only learn domain-specific signals, then we should expect performance on individual domains to be negatively correlated with one another.

We train 100 models on the union of our datasets. We use a random 80/10/10 train/validate/test split for each dataset with uniformly drawn hyperparameters. We use BERT-base and RoBERTa-base for our base models, dropout percentages between 0.1 and 0.5, and the AdamW optimizer with learning rates between 0.00001 and 0.0001.

We then evaluate each model on the individual test sets. We exclude models that failed to converge and models that have an outlier F1 score using the IQR test and perform pairwise linear regression on the remaining F1 scores.

We present our results without outliers in Figure 4. All pairs of tasks except for product reviews and phishing are positively correlated, with five of them significant at the 0.05 level.

9.3 Cross-Domain Generalization Experiment

If a deep-learning model primarily learns a universal deception signal, then it should generalize to deception domains that it has not yet seen. In particular, they should be able to achieve a higher F1 score than a coin flip classifier, which we can calculate using the formula CF F1 $= q/(0.5 + q)$, where q is the portion of the dataset that is deceptive.

On each dataset, we train 100 models with hyperparameters drawn from uniform distributions. We use BERT-base and RoBERTa-base for our base models and values between 0.1 and 0.5 for dropout percentage. For our learning rate, we use a different range for each task to minimize divergence: $[0.00001, 0.00006]$ for product reviews, $[0.00001, 0.000025]$ for job scams, $[0.00001, 0.00010]$ for phishing, $[0.00001, 0.00004]$ for political statements, and $[0.00001, 0.00010]$ for fake news.

Each model is evaluated on each dataset, ignoring models that fail to converge. We perform a 1-sample t-test with the alternative hypothesis "the mean F1 in domain Y of models trained on X is less than or equal to the coin flip F1 of Y." We report the resulting p-values in Table 6. In ten cases, models trained on one domain manage to beat the coin-flip baseline at a 0.01 significance level, with nine cases beating the coin-flip baseline at the 10σ ($p < 7.62 \times 10^{-24}$) level. However, we also find that eight pairs have a p-value of 1.00, meaning they performed worse than the coin-flip baseline.

Interestingly, we also find that the fake news models manage to beat the coin flip on all domains. We suspect that this is due in part to its larger size but leave this as a direction for future research.

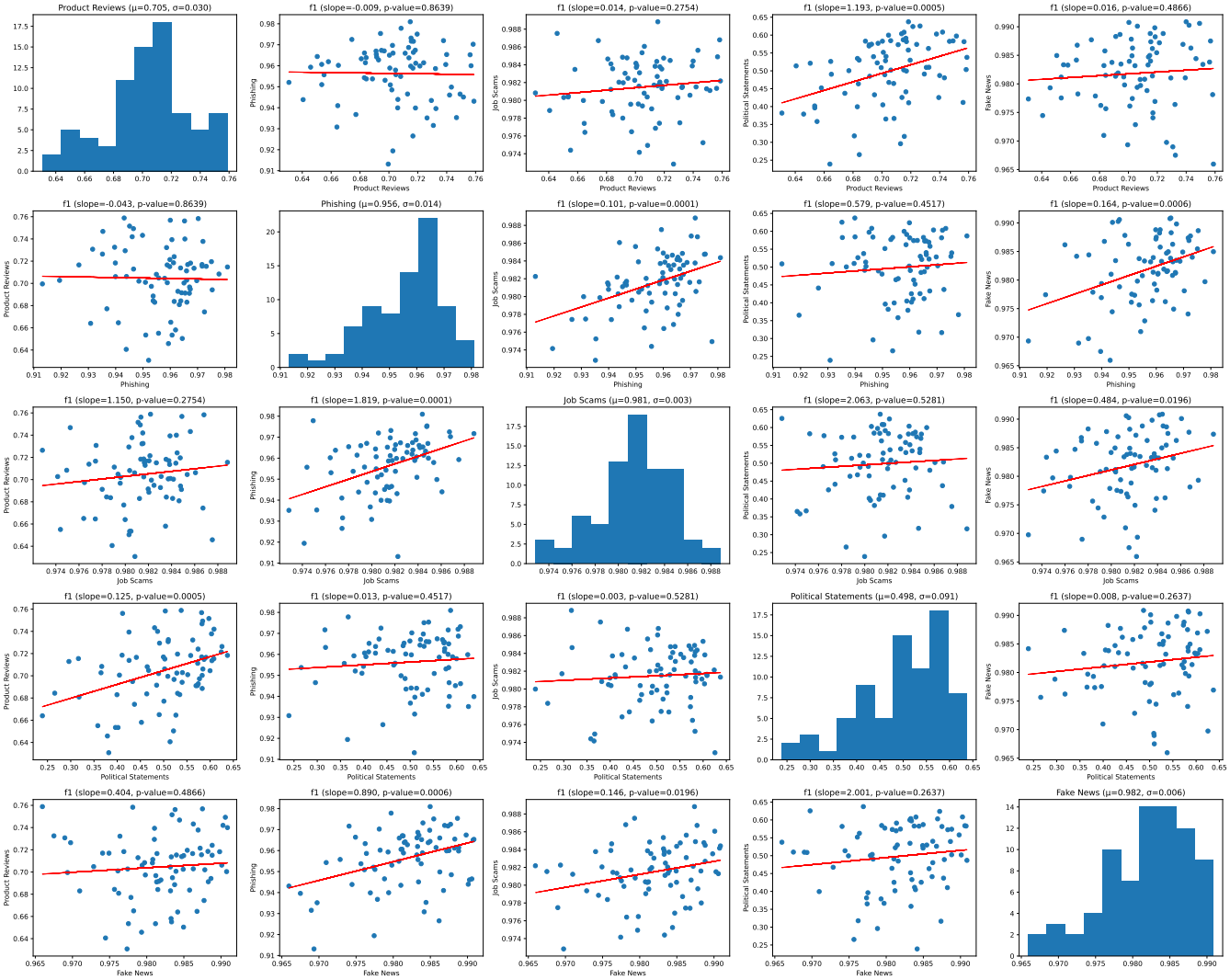


Figure 4: Pairwise F1 score scatter matrix of converged combined models. Outliers are excluded.

Table 6: p-values in the Cross Dataset Experiment. Values below 0.01 are considered significant. A dagger indicates that the 0.00 values are correct to two decimal places. The 0.00 values without the dagger have a 0 in at least the third place after the decimal.

	Product Reviews	Phishing	Job Scams	Political Statements	Fake News
Product Reviews	$0.00^{\dagger}$	1.00	$0.00^{\dagger}$	1.00	1.00
Phishing	1.00	$0.00^{\dagger}$	$0.00^{\dagger}$	1.00	1.00
Job Scams	$0.00^{\dagger}$	0.98	$0.00^{\dagger}$	$0.00^{\dagger}$	$0.00^{\dagger}$
Political Statements	1.00	1.00	0.00	0.00	0.96
Fake News	$0.00^{\dagger}$	$0.00^{\dagger}$	$0.00^{\dagger}$	$0.00^{\dagger}$	$0.00^{\dagger}$

9.4 Discussion

The multi-domain experiment provides strong support for the existence of universal deception signals. All but one pair are positively correlated. Five are statistically significant, and the one negative correlation is not statistically significant. In contrast, the results of our cross-domain generalization experiment are mixed. While some pairs beat the coin-flip baseline, others performed worse than the baseline.

Taken together, these results suggest that both universal and domain-specific deception signals exist. Models trained on a single task will learn both universal and task-specific signals, potentially resulting in poor generalization to other deception domains. Therefore, training a domain-independent deception detector will likely require a diverse domain-independent dataset.

10 Conclusions

We have provided new definitions for deception based on explanations and probability theory. We gave a new taxonomy of deception that clarifies the explicit and implicit elements of deception. We have given sound desiderata for systematic review and meta-analysis, which we hope will help researchers conduct high-quality analyses of the literature and devise new domain-independent deception detection techniques.

We have argued against hasty conclusions regarding linguistic cues for deception detection and especially their generalizability. The Critiques contained in Fitzpatrick et al. [2015], Vogler and Pearl [2020], Vrij [2008] may present a valid point, namely that some linguistic cues might not generalize across the broad class of attacks. However, over-generalizations should be made with caution, as they discourage future domain-independent deception research. Moreover, we have presented evidence showing that there do exist common linguistic cues in deceptive attacks with widely varying goals and topical content.

Our linguistic analysis of four datasets and cross-dataset analysis of five different deception datasets shows that there are linguistic features, some at the surface level and some deeper, that can be used to build classifiers for more general deception datasets. With all the new developments in machine learning and NLP, we believe that research on linguistic deception detection is poised to take off and could result in significant advances.

acknowledgments

We thank all those who supplied datasets for this research and Vu Minh Hoang Dang for his comments on a previous draft of this article.

Verma's research was partially supported by NSF grants 1433817, 1950297, 2210198, and 2244279, ARO grants W911NF-20-1-0254, W911NF-23-1-0191, and ONR award N00014-19-S-F009. He is the founder of Everest Cyber Security and Analytics, Inc. Boumber's research was partly supported by ARO award W911NF-20-1-0254. Zeng's research was supported by ONR award N00014-19-S-F009. Liu's research was supported by NSF award 1950297.

References

Shlomo Argamon and Shlomo Levitan. Measuring the usefulness of function words for authorship attribution. In *Proceedings of the 17th Joint International Conference on Humanities Computing and Digital Scholarship*, pages 4–6. The Association for Computers and the Humanities and The Association for Literary and Linguistic Computing, June 2005.

Shivam Bansal and Chaitanya Aggarwal. TextSTAT, 2019. Online `https://pypi.org/project/textstat`; accessed 31 July 2023.

J. Bowyer Bell and Barton Whaley. *Cheating and Deception*. Routledge, 2017.

James Bergstra, Dan Yamins, David D Cox, et al. Hyperopt: A python library for optimizing the hyperparameters of machine learning algorithms. In *Proceedings of the 12th Python in science conference*, volume 13, page 20. Citeseer, 2013.

Judith Bogert. In defense of the fog index. *The Bulletin of the Association for Business Communication*, 48 (2), June 1985.

Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. The development and psychometric properties of LIWC-22. Technical report, University of Texas at Austin, Austin, TX, 2022.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL `http://dx.doi.org/10.1023/A%3A1010933404324`.

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.

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. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.178`.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL `http://doi.acm.org/10.1145/2939672.2939785`.

Robert B. Cialdini. *Influence: The Psychology of Persuasion*. William Morrow, New York, 2006. revised edition.

Giovanni Da San Martino, Preslav Nakov, Jakub Piskorski, and Nicolas Stefanovitch. SemEval 2023 task 3: "Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup"', 2023. `https://propaganda.math.unipd.it/semeval2023task3`.

Avisha Das, Shahryar Baki, Ayman El Aassal, Rakesh M. Verma, and Arthur Dunbar. SoK: A comprehensive reexamination of phishing research from the security perspective. *IEEE Commun. Surv. Tutorials*, 22(1):671–708, 2020.

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.

Daniel Druckman and Robert A. Bjork, editors. *In the Mind's Eye: Enhancing Human Performance*. National Academy Press, 1992. National Research Council (U.S.). Committee on Techniques for the Enhancement of Human Performance.

James F. Dunnigan and Albert A. Nofi. *Victory and Deceit: Deception and Trickery at War*. Writers Club Press, 2001.

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.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, 2020.

Song Feng, Ritwik Banerjee, and Yejin Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 171–175, 2012.

Charles J. Fillmore. The case for case. In *Universals in Linguistic Theory*. Holt, Rinehart & Winston, New York, 1968.

Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. *Automatic Detection of Verbal Deception*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2015. doi: 10.2200/S00656ED1V01Y201507HLT029. URL https://doi.org/10.2200/S00656ED1V01Y201507HLT029.

Dariusz Galasinski. *The Language of Deception: A Discourse Analytical Study*. Sage Publications, 2000.

Liev Garcia. Amazon-reviews-dataset, 2019. Online; https://www.kaggle.com/lievgarcia/amazon-reviews; accessed 2 August 2023.

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.

Maria Glenski, Ellyn Ayton, Robin Cosbey, Dustin Arendt, and Svitlana Volkova. Towards trustworthy deception detection: Benchmarking model robustness across domains, modalities, and languages. In *Proceedings of the 3rd International Workshop on Rumours and Deception in Social Media (RDSM)*, pages 1–13, 2020.

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. URL https://aclanthology.org/W12-0404.

John C. Gower and Gavin J. S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of The Royal Statistical Society Series C-applied Statistics*, 18:54–64, 1969. URL https://api.semanticscholar.org/CorpusID:18902751.

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.

Marco Guerini, Oliviero Stock, and Massimo Zancanaro. A taxonomy of strategies for multimodal persuasive message generation. *Applied Artificial Intelligence*, 21(2):99–136, 2007.

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

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. doi: 10.1007/s00500-016-2409-2. URL https://doi.org/10.1007/s00500-016-2409-2.

Matthew Honnibal. Introducing spaCy. https://explosion.ai/blog/introducing-spacy, 2015.

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.

Timothy R. Levine. *Encyclopedia of Deception*, volume 2. Sage Publications, 2014.

Edward Loper and Steven Bird. NLTK: The natural language toolkit. *arXiv*, cs/0205028, 2002.

Stephan Ludwig, Tom Van Laer, Ko De Ruyter, and Mike Friedman. Untangling a web of lies: Exploring automated detection of deception in computer-mediated communication. *Journal of Management Information Systems*, 33(2):511–541, 2016.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, jun 1993. ISSN 0891-2017.

Rada Mihalcea and Carlo Strapparava. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference: Short Papers*, pages 309–312, 2009.

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.

Onyekware U. Oluoha, Terungwa S. Yange, George E. Okereke, and Francis S. Bakpo. Cutting edge trends in deception based intrusion detection systems – a survey. *Journal of Information Security*, 12(4):250–269, 2021.

Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. *arXiv*, 1107.4557, 2011.

Subhadarshi Panda. *Deception Detection Across Domains, Languages and Modalities*. PhD thesis, City University of New York, USA, 2022. URL https://academicworks.cuny.edu/gc_etds/5015.

Subhadarshi Panda and Sarah Ita Levitan. Deception detection within and across domains: Identifying and understanding the performance gap. *ACM J. Data Inf. Qual.*, 15(1):7:1–7:27, 2023.

Jeffrey Pawlick and Quanyan Zhu. *Game Theory for Cyber Deception*. Springer, 2021.

Jeffrey Pawlick, Edward Colbert, and Quanyan Zhu. A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Comput. Surv.*, 52(4), aug 2019. ISSN 0360-0300. doi: 10.1145/3337772. URL https://doi.org/10.1145/3337772.

Veronica Perez-Rosas. *Exploration of Visual, Acoustic, and Physiological Modalities to Complement Linguistic Representations for Sentiment Analysis*. PhD thesis, University of North Texas, 2014. URL http://search.proquest.com.ezproxy.lib.uh.edu/dissertations-theses/exploration-visual-acoustic-physiological/docview/1725125492/se-2?accountid=7107.

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.

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*, pages 101–108, July 2020. URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Simone Raponi, Zeinab Khalifa, Gabriele Oligeri, and Roberto Di Pietro. Fake news propagation: a review of epidemic models, datasets, and insights. *ACM Transactions on the Web (TWEB)*, 16(3):1–34, 2022.

Yafeng Ren and Donghong Ji. Learning to detect deceptive opinion spam: A survey. *IEEE Access*, 7: 42934–42945, 2019.

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.

Neil Rowe. A taxonomy of deception in cyberspace. In *International Conference on Information Warfare and Security*, pages 173–181, 2006.

Sadat Shahriar, Arjun Mukherjee, and Omprakash Gnawali. A domain-independent holistic approach to deception detection. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, pages 1308–1317, 2021.

Sadat Shahriar, Arjun Mukherjee, and Omprakash Gnawali. Deception detection with feature-augmentation by soft domain transfer. In Frank Hopfgartner, Kokil Jaidka, Philipp Mayr, Joemon Jose, and Jan Breitsohl, editors, *Social Informatics*, pages 373–380, Cham, 2022. Springer International Publishing. ISBN 978-3-031-19097-1.

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.

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. ISSN 1936-1955. doi: 10.1145/3349536. URL https://doi.org/10.1145/3349536.

Charles Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15(1):72–101, 1904.

Mark Staples and Mahmood Niazi. Experiences using systematic review guidelines. *Journal of Systems and Software*, 80(9):1425–1437, 2007. ISSN 0164-1212. doi: https://doi.org/10.1016/j.jss.2006.09.046. URL http://www.sciencedirect.com/science/article/pii/S0164121206002962.

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.

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. doi: 10.1177/0261927X09351676. URL https://doi.org/10.1177/0261927X09351676.

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. doi: 10.1109/TCSS.2021.3068519.

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. doi: 10.1145/3319535.3363267. URL `https://doi.org/10.1145/3319535.3363267`.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

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

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

Barton Whaley and Susan Stratton Aykroyd. *Textbook of Political-Military Counterdeception: Basic Principles & Methods*. National Defense Intelligence College, 2007.

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. doi: 10.3390/s21165496. URL `https://doi.org/10.3390/s21165496`.

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.

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, pages 2077–2079, New York, NY, 2020. Association for Computing Machinery. ISBN 9781450370899. doi: 10.1145/3372297.3420017. URL `https://doi.org/10.1145/3372297.3420017`.

Victor Zeng, Xuting Liu, and Rakesh M. Verma. Does deception leave a content independent stylistic trace? In *Proceedings of the Twelfth ACM Conference on Data and Application Security and Privacy (CODASPY '22)*, New York, NY, 2022. Association for Computing Machinery. doi: 10.1145/3508398.3519358. URL `https://doi.org/10.1145/3508398.3519358`.

Lina Zhou and Dongsong Zhang. An ontology-supported misinformation model: Toward a digital misinformation library. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37 (5):804–813, 2007.

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. ISSN 0926-2644. doi: 10.1023/B:GRUP.0000011944.62889.6f.

Xinyi Zhou, Reza Zafarani, Kai Shu, and Huan Liu. Fake news: Fundamental theories, detection strategies and challenges. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 836–837, 2019.

Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. Fake news early detection: A theory-driven model. *Digital Threats: Research and Practice*, 1(2):1–25, 2020.

Table A7: Statistics of the four available datasets covering different domains.

Dataset	Size	Truthful / Deceptive	Category
Amazon Reviews	20,976	10,481 / 10,495	Fake reviews
DecOp	1,250	625 / 625	Deceptive opinions
IWSPA-AP	5,026	4,429 / 597	Phishing emails
WELFake	62,101	34,615 / 27,486	Fake news
Total	89,353	50,150 / 39,203	Combined

A Significance Testing of Linguistic Cues from Deception Literature

For significance-testing of the linguistic cues from the deception literature Zhou and Zhang [2007], Verma et al. [2021], we did a preliminary analysis of these four datasets. DecOp Capuozzo et al. [2020] containing deceptive opinions, the WELFake dataset Verma et al. [2021] containing fake news, the IWSPA-AP dataset Verma et al. [2019] containing phishing/legitimate emails, and the Amazon Reviews dataset Garcia [2019] consisting of truthful and fake product reviews. All are publicly available, but one of them, DecOp, is a small laboratory dataset with its limitations.

The goals in these datasets are quite diverse. Phishing email attackers wish to install malware or steal identity/money. Deceptive opinion/review authors wish to sway opinions on services or products. Fake news authors wish to sway elections, divide people, or cause chaos. Fake product reviews are designed to sell more of a certain product or depress the sales of competitors.

- The DecOp Dataset: This dataset is from Capuozzo et al. [2020]. It contains truthful/deceptive opinions on several topics such as abortion and cannabis legalization. These opinions were collected using crowdsourcing in the US and Italy. The researchers also trained transformer models that achieved 0.62–0.90 accuracies with different settings.
- The WELFake Dataset: This dataset is from Verma et al. [2021]. It draws upon multiple true/fake news datasets.
- The IWSPA-AP Dataset: This is the IWSPA Anti-Phishing competition dataset of emails Verma et al. [2019].
- The Amazon Reviews Dataset: This comprises real and fake Amazon reviews from a Kaggle repository Garcia [2019].

Dataset statistics are shown in Table A7. Of these, the WELFake and Amazon Reviews Dataset are also included in the main sections.

We analyzed each dataset for any artifacts of data collection and cleaned them to remove such artifacts. The cleaning procedures include two parts: text removal and text cleaning. We removed a total of 10,728 duplicate, non-English, or empty bodies, giving a total of 89,353 items. We then sanitize the texts using the methods discussed in Zeng et al. [2022]. We remove meta-data in emails and source leaks in news and replace HTML break tags with new lines. Additionally, the authors of Zeng et al. [2022] found that the provided labels in WELFake Verma et al. [2021] are flipped, so we flip its labels as a final cleaning step. We removed a total of 10,728 duplicate, non-English, or empty bodies, giving a total of 89,353 items.

A.1 Features: Linguistic Cues

We extracted 27 total textual features from the literature shown in Table A8. Features 1–27, with 15, 16 and 23 skipped, are from Zhou et al. [2004], and features 28–30 are from Verma et al. [2021]. We then measure their values on the deceptive and legitimate samples of each dataset and count the number of datasets with

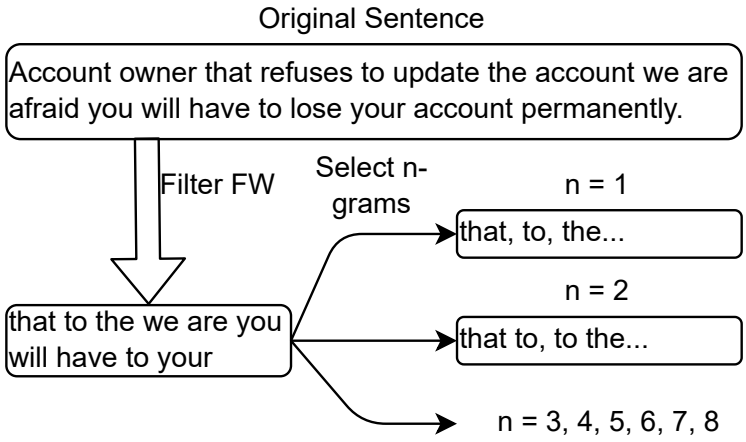


Figure A5: An example of extracting function-word (FW) n-grams from a sentence.

statistically significant differences (using appropriate statistical tests). A difference is significant if and only if its p-value, after Bonferroni-Holm correction, is smaller than the threshold 0.01. Since there is some debate on the multiple comparisons issue (e.g., see Gelman et al. [2012]), we report statistically significant features both with and without the correction in Table A9.

A.2 Results: Selected Features

We find that the following features are statistically lower for deceptive samples in three or more datasets:[4] number of words, number of verbs, number of sentences, sentence length, word length, pausality (number of punctuation marks per sentence), number of modifiers (adjectives and adverbs), self reference (number of first-person plural pronouns), redundancy (number of function words per sentence), and the Automatic Readability Index.

We find that the following features are statistically higher for deceptive samples in three or more datasets: modal verbs, certainty (number of words that have the certainty tag in the LIWC 2015 Dictionary per word), and the SMOG readability index. We used Stanza Qi et al. [2020] for POS tagging. Removing positive affect, since it is covered by our LIWC list of features below, gives us 13 features.

B Function Word n-grams

In this section, we describe our initial analysis of function word n-grams for the same four public datasets as in the previous section of the appendix.

B.1 Method

We combine function words and n-grams by looking for n-grams of function words that appear significantly more or less often in deceptive texts than truthful texts.

We first extract the function words from our texts using a list from the publicly-available PhishBench 2.0 Zeng et al. [2020] and compute function-word n-grams for n from one through eight (an example of this process is in Figure A5).[5]

[4]The precise number of datasets with fewer is in Table A9.

[5]We also explored POS tagging, which produced similar results at greater computational cost.

Table A8: List of linguistic cues. Features 1–27 are from Zhou et al. [2004], and features 28–30 are from Verma et al. [2021]. Features with an asterisk (∗) are selected after testing on the four available datasets. Features in **bold** remain qualified after adjusting their p-values per the Bonferroni-Holm method.

1. **words*** (words)	W(D). NLTK's word tokenizer Loper and Bird [2002] 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*** (verbs)	Num-verbs(D). NLTK's word tokenizer was used to identify verbs.	17. **self reference*** (self_ref)	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 Honnibal [2015] was used to identify noun phrases.	18. group reference	Num-first person plural pronoun(D) (i.e., Num-$\{we, us\}$)/W(D).
4. **sentence*** (sens)	S(D). NLTK's sentence tokenizer was used.	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 Qi et al. [2020] was used for 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*** (sen_len)	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 opposed to function words. Methods to identify content/function words are discussed in the Function Word n-gram section in the Appendix.
7. average word length* (word_len)	Num-characters(D) / W(D). Characters include digits, punctuation, and spaces.	22. redundancy* (redun)	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 Feature (3).	23. typographical error ratio [skipped]	*This feature is skipped because exploratory analysis showed that the typographical error ratio is zero for most texts in both categories. The popularity of the auto-correct feature on browsers and text editing software has probably diminished the effectiveness of this feature.*
9. **pausality*** (paus)	Num-punctuation marks(D) / S(D)	24. spatiotemporal information	Num-('space' + 'time') / W(D), where 'space' and 'time' refer to the Num-words with the tag 'space' and 'time' in the LIWC2015 dictionary. LIWC is a dictionary that associates words with various tags; we used a commercial program from Pennebaker Conglomerates Tausczik and Pennebaker [2010].
10. **modifier*** (modi)	Num-adjectives and adverbs(D). A word is an adjective or adverb iff word.upos = 'ADJ' or word.upos = 'ADV'. Stanza was used for POS tagging.	25. perceptual information	Num-'percep' / W(D). 'percep' is defined similarly, as in Feature (24).
11. modal verb* (modal)	Num-modal verbs(D) / W(D). A word is a modal verb iff word.upos = 'AUX' and word.xpos = 'MD'. Stanza was used for POS tagging.	26. positive affect	Num-'posemo' / W(D). 'posemo' is defined similarly, as in Feature (24).
12. certainty* (cert)	Num-words that have the tag 'certain' in the LIWC2015 dictionary(D) / W(D).	27. negative affect	Num-'negemo' / W(D). 'negemo' is defined similarly in Feature (24).
13. other reference	Num-third person pronoun(D) / W(D). A word is a third person pronoun iff word.xpos = 'PRP' and word.feats = 'Person=3'. We used Stanza for POS tagging.	28. Gunning fog grade readability index Bogert [1985]	An index to quantify the readability of a text by estimating the years of education required to understand the text. We used TextSTAT Bansal and Aggarwal [2019] to calculate it.
14. passive voice	Num-passive voice verb(D) / W(D). A word is a passive voice verb iff word.deprel = 'aux:pass'. We used Stanza for POS tagging.	29. **SMOG readability index*** (smog)	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* Zhou et al. [2004]. 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* (ari)	Similar to Features (28) and (29). We also used TextSTAT for this.

Table A9: Behavior of features that show statistically significant differences between the truthful and deceptive classes. *# sig* is the number of datasets where the feature shows a significant difference. A positive number means the feature value is higher in the deceptive class. Emboldened features are those still qualified after the Bonferroni-Holm p-value adjustment.

	Feature	# sig		Feature	# sig
1	**word**	−3	8	modal verb	+3
2	**verb**	−3	9	certainty	+3
3	**sentence**	−3	10	**self ref**	−3
4	**sen_len**	−3	11	redundancy	−3
5	word_len	−3	12	**SMOG**	+4
6	**pausality**	−4	13	ARI	−3
7	**modifier**	−3			

Table A10: Table of n-grams that show a consistent significant difference in *Occ* in *# sig* datasets. A positive *# sig* means the n-grams is generally more frequent in deceptive texts. Emboldened unigrams are still qualified after the Bonferroni-Holm correction.

	n-gram	# sig		n-gram	# sig
1	**I**	+3	11	their	+3
2	**they**	+3	12	**if**	+3
3	out	+3	13	both	+3
4	is a	−3	14	at the	+3
5	**only**	+3	15	**will**	−3
6	do	+4	16	**me**	+3
7	**at**	+3	17	**but**	+4
8	about	+3	18	than	+3
9	**that**	+3	19	**and**	−3
10	**them**	+3	20	for	−3

We then calculate the frequency of occurrence of every n-gram x in each text t using the formula

$$Occ_n(x,t) = \sum_{s \in t} \frac{\#_x(s)}{|s| - n}$$

where $|s|$ denotes the number of words in sentence s of t, and $\#_x(s)$ denotes the number of times x occurs in s.

We then ranked the function word n-grams by the difference of the aggregate occurrences in the two classes (truthful versus deceptive) and selected the top 100 for statistical significance testing. We then ran two-sample t-tests on each dataset, comparing occurrence scores between legitimate and deceptive texts, and identified ones that showed a consistent significant difference, both with and without Bonferroni-Holm correction.

B.2 Results

The results are in Table A10. Only two n-grams, "do" and "but" show significant differences between truthful/deceptive groups in all four datasets. Both unigrams ($n = 1$) are more frequent in deceptive texts. Twenty n-grams, including "I", "they," "is a," and "at the," show significant differences between truthful and deceptive texts in three out of four datasets. After Bonferroni-Holm correction, 11 n-grams still qualified.

An individual dataset has thousands of FW n-grams, of which between 100 to 400 of them are significant discriminators of deceptive texts. In our experiments, 20 terms, 18 of them unigrams, show a common

behavior across at least three out of the four datasets, and four across all four. After Bonferroni-Holm correction, 11 out of 20 are still qualified. We also notice that the lower the n, the higher the chance that the n-gram will show a significant difference between truthful and deceptive groups.

C LIWC and BERTAA Features

The 55 stylistic features from Fabien et al. [2020] are listed below:

- Length of text: len-text
- Number of words: len-words
- Average length of words: avg-len
- Number of short words: num-short-w
- Proportion of digits and capital letters: per-digit, per-cap
- Individual letters and digits frequencies: f-a, f-b, f-c, f-d, f-e, f-f, f-g, f-h, f-i, f-j, f-k, f-l, f-m, f-n, f-o, f-p, f-q, f-r,f-s, f-t, f-u, f-v, f-w, f-x, f-y, f-z, f-0, f-1, f-2, f-3, f-4, f-5, f-6, f-7, f-8, f-9
- Hapax-legomena: richness
- Frequency of 12 punctuation marks: f-e-0, f-e-1, f-e-2, f-e-3, f-e-4, f-e-5, f-e-6, f-e-7, f-e-8, f-e-9, f-e-10, f-e-11

LIWC denotes the Linguistic Inquiry and Word Count program. It uses dictionaries of words that fit into different categories like tone_pos (positive tone) to identify words in the text that fall into that category (happy, elated, excited, ...) and reports *percentage of words in the text* that fall into those categories except for WC (word count) and WPS (words per sentence). More information about LIWC can be found at `https://www.liwc.app/help`.

The 86 LIWC features that we used along with their abbreviations are as follows:

- Summary Variables (8 features): Word count – WC, Analytical thinking – analytic, Clout – clout, Authentic – authentic, Emotional tone – tone, Words per sentence – WPS, Big words – BigWords, Dictionary words – Dic
- Linguistic Dimension (17 features): Total function words – function, Total pronouns – pronoun, Personal pronouns – ppron, 1st person singular – I, 1st person plural – we, 3rd person plural – they, Impersonal pronouns – ipron, Determiners – det, Articles – article, Numbers – number, Prepositions – prep, Auxiliary verbs – auxver, Adverbs – adverb, Conjunctions – conj, Negations – negate, Common verbs – verb, Quantities – quantity
- Psychological Processes (28 features): Drives – drives, Affiliation – affiliation, Power – power, Cognition – cognition, All-or-none – allnone, Cognitive processes – cogproc, Insight – insight, Causation – cause, Discrepancy – discrep, Tentative – tentat, Certitude – certitude, Differentiation – differ, Affect – affect, Positive tone – tone_pos, Negative tone – tone_neg, Emotion – emotion, Positive emotion – emo_pos, Negative emotion – emo_neg, Anxiety – emo_anx, Sadness – emo_sad, Social processes – Social, Social behavior – socbehav, Prosocial behavior – prosocial, Politeness – polite, Moralization – moral, Communication – comm, Social referents – socrefs, Friends – friend
- Punctuation (5 features): All Punctuation – allpunc, Apostrophes – apostro, Periods – period, Commas – comma, Other punctuation – OtherP

- Expanded Dictionary (28 features): Culture – culture, Politics – politic, Ethnicity – ethnicity, Technology – tech, Lifestyle – lifestyle, Home – home, Work – work, Religion – relig, Physical – physical, Health – health, Mental health – mental, Need – need, Lack – lack, Fulfilled – fulfill, Risk – risk, Curiosity – curiosity, Allure – allure, Perception – perception, Attention – attention, Motion – motion, Space – space, Visual – visual, Auditory – auditory, Feeling – feeling, Time – time, Past focus – focuspast, Present focus – focuspresent, Future focus – focusfuture

Note that a few of the features between these three lists: BERTAA, the 13 selected features from Table A9 and LIWC list are duplicates. Specifically, WD from LIWC, words in Table A9, and len-text from BERTAA are duplicative, and WPS from LIWC, average sentence length in Table A9 form another duplicate group. These are removed by the colinearity check mentioned in the feature analysis section, Section 8.