Gröndahl, Tommi; Asokan, N.

Text analysis in adversarial settings

Published in:
ACM Computing Surveys

DOI:
10.1145/3310331

Published: 01/06/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Gröndahl, T., & Asokan, N. (2019). Text analysis in adversarial settings: Does deception leave a stylistic trace? *ACM Computing Surveys*, *52*(3), 1-36. Article 45. https://doi.org/10.1145/3310331

Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace?

TOMMI GRÖNDAHL and N. ASOKAN

Textual deception constitutes a major problem for online security. Many studies have argued that deceptiveness leaves traces in writing style, which could be detected using text classification techniques. By conducting an extensive literature review of existing empirical work, we demonstrate that while certain linguistic features have been indicative of deception in certain corpora, they fail to generalize across divergent semantic domains. We suggest that deceptiveness as such leaves no *content-invariant stylistic trace*, and textual similarity measures provide superior means of classifying texts as potentially deceptive. Additionally, we discuss forms of deception beyond semantic content, focusing on hiding author identity by *writing style obfuscation*. Surveying the literature on both author identification and obfuscation techniques, we conclude that current style transformation methods fail to achieve reliable obfuscation while simultaneously ensuring semantic faithfulness to the original text. We propose that future work in style transformation should pay particular attention to disallowing semantically drastic changes.

1 INTRODUCTION

Deception is rampant in online text, and its detection constitutes a major challenge at the crossroads of natural language processing (NLP) and information security research. Multiple studies have contended that leading machine learning techniques are able to extract features that can distinguish between deceptive and normal text. In order for such features to truly reflect deceptiveness instead of domain-specific lexical content, the features discovered should generalize across domains. Variants of deception also extend beyond textual content. In particular, metalinguistic information can be obfuscated to deceive a classifier while retaining the original content. Of such endeavours, the most prominently discussed has been *adversarial stylometry* [14, 15], consisting of techniques that attempt to provide author anonymity by defeating identification or profiling. In this survey we review prior research on textual deception and its detection, focusing on deceptive content in Section 2 and adversarial stylometry in Section 3.

Modern NLP techniques offer a large variety of methods for classifying texts based on the distribution of *linguistic information*: features that are detectable from *text alone*, without extra-linguistic knowledge concerning author behavior, metadata etc. Depending on the task, target categories can be delineated by semantic content, grammar, or any combination of these. Identifying or profiling authors based on writing style comprises the field of *stylometry*. As a scientific endeavour it dates back at least to the 19th century [104, 112], and was formulated as a computational task in the 1960s [118, 163]. In contemporary work, the traditional focus on literary documents has largely been overshadowed by the increased use of online datasets, such as blog posts [121], e-mails [32, 37], forum discussions [183], SMS messages [138], and tweets [25]. Neal et al. [123] comprehensively survey the state-of-the-art in stylometry.

Stylometry uses linguistic information to extract a *non-linguistic* property of the author of a text, such as identity, gender or age. Within NLP, it thus belongs to the field of *metaknowledge extraction* [33], which relies on linguistic information systematically correlating with the relevant property under investigation, despite that property itself not

2019. Manuscript submitted to ACM

being linguistic. The conjecture that author identity can be reliably inferred from his/her stylistic "fingerprint" is known as the *Human Stylome Hypothesis* (HSH) [168].

Motivated by the HSH, we can also formulate an analogous question about any other property of a text: does the property leave a *linguistic trace*, and if so, to what extent does it leave a content-independent *stylistic trace* that could be recognized across semantically distant texts? In this paper we discuss this question with respect to a class of properties that fall under the umbrella term of *deception*. We investigate the issue both from the perspective of detecting deception in text, and from the *adversarial* perspective of creating deceptive data that can evade classification. Our focus is on information security applications in particular. Terminologically, we call non-deceptive text "normal".

If reliable linguistic cues of deception existed, they could be used to detect security breaches such as fake reviews [98, 128, 129, 177, 179], troll-messages [23, 115, 153], or even fake news [126, 132]. A number of prior studies have attempted to demonstrate the potential of stylometry for deception detection, and to find the major linguistic determinants of deceptive text. We review and discuss this research in Section 2. A common assumption behind all these studies is that deception leaves a stylistic trace comparable to an author's "stylome". If true, this would allow detecting deceptiveness *from the text alone*, without recourse to extra-linguistic information. If, on the other hand, deception leaves no major stylistic trace, its reliable detection would require linguistic analyses to be augmented with other techniques. Based on the survey, we argue for the latter position, and suggest alternative methods based on *content-comparison* that provide more promising approaches for this task (Section 2.3).

In Section 3 we turn to adversarial stylometry. From a security perspective, it simultaneously functions as an *attack* against authorship classification, and as a *defence* against non-consensual deanonymization or profiling. The latter scenario has been called the *deanonymization attack* [121], and its feasibility is conditional on the HSH. Therefore a major question is whether current author identification techniques pose a realistic privacy threat. Based on a review of state-of-the-art stylometry research in Section 3.1, we argue that while the HSH has not always been validated, the deanonymization attack constitutes a genuine privacy concern especially when the candidate authors are few in number. In Section 3.2 we discuss the attack scenario in more detail.

Methods for style transformation can be divided into manual, computer-assisted and automatic techniques. For *ordinary* users, only the last would constitute a practically effective mitigation against the deanonymization attack. Manual obfuscation is difficult and time-consuming, and requires a good grasp of linguistic subtleties, which makes the task unsuitable for users lacking extra time and resources. An additional difference can be made between *obfuscation* and *imitation*, where the latter targets a particular style instead of simply avoiding detection. Section 3.3 reviews existing work on style obfuscation and imitation techniques. We argue that while some methods show potential in principle, all face serious problems with balancing between obfuscation and maintaining semantic content.

Given that style obfuscation and imitation constitute types of deception, we then return to the original question of whether deception leaves a stylistic trace, and apply it to this special case. Even if obfuscation was successful, the property of *being obfuscated* could itself be stylometrically traceable. Problematically, our review in Section 3.3.3 demonstrates that this question has typically not been tested. Studies attempting such "fingerprinting" of the obfuscation method have succeeded [22, 36], but have only experimented on a small subset of possible methods. As different techniques require different recognition methods, a general detector of style obfuscation is likely difficult to attain.

In summary, this survey addresses three major questions:

Q1 Does deception leave a content-independent stylistic trace?
Q2 Is the deanonymization attack a realistic privacy concern?

Q3 Can the deanonymization attack be mitigated with automatic style obfuscation?

Q1 provides the common theme of the survey. Section 2 discusses the linguistic detection of *deceptive content*, with a particular focus on online text. Section 3 then moves on to the topic of adversarial stylometry, i.e. mitigating the deanonymization attack (Q2) via style transformation (Q3).

We summarize our findings and suggestions below.

- There is no evidence that deception leaves a *content-invariant* stylistic trace. Instead, detection should involve the *comparison of semantic content* across texts.
- While the validity of the HSH is uncertain, the deanonymization attack is a realistic privacy concern.
- As of yet, automatic style transformation techniques do not secure semantic faithfulness.

2 DECEPTION DETECTION VIA TEXT ANALYSIS

In this section we review the research on textual deception detection, and discuss the linguistic features associated with writing intended to deceive the reader. Multiple studies have indicated that at least non-expert human accuracy in detecting textual deception is approximately on a chance level, or even worse [13, 46, 47, 124]. As Fitzpatrick et al. [52] note, this makes deception detection a somewhat exceptional topic for NLP, since human performance in most other text classification problems tends to be more accurate than computational solutions. In contrast, automated classification of deceptive text should increase not only the efficiency but also the accuracy of human performance. However, we argue that the divergence of features deemed relevant by different studies indicates that classification has been too content-specific to generalize across semantic domains. Relevant features tend to be lexical correlates of deceptive text in particular corpora rather than general "deception markers" as such.

Most research in deception detection has concerned face-to-face discussion [39, 44, 46, 47, 51, 52]. As Crabb [32] notes, such results do not always directly apply to communication via electronic devices, which are the most relevant for information security concerns. In particular, physiological data is unavailable to the receiver in text-based communication. We limit our discussion to deceptive communication in written English. Following DePaulo et al. [39], we dissociate deceptiveness as a *communicative intention* from falsity as a semantic property.[1] Utilizing a famous formulation by Paul Grice, *communication* can roughly be characterized as behavior with the deliberate goal of causing certain thoughts in the (intended) receiver [60, 162]. Deception thus constitutes a specific type of communication, where the speaker intends the hearer to form thoughts which the speaker believes to be false. The notion of deception as an author intention is also shared by Buller and Burgoon's *Interpersonal Deception Theory* [17]. For our purposes, we can use the following general characterization:

Deception

A deceives B if for some proposition P:

A believes that P is false

A attempts to make B believe that P is true

The deceptiveness of a communicative act makes no restrictions on the nature of the proposition P. In particular, P might not belong to the semantic content of the expression E. We can thus separate between *explicit* and *implicit deception* as follows:

[1] Literal truths with a deceptive intention include cases where the speaker believes the hearer to *infer* a falsity from a literal truth. If Bob asks: "Where is Jim?", and Alice answers: "I saw him in the cafeteria", in normal circumstances Alice assumes Bob to infer that Jim may still be there. Hence, if Alice believes Jim not to be there anymore (e.g. she also saw him leave the cafeteria), she is deceiving Bob by telling a literal truth. Assumed inferences can be understood as belonging to communicated content as *implicatures* [162].

Explicit deception

A explicitly deceives B if for some proposition P:

A believes that P is false

A attempts to make B believe that P is true by uttering an expression E

The semantic content of E contains P

Implicit deception

A implicitly deceives B if for some proposition P:

A believes that P is false

A attempts to make B believe that P is true by uttering an expression E

The semantic content of E does not contain P

A assumes B to infer P from the explicit content of E and context information that A assumes B to know

In both cases, B infers P from A's utterance E. In explicit deception, P can be found directly from E itself without consulting other assumptions or beliefs within the discourse. In implicit deception, further inferences are needed to come to the conclusion P. As an example, consider fake online reviews, which Section 2.2.1 will discuss in detail. Some fake reviews contain explicit falsities: if a TripAdvisor user claims to have been in a hotel and (dis)liked it, this is explicitly deceptive if he actually has not visited the hotel. However, suppose the reviewer merely makes general claims about the hotel ("This hotel is excellent/horrible!" etc.). Here, the deception concerns the reviewer's first-hand experience, which he lacks, and is independent of the reviewer's actual beliefs of his review's correctness. Therefore, it can be treated as a variant of implicit deception.

We further divide different types of deception reviewed in Sections 2.1–2.2 to the following three groups, the first being explicit and the two latter implicit:

Deception of literal content: the semantic content of the text is deceptive

Deception of authority: the deceiver implies having authority concerning the issue, which he lacks

Deception of intention: the deceiver has an ulterior deceptive motive for writing the message

While not meant to be exhaustive, this taxonomy is useful in accounting for disparities between different studies. Most studies reviewed concern deception of literal content. However, as argued above, fake reviews can exhibit deception of *authority* instead. Deception of intention is exemplified by *trolling*, where the author writes something to advance a particular view or to harass another person. Here, deception does not necessarily concern the literal content (which may sometimes be sincerely believed by the troll), but instead the ulterior motive behind the message.

Ultimately, the issue at hand is whether it would be possible to develop a "textual lie detector" that takes a text as an input and outputs a classification label that reliably tracks the real-world property of deceptiveness. The main problem for such a goal is that even if deceptive texts differ from non-deceptive texts in particular corpora, the features may not generalize across different text types. Deception could leave *some* stylistic cues in e.g. in online discussions, fake news, fake reviews, or scientific papers; but in order for the hypothetical "lie detector" to work, these cues should be sufficiently *similar* across them all. To evaluate whether existing methods are applicable for such general deception detection, we need to compare empirical results from different studies and see if common patterns emerge.

Table 1 shows the linguistic properties that appear three or more times in the studies reviewed in Sections 2.1-2.2. Based on these, we formulate the following hypotheses:

H1: deceptive text is emotionally laden

H2: deceptive text contains certainty-related terminology

Studies	Cue
[19, 64, 80, 98, 124, 128, 129, 184, 185]	High emotional load
[19, 64, 100, 128, 129, 177, 184]	Generality / abstractness / lack of specificity
[95, 98, 100, 128, 179]	High use of first-person pronouns
[63, 80, 113, 124]	Low use of first-person pronouns
[32, 128, 129, 186]	High use of verbs
[64, 95, 113]	High use of certainty-related words

Table 1. The most common linguistic cues to deception from all studies reviewed in Sections 2.1-2.2

H3: deceptive text lacks in detail

H4: deceptive text lacks a first-person narrative

Table 1 contains two contradictory properties: high and low use of first person pronouns. We choose the low use hypothesis as the default (H4), since it would be predicted by the lack of first-hand experience of the situation. This empirical divergence is indicative of the context-dependency of suggested deception cues. Nevertheless, H1–H4 are intuitively understandable and fit well within standard psychological models of deception [17]. H1 can be explained either by the stress caused by lying [45], or from attempted emotional persuasion of the audience. Experimental results reviewed in Section 2.2.1 point to the latter [128, 129]. H2 also likely results from the persuasive purpose of deception. H3 and H4 are motivated by the fact that deceivers often have no first-hand experience of the situation they are describing.

H1–H4 can thus be argued to follow from two basic tendencies present in deception: *attempted persuasion* and *lack of first-hand knowledge*. Interestingly, these can sometimes motivate the deceiver to behave in *opposite* ways, which may partly explain our seemingly inconsistent finding concerning first-person pronoun usage. Increased use of the first-person pronoun indicates a personal narrative and hence emphasizes the notion of the author having actually experienced the situation under discussion. It can therefore be used in an attempt to increase the credibility of the text. On the other hand, the lack of first-hand experience makes it more difficult for deceivers to credibly describe something they do not know in detail, and hence can motivate them to stick to a more general, third-person narrative.

Section 2.1 reviews the literature on deception detection from a general perspective not specific to information security concerns. Section 2.2 focuses specifically on deception in online text, discussing fake reviews (2.2.1) and troll comments (2.2.2). We summarize our analyses and give recommendations for future research in Section 2.3.

2.1 General deception detection

In this section we focus on deception detection outside of online text datasets. We further distinguish between *experimentally elicited* deception and natural or *non-elicited* deception, devoting Section 2.1.1 to the former and Section 2.1.2 to the latter.

2.1.1 Experimentally elicited deception. In this section we present results from experimental research on elicited deceptive text. By *elicited* we mean that the texts were produced at the command of a test instructor, and that their deceptiveness or truthfulness was explicitly requested.

Burgoon et al. [19] formulated eight hypotheses concerning deception:

> "deceptive senders display higher
> (a) quantity,

(b) non-immediacy,
(c) expressiveness,
(d) informality, and
(e) affect;

and less

(f) complexity,
(g) diversity, and
(h) specificity of language" [19][2]

They based these hypotheses on two experiments, where truthful and deceptive text was gathered from participants. The messages were obtained via e-mail in the first experiment, and via either face-to-face communication or text chat in the second experiment. However, the results of these experiments were contradictory, as deceivers used longer but less complex messages in the first test, and shorter but more complex messages in the second (although the results from the second test were not statistically significant). The hypotheses (a–h) reflect the results of the first experiment. In relation to H1–H4, (e) indicates *emotional load* (H1) and (f–h) fall into the broader category of *lacking detail* (H3). However, the latter is partly at odds with (c), which indicates that deceivers also use higher amounts of descriptive words. This effect could arise from the attempted persuasion involved in deception.

Based on nine linguistic properties similar to those suggested by Burgoon et al. [19] (each composed of many features, 27 altogether), Zhou et al. [184] classified experimentally elicited texts as deceptive or truthful. All features were relevant with the exception of specificity. In a subsequent study, Zhou et al. [185] report 22 linguistic features as indicative of deception (see Table 2). Using these features, they compared four machine learning methods in the classification task: discriminant analysis, decision trees, neural networks and logistic regression. The methods fared roughly equally well, and at best achieved an accuracy of ca. 80%.

Newman et al.'s [124] results on classifying experimentally elicited deceptive and truthful texts indicated that deception was characterized by the reduced use of first- and third-person pronouns and exclusive words (e.g. *but*, *except*), along with the increased use of negative emotion words (e.g. *hate*, *anger*) and motion words (e.g. *walk*, *go*). These findings are partly in line with more general observations concerning deception, but not fully. While emotional load (H1) and reduced first-person pronoun use (H4) are expected , it is unclear why deception should correlate with the reduction of *both* first- and third-person pronouns, a decrease in exclusive words, or an increase in motion words. Typically, first- and third-person pronoun usage can be seen as *complementary* ways to talk about a situation, the first-person indicating a personal narrative and the third-person an impersonal one. It is therefore unclear what properties the reduction in *both* would coincide with. Also, exclusion words are likely too abstract to be closely related to particular communicative functions, and hence it would be surprising if their prevalence in deceptive text were to generalize across different datasets. Finally, motion words generally denote concrete events, and therefore contradict the general finding of deception lacking in detail (H3). It is therefore relatively unsurprising that, aside of reduced first-person pronoun use and emotional load, Newman et al.'s [124] results do not resurface in other studies.

Based on a review of prior research, Hancock et al. [63] formulated seven hypotheses on linguistic deception cues:

"(a) Liars will produce more words during deceptive conversations than during truthful conversations.

[2] *Quantity* means the amount of text produced, *non-immediacy* refers to the lack of directness and intensity between the author and receiver, *expressiveness* is the amount of descriptive material in the text (e.g. adjectives and adverbs), *informality* is indicated e.g. by the amount of typos, *affect* refers to the emotional load of the text, *complexity* is measured by readability indices [160], *diversity* is the type-token ratio among words, and *specificity* denotes the level of detail in the text.

(b) Liars will ask more questions during deceptive conversations as compared to truthful conversations.
(c) Liars will use fewer first-person singular but more other-directed pronouns in deceptive conversations than in truthful conversations.
(d) Liars will use more negative emotion words during deceptive conversations than during truthful conversations.
(e) Liars will use fewer exclusive words and negation terms during deceptive conversations as compared to truthful conversations.
(f) Liars will avoid causation phrases during deceptive interactions relative to truthful interactions.
(g) Liars will use more sense terms during deceptive interactions as compared to truthful interactions." [63]

The test subjects were divided between *motivated* and *unmotivated* liars based on whether the experimenter had provided false information (later revoked) about the importance of the ability to lie for success in life. Some hypotheses received confirmation from all liars (a–c, g), some only from motivated liars (f), and others from neither (d, e). Hypothesis (c) is indicative of a more general property of deception: the lack of a personal narrative (H4). However, a contrasting result is provided by the confirmation of (g): the increase of sense-related terminology. Sensation indicates a personal narrative, making this result contrast with the more general finding that deception tends to correlate with the lack of first-hand knowledge and detail (H3).

Lee et al. [95] tested the ability of various linguistic features to predict deception in data from 30 deceptive and 30 truthful participants answering questions. While their initial hypothesis contained eight conglomerate properties, only one was statistically significant: *certainty*, as calculated with a five-feature proxy measure comprised of causation words (e.g. *because*, *hence*), insight words (e.g. *think*, *know*), certainty words (e.g. *always*, *never*), first-person singular pronouns, present-tense verbs, and tenacity verbs (e.g. *is*, *has*). All five predicted deception in a statistically significant manner. These results are partly contradictory with Hancock et al's [63], who found that (motivated) liars tended to avoid causation phrases. Further, the increase of first-person pronouns contrasts with many other studies, where their high use has correlated negatively with deception (H4).

Mihalcea and Strapparava [113] classified truthful and deceptive opinions concerning political and personal issues (abortion, capital punishment, and the responder's best friend) gathered via Amazon Mechanical Turk. At best they achieved a 70% accuracy with a Naïve Bayes classifier. They report a decrease of self-related words and an increase of certainty-related words as indicative of deception. Both results are in line with general findings of deception typically instantiating attempted persuasion (H1–H2) and a lack of first-hand experience (H3–H4).

In summary, the studies reviewed in this section suggest certain common features of experimentally elicited deception, but also include some unclear and even contradictory results. The results are collected in Table 2. The table additionally shows which of the hypotheses H1–H4 receive support or are contradicted by the findings. A general trend is visible: deceivers often try to artificially emphasize what they say by using emotional and certainty-related terminology (H1–H2), while not providing detailed information about the topic they address (H3). However, contradictory results exist especially with respect to features related to the first-person narrative (H4). Many studies also support some of the hypotheses but oppose others. For instance, in Hancock et al.'s [63] data deception correlated both with reduced first-person pronoun usage and increased sense-terminology. The first of these features supports H4, but the latter points to the opposite direction, as sense-terminology often relates to descriptions of first-hand encounters. A similar case is found in Newman et al. [124], who detected both reduced first-person pronoun usage and increased motion-word usage as indicators of deception.

Study	Test setting	Deception cues	Support	Oppose
[19] [184]	A theft-based game [19]; a variant of the Desert Survival Problem [92, 184]	quantity, **reduced immediacy**, expressiveness, informality, **affect**, reduced complexity, reduced diversity, **reduced specificity**	H1, H3	
[185]	Two variants of the Desert Survival Problem [92]	verbs, modifiers, word length, punctuation, modal verbs, individual reference, group reference, **emotiveness**, content diversity, redundancy, *perceptual information*, *spatiotemporal information*, errors, **affect**, imagery, pleasantness, positive activation, positive imagery, negative activation	H1	H3
[124]	Reported views about abortion, friendship, and a mock crime scenario.	**reduced first person pronouns**, *reduced third person pronouns*, reduced exclusive words, **negative emotion words**, *motion words*	H1, H4	H4
[63]	Conversations between two participants	quantity, questions, **reduced first person singular pronouns**, **other-directed pronouns**, *sense terms*	H4	H4
[95]	A questioner-responder game	causation words, **insight words**, **certainty words**, *first-person singular pronouns*, present-tense verbs, tenacity verbs	H2	H4
[113]	Reported views about abortion, capital punishment, and friendship	**reduced self-related words**, **certainty-related words**	H2, H4	

Table 2. Linguistic cues of experimentally elicited deception
Bold: support H1–H4
Italics: do not support H1–H4

2.1.2 Non-elicited deception. We now move on to deception in texts which have not explicitly been requested to be deceptive. These include both real-world corpora, as well as texts produced in experimental conditions where deceptiveness was not asked but was later evaluated based on independent criteria.

A common dataset for real-life deception has been the Enron e-mail corpus [85].[3] Keila and Skillicorn [80] detected deceptive text from the Enron corpus, using features drawn from Zhou et al. [187] and Newman et al. [124] on the linguistic cues of deception: reduced use of first and third person pronouns and exclusive words, and increased use of negative emotion words and motion words. As discussed in Section 2.1.1, assuming these features to always indicate deception is not unproblematic. Further, while Keila and Skillcorn's manual evaluation indicated that the e-mails ranked high by these properties contained deceptive e-mails, the lack of ground truth makes it impossible to properly evaluate their results. Keila and Skillicorn additionally note that not only deception but other "marked" types of communication were also indicated by these features, such as otherwise inappropriate messages.

Louwerse et al. [100] predicted fraud in the Enron corpus using a five-point abstractness scale based on prior work by Semin and Fiedler [154], who classified verbs and adjectives on the following scale, (a) being the most concrete and (e) the most abstract (examples from Semin and Fiedler [154]):

(a) Descriptive Action Verbs: *hit*, *yell*, *walk*
(b) Interpretative Action Verbs: *help*, *tease*, *avoid*

[3] http://www.cs.cmu.edu/~enron/

(c) State Action Verbs: *surprise*, *amaze*, *anger*
(d) State Verbs: *trust*, *understand*
(e) Adjectives: *distraught*, *optimal*

Louwerse et al. [100] further divided adjectives to four analogical classes based on (a)–(d). Using Semin and Fiedler's [154] assessment that abstractness indicates low verifiability and low informativity, they predicted that high abstractness would correlate with deception. The email database was divided into sixteen events based on sending times, some of which were highly correlated with deception taking place within the Enron corporation. Regression analysis demonstrated that these events correlated with linguistic cues of high abstractness, providing support for the hypothesis.

Additionally, based on the results of Newman et al. [124] and Hancock et al. [63] (see Section 2.1.1), Louwerse et al. [100] further investigated the correlation of deceptive events in the Enron corpus with first and third person pronouns, causal adverbs, negation, the connective "but", and email length. Of these, first person pronouns and negations were partially indicative of deceptive events, but the results were *contrary* to the prior studies [63, 124], as first person pronouns were used *more* in deceptive emails rather than less.

Larcker and Zakolyukina [93] studied linguistic properties of fraudulent and truthful financial statements by Chief Executive Officers (CEOs) and Chief Financial Officers (CFOs) in conference calls. Their results diverged significantly between CEOs and CFOs. Differences were found in e.g. negations and extremely negative emotion words, which correlated positively with deception for CFOs but not CEOs. Some cues were even contrastive, as deception correlated with certainty-related words among CFOs, but hesitation-related words among CEOs. One possible reason for these differences could be that the features reflect the personal style of the CEOs/CFOs themselves rather than their deceptiveness. However, some commonalities were found: deceptive CEOs and CFOs both used more general group references, less non-extreme positive emotion terms and less third-person plural pronouns. While the prevalence of general group references indicates distance and thus supports H4, the other indicators seem particular to this study, as they are not replicated in other studies. They also bear no clear relation to H1–H4.

Toma and Hancock [166] compared the linguistic properties of fraudulent and truthful online dating profiles. While the profiles were written in experimental settings, deception was not encouraged, and was only detected by comparing the profiles to ground-truth gathered about the users. Deception correlated significantly with reduced first-person singular pronouns, increased negations, a lower word count, and a decrease in negative emotion vocabulary. While the last feature stands in opposition to many other studies [80, 124, 187], it is unsurprising considering the context: a deceptive dating-profile would most likely exaggerate positive qualities and downplay negative ones. Hence, it is unlikely that this result would generalize across different text types.

Crabb [32] used POS-tags and lexical diversity for deception detection from the Enron corpus. She used two methods: clustering with the Expectation-Maximum algorithm, and calculating means for each feature in isolation to detect statistically significant differences with respect to deception-cues identified in prior research [2, 50, 63, 80, 95, 100, 186, 187]. Two clusters were deemed most relevant due to the high occurrences of modal, base and present tense verbs, second-person pronouns, and function words. However, while emails in these clusters generally had higher values for such features than those in other clusters, not all such values were statistically significant. Further, the lack of ground truth in the Enron corpus prevented any conclusive inferences to be made concerning the prevalence of deception in the clusters.

Studies	Data	Deception cues	Support	Oppose
[100]	Enron e-mails [85]	**abstractness**, negations, *first person pronouns*	H3	H4
[93]	Conference call transcripts	**general group references**, reduced non-extreme positive emotion terms, reduced third-person plural pronouns	H3	
[166]	Online dating profiles	**reduced first-person singular pronouns**, negations, reduced word count, *reduced negative emotion words*	H4	H1
[64]	Fraudulent scientific papers	words related to scientific methodology, amplifying terms, **certainty-related words**, **emotional words**, reduced diminisher terms, reduced adjectives	H1, H2	
[32]	Enron e-mails [85]	modal, base and present tense verbs, **second-person pronouns**, function words	H4	

Table 3. Linguistic cues of non-elicited deception
Bold: support H1–H4
Italics: do not support H1–H4

Hancock and Markowitz [64] used linguistic information to classify papers by the social psychologist Diederik Stapel, who famously fabricated data to many publications. They observed the following tendencies in Stapel's fraudulent papers in comparison to truthful ones:

- more terms related to scientific methodology
- more amplifying terms (e.g. *extreme*, *exceptionally*, *vastly*)
- more certainty-related terminology
- more emotional terminology
- fewer diminisher terms (e.g. *somewhat*, *partly*, *slightly*)
- fewer adjectives

Hancock and Markowitz' [64] results thus provide support for the hypotheses that deceivers generally exaggerate the content they want the receiver to believe (H1) and their level of certainty (H2), while providing less qualitative descriptions (H3). Their model correctly classified 71% of Stapel's papers. While this was significantly better than random choice, the authors express caution about the feasibility of their method for broader forensic use, citing the large error rate and the domain-specificity of scientific discourse.

Studies on non-elicited deception are summarized in Table 3. The results are mostly in line with experimental research (Table 2): common features include high emotional load (H1), certainty-related terminology (H2), abstractness (H3), and the reduced use of first-person pronouns (H4). However, as in experimental studies, the evidence is contradictory concerning emotional words and first-person pronouns.

2.2 Deception detection from online text

In this section we focus on two specific topics relevant for online security: *fake reviews* and *troll comments*. We argue that both present unique properties not inferrable from the results reviewed in Section 2.1. We further discuss alternative methods for their detection, and evaluate the importance of pure text analysis as a tool for these tasks.

2.2.1 Fake reviews. One major source of deceptive online text is *fake reviewing*, where the reviewer deliberately attempts to (mis)lead the audience into believing something about a product [146, 169]. Fake reviews may have special properties in comparison to other forms of deception, and are therefore allocated a separate section in this survey. As Yoo and Gretzel [179] point out, fake reviewers are often professionals, and can typically model their writing on real reviews. Additionally, a fake review does not need to be fraudulent with respect to the author's actual opinions. Instead, its deceptiveness stems from the *purpose* of the author to spam a site for some ulterior reason instead of providing informative reviews. Hence, fake reviews are not necessarily disbelieved by the author, but the content is irrelevant to the author's true goal: they instantiate *deception of intention*.

For supervised methods, obtaining labeled data constitutes a major challenge, and studies have typically collected their own data. The largest corpus of elicited fake reviews has been compiled by Ott et al. and contains 400 fake and 400 truthful reviews of both the positive [128] and negative [129] kind.[4] Additionally, the website Yelp provides a corpus of filtered reviews suspected to be fake.[5]

Human written reviews Ott et al. [128] detected deceptive pieces in TripAdvisor hotel review data generated via Amazon Mechanical Turk. Combining psycholinguistic features from the LIWC software [131] and word bigrams, they achieved an accuracy of 89.8%, and summarize their results as follows:

> "(...) truthful opinions tend to include more sensorial and concrete language than deceptive opinions;
> (...) we observe an increased focus in deceptive opinions on aspects external to the hotel being reviewed
> (...) our deceptive reviews have more positive and fewer negative emotion terms.
> (...) we find increased first person singular to be among the largest indicators of deception" [128]

These findings stand in stark contrast to H4, since here deception is indicated by an *increase* in first-person pronouns and hence a more personal narrative. This trend turns out to be prevalent in fake reviews, providing support for Yoo and Gretzel's [179] contention that fake reviews differ from other forms of deception. On the other hand, Ott et al. also found that fake reviews were more abstract and less specific, in line with H3 and against Yoo and Gretzel's analysis of fake reviews having a special status due to the availability of information.

Feng et al. [50] further improved Ott et al.'s [128] results by adding syntactic phrase structure to the stylometric evaluation of the same dataset, reaching 91.2% accuracy. As features they used both word bigrams and abstract syntactic relations derived from a context-free grammar parse.

Ott et al.'s first study [128] was concerned with *positive* hotel reviews. In a subsequent study [129], the same authors applied the method to *negative* reviews, also gathered via Amazon Mechanical Turk. They achieved a F-score of c.a. 86% with n-gram-based support vector machines (SVMs). Negative fake reviews contained less spatial information and had a larger verb-to-noun ratio than truthful reviews. They also manifested an excess of negative emotion terms, in direct contrast with the high use of positive terms in the prior study. Ott et al. interpret these results as opposing the hypothesis that negative words indicate the emotional distress involved in lying [45]. Rather, the increased use of emotional terminology can be explained by the intention of the deceiver to communicate certain contents, which is why the prevalent emotions will vary along with these intentions. High emotional load may still be a useful deception cue, but it results from a more general property of *emphasis*, and is not ubiquituously negative.

Yoo and Gretzel [179] tested seven hypotheses on the linguistic properties of fake hotel reviews:

> "(a) Deceptive reviews contain more words.

[4]The corpus is available at http://myleott.com.
[5]https://www.yelp.com/dataset

(b) Deceptive reviews are less complex.
(c) Deceptive reviews are less diverse.
(d) Deceptive reviews contain less self-references (immediacy).
(e) Deceptive reviews contain a greater number of references to the hotel brand.
(f) Deceptive reviews contain a greater percentage of positive words.
(g) Deceptive reviews contain a smaller percentage of negative words."

Hypotheses (e–g) were confirmed, while (a–d) were not. In fact, the opposite hypotheses to (b) and (d) received support: fake reviewers used more complex language and more self-references than truthful reviewers. These results imply that fake reviews may differ from other types of deception by often being conducted by experts. Further, based on Ott et al.'s results [128, 129], it seems likely that (f–g)'s success was due to the reviews' promotional nature, and would plausibly not be replicated on negative review data.

Hue et al. [68] base their analysis of deceptive reviews on two properties: *sentiment* and *readability*. Sentiment is relevant since fraudulent reviewers likely have the intention of slanting the review either in favour of or against the product. Hue et al. further argue that in addition to sentiment varying randomly across different reviews by a genuine author, the same should be true of *readability*, measurable by e.g. the Automated Readability Index (ARI) based on the amount of characters within words and the amount of words within sentences [160]. In contrast, they maintain that readability should remain high and consistent across fraudulent reviews, since these aim at a maximally general audience. Using the Wald-Wolfowitz Runs test to detect non-randomness in manually labelled data from Amazon reviews, they provide empirical confirmation for constancy in both sentiment and readability as indicators of fake reviews.

Li et al. [98] studied linguistic generalities across fake reviews, which they divided between expert-generated and crowdsourced spam. They note that the commmon assumption of fake reviews lacking in detail [97, 128] is not true of expert-generated reviews. For crowdsourced reviews, their results accorded with previous studies indicating that fake reviews are less specific, and thus contain less descriptive terms like nouns or adjectives [10, 17, 18, 145]. This, however, was not the case for expert-generated fake reviews, which were highly informative and descriptive. Other linguistic cues Li et al. discovered were exaggerated sentiment and the overuse of first person singular pronouns. The latter result was contradictory to many previous studies proclaiming that deceivers avoid talking about themselves [18, 86, 124, 185].

Xu et al. [177] based their unsupervised fake review classifier on the text's *generality*, i.e. lack of informativity. The model ranked reviews based on "spamicity", the top reviews being most spam-like. Based on Ott et al.'s claim that online review sites typically contain $8\% - 15\%$ spam, they tested their model by treating the top k % as spam, where the value of k was varied between 5%, 10% and 15%. Accuracy was tested by comparing the top k % to its supervised classification by SVMs [26]. Applying the model to three review datasets, Xu et al. obtained F-scores of $75.2\% - 78.8\%$ with $k = 5\%$, $72.2\% - 76.6\%$ with $k = 10\%$, and $69.4\% - 71.7\%$ with $k = 15\%$. As they note, their method only works for reviews for products that are unavailable for the fake reviewer to investigate, such as restaurants or hotels. The assumption of fake reviews lacking specificity does not hold for products of which much information is available via commercials or other descriptions, since the reviewer could use such information in constructing the spam [179].

The results from fake review studies are summarized in Table 4. A recurring theme is the lack of specificity, but this depends on the assumption that the reviewer does not access information about the product [177]. High or low sentiment has also been demonstrated to be relevant as in other forms of deception, but its direction depends on the

Studies	Data	Deception cues	Support	Oppose
[179]	Hotel reviews (positive, experimentally elicited)	high complexity, *first person pronouns*, brand names, positive words, decreased negative words		H4
[128]	Hotel reviews (positive, crowdsourced)	**reduced specificity**, **external information**, positive sentiment, reduced negative sentiment, *first person singular pronouns*, high verb-to-noun ratio	H3	H4
[129]	Hotel reviews (negative, crowdsourced)	(in negative reviews:) **reduced specificity**, negative emotion terms, high verb-to-noun ratio	H3	
[68]	Amazon.com reviews	high readability, constancy of sentiment		
[98]	Hotel, restaurant, and doctor reviews (crowdsourced)	unspecificity (non-expert reviews), *specificity* (expert reviews), **exaggerated sentiment**, *first person singular pronoun*	H1	H3, H4
[177]	Amazon audioCD, TripAdvisor (hotels), Yelp (restaurants)	**text generality**	H3	

Table 4. Linguistic properties of fake reviews
Bold: support H1–H4
Italics: do not support H1–H4

nature of the review (positive or negative). Increased use of the first person pronoun stands in contrast to results received on other forms of deception (see Section 2.1), supporting Yoo and Gretzel's [179] contention about fake reviews constituting a *sui generis* type of deception. Yoo and Gretzel's analysis of fake reviews being special due to the amount of detailed information available receives partial support from Li et al. [98], but only for expert-generated reviews.

Automatically generated reviews *Automatic text generation* is a vast field within NLP [34], and poses an additional threat to review sites. Detecting automatically generated reviews is a different task than detecting man-made fake reviews, due to the different nature of the deception. In automatically generated reviews, the deception concerns *identity*: the message is meant to look like it is written by a human, but is in fact machine-generated.

Hovy [66] automatically generated fake reviews using a 7-gram Markov chain trained with data from the review site Trustpilot. For classification, he used logistic regression with word n-grams ($1 \leq n \leq 4$) as features. The classifier additionally sought irregularities between age, gender, review category, and n-grams. Adding such meta-information to the model significantly improved its ability to fool the classifier. However, while exact copies of training reviews were removed, a 7-gram model will likely reproduce large chunks of the training data. Duplicate or similarity detection between the training data and the generated reviews was not conducted by Hovy.

Yao et al. [178] generated fake reviews with a character-level Recurrent Neural Network (RNN) trained with restaurant reviews from Yelp.[6] They were unable to distinguish RNN-generated reviews reliably from those in the Yelp corpus, using linear SVMs with various linguistic features, the plagiarism detection method Winnowing [151], or human evaluators from Amazon Mechanical Turk ($n = 594$). These results demonstrate that machine-generated fake reviews can resist classification by common methods. However, Yao et al. suggest an alternative defence against their RNN-generated fake reviews, based on statistical differences in character distributions between generated reviews and the training corpus.

[6]https://www.yelp.com/dataset

Juuti et al. [77] utilized Neural Machine Translation (NMT) to generate context-appropriate restaurant reviews. They demonstrated a superior performance to Yao et al. [178] in fooling human users trying to distinguish between genuine and generated reviews. In their user study, Juuti et al. were able to avoid detection at a rate of 3.5/4, as opposed to 0.8/4 with Yao et al.'s method. By controlling the context (e.g. restaurant name, type of food, review rating etc.) they can further generate reviews of a specific type with a single NMT model. Despite successfully deceiving human readers, they were able to detect generated reviews with a very high F1-score of 97%, using an AdaBoost classifier trained on words, POS n-grams, dependency tag n-grams, and NLTK's [11] readability features.

Recent developments in automatic text generation demonstrate that automating the task of fake reviewing is an increasing threat. As generated reviews do not display particular similarities to human-written fake reviews [77], there is no reason to believe that the hypotheses H1–H4 have any particular relevance here. Text generators mimic the writing style of their training corpus, which by assumption contains mostly genuine reviews. Hence, standard deception detection has no bearing on this issue. Instead, classifying generated reviews requires knowledge of the generation model itself, in which case they remain detectable [77, 178]. However, as such knowledge is not always available, the problem cannot be considered solved.

2.2.2 Trolling and cyberbullying. *Troll users* deliberately post malicious content to online forums, either to harass others for amusement or with the intention of advancing an agenda. *Paid trolls* post professionally on behalf of an institution (e.g. a political candidate, government, or corporation), while *mentioned trolls* are identified as trolls by other users [115]. *Cyberbullying* is a related phenomenon, where the author targets a particular victim instead of an entire forum. While trolls or cyberbullies are not exclusively dishonest, there is major overlap in the purposes of a deceiver and a troll: both write content with a purpose other than its truthful communication. Especially professional trolls have no necessary connection between their actual opinion and what they write, and therefore are likely to write content they believe to be false. Additionally, even if a troll writes something he believes, his *intention* is nevertheless fraudulent. Similar considerations apply for cyberbullying. It is therefore initially plausible that trolling/cyberbullying and other forms deception detection might overlap in linguistic features.

Cambria et al. utilized *sentic computing* to classify texts according to the likelihood of being authored by a troll [23]. As a knowledge-based method, sentic computing is more grammatically and semantically oriented than many other current NLP approaches, as it is built on a pre-programmed set of "common-sense" concepts and inference patterns [24]. Cambria et al. used the method to attest the emotional content of the data, based on the following scales:

> "1. the user is happy with the service provided (Pleasantness)
> 2. the user is interested in the information supplied (Attention)
> 3. the user is comfortable with the interface (Sensitivity)
> 4. the user is disposed to use the application (Aptitude)" [23]

Cambria et al., classify human emotions by these four dimensions together with *polarity*, i.e. whether the emotion is positive or negative. They found that troll posts had a high absolute value of Sensitivity and a generally negative polarity. Trolls manifested either significantly high or low levels of comfort with the interface, together with a negative sentiment. Testing on a manually classified test set of troll and non-troll Twitter messages, Cambria et al. received an F-score of 78% (82% precision, 75% recall).

J.-M. Xu et al. [175] used sentiment analysis to detect cyberbullying from Twitter. They manually classified seven emotions relevant for bullying: anger, embarrassment, empathy, fear, pride, relief and sadness. Fear was by far the most

Studies	Data source	Cues to trolling/bullying
[23]	Twitter	negative sentiment
[153]	Discussion forum	negative sentiment
[115]	Discussion forum	bag-of-words, negative sentiment
[27]	Youtube comments	offensive words, intensifiers
[175]	Twitter	fear-related words

Table 5. Linguistic properties of troll comments

common emotion in their cyberbullying dataset, whether the author was identified as a bully, a victim, an accuser or a bystander. These results indicate that fear-related terminology may be informative of bullying as a *topic*, but not of the status of the author as the bully.

Troll posts are commonly negative, being targeted against some viewpoint or a person. Using the hypothesis that negative sentiment is indicative of trolling, Seah et al. [153] applied sentiment analysis to online forum posts. They received a generalized receiver operating characteristic of 78% with binary classification and 69% with ordinal classification.

Mihaylov and Nakov [115] used various linguistic and metalinguistic features to detect both paid and mentioned troll comments in news community forums. They received an F1-score of 78% for mentioned trolls and 80% for paid trolls. Despite the slight differences between the troll types, they conclude that both paid and mentioned trolls behave similarly in comparison to non-trolls. Among linguistic features, bag-of-words fared well overall, as opposed to more abstract grammatical properties like POS-tags. However, metalinguistic features were more effective than any linguistic feature.

Offensive language is an important factor in trolling and cyberbullying. Following Jay and Janschewitz [70], Chen et al. [27] characterize offensive language as vulgar, pornographic or hateful. For evaluating the overall offensiveness of sentences, Chen et al. used an offensive word lexicon that included manual measures for words collected from Youtube comments, and further measurements based on a word's syntactic context. Detecting offensive users in an online discussion corpus, they receive 78% in both precision and recall.

A related issue is *hate speech*, the detection of which has been explored in a number of studies [6, 20, 35, 41, 57, 125, 152, 170, 174, 181]. Hate speech is only occasionally deceptive, which is why we do not discuss it in detail here. However, a brief summary of the findings in this field is worth taking into account. First, character n-grams have generally performed well across hate speech datasets, which is likely due to their flexibility across spelling variants [62, 111, 152]. Second, offensive word lexicons have not performed well in the absence of n-gram features [125, 152]. Third, while deep learning approaches have become more popular than more basic machine learning methods [6, 181], a comparative study by Gröndahl et al. [62] demonstrated that their performance did not significantly differ when trained on the same datasets. Finally, the same study showed that even state-of-the-art approaches are highly vulnerable to simple text transformations like removing spaces or adding innocuous words. Such evasion techniques are similar to earlier methods of evading spam detection [102, 188].

Reviewing the main results of the troll detection studies discussed in this section, the most prevalent cue is *negative sentiment*. It clearly does not suffice, as non-troll messages can also have negative sentiment, and not all trolls are negative. At most the results indicate that negative sentiment is indicative of an increased probability of trolling. Like with fake reviews, studies on troll and cyberbullying detection reflect the fact that *content*, much more than writing style,

has determined the success of classification. Hence, the results do not support the plausibility of a content-invariant detection scheme.

2.3 Deception detection: future prospects

Summarizing the studies reviewed in Sections 2.1–2.2, some results have been replicated in multiple studies. In particular, deceptive texts often have a high emotional load and a large frequency of certainty-related terms, while troll posts tend to have a negative sentiment. However, a more fine-grained analysis demonstrates that the relevant features have been highly content- and context-sensitive. Hence, they are unlikely to scale beyond the semantic domains of particular datasets. Therefore, we suggest that detecting deception is more efficient with methods outside of purely linguistic analysis. Specifically, we recommend *semantic comparison* between different documents. For example, Mihaylov et al. [114] used various measures to detect troll users from an online news community forum, among which was *comment-to-publication similarity*. Their hypothesis was that trolls may be prone to deliberately cite news articles in a misleading fashion to support their own perspective. This feature had a positive effect on classification, and links troll-detection to *rumor-debunking*, where similar content-comparison methods prevail [147].

A related approach to unsupervised fake review classification is the detection of semantic and grammatical similarities between reviews. Such methods rely on the assumption that spammers tend to repeat the same message in multiple places. Narisawa et al. [122] classified spam based on the similarity measure of *string alienness*, obtaining F1-scores between 50% and 80%. Uemura et al. [167] detected review spam using *document complexity* (based on the amount of similar documents within the corpus), and received F1-scores between 66% and 73%. Lau et al.'s [94] unsupervised model was also based on duplicate detection based on semantic overlap.

Of course, semantic comparison measures do not detect author intentions, such as deceptiveness. This task may well be *impossible in principle* if only text data is used. Our literature review indicates that deceptiveness as an author intention does not leave a content-invariant linguistic trace. Deception may, at most, correlate with certain linguistic properties in particular semantic domains. Restricted to a domain, linguistic features may still be useful in aiding deception detection, at least when used in combination with metalinguistic data concerning e.g. user behavior on the forum.

3 AUTHOR IDENTIFICATION AND ADVERSARIAL STYLOMETRY

In this section we discuss *author identification* from an *adversarial* perspective, where detection and its evasion are treated as competing tasks. Avoiding deanonymization or profiling involves obfuscating writing style, for which a variety of techniques has been suggested. Style transformation for anonymization or imitation purposes constitutes a type of *deception*, albeit different in kind from those reviewed in Section 2. There, we characterized deception, broadly understood, as attempting to lead the reader into believing something false. In style transformation, the relevant information involves *author identity* or *profile*, the first concerning individual identity and latter membership in a broader group. Unless mentioned otherwise, the studies reviewed have concerned author identification. With respect to profiling, features are likely to vary depending on the classes under interest (age, gender, occupation etc.), which makes results less generalizable. However, some style transformation studies have concentrated on profiling instead of identification [136, 157].

We begin by reviewing the state-of-the-art in stylometry research in Section 3.1. We then advance to information security -related uses of author identification, to which we devote Section 3.2. After introducing the *deanonymization attack* [121], we dedicate Section 3.3 to discussing its mitigation by style obfuscation or imitation.

3.1 Author identification

The success of author identification depends on the validity of the *Human Stylome Hypothesis* (HSH) [168], which maintains that authors have a unique writing style that is retained to a significant extent between different texts, even across variation in semantic content. Its validity is obviously not a binary matter, and will inevitably differ between authors and datasets. Nevertheless, general trends found in empirical work contribute useful indicators of its suitability for real-world applications. In this section we provide a concise review of existing work in author identification. For further discussion, we refer to prior surveys dedicated solely to this topic [123, 163].

There is a close affinity between writing styles and *idiolects* as speaker-specific (mental) grammars, which can differ among members of the same language community. The idea that lexical repositories and grammatical rules vary between individual speakers is a well-attested linguistic fact [12, 31, 127, 150]. While this gives the HSH initial plausibility, it is worth bearing in mind that idiolects reflect a large variety of factors, not limited to choices between content-equivalent stylistic variants. Indeed, the linguistic literature on idiolects has often focused on *semantic* variation between authors [53, 101]. Another problem for the HSH is the prevalence of *style-shift*. As the sociolinguist William Labov stated: "There are no single style speakers" [91]. If a speaker can change between styles in different contexts, stylometric classification might not capture author identity but rather "style clusters" spanning many authors, who conversely can belong to multiple clusters. Recent results on large-scale stylometric clustering accord with this hypothesis [123].

The problematicity of HSH notwithstanding, concrete examples of author identification can be found outside academic research. In 2011, an American man was found to be the true author of a blog supposedly written by a Syrian woman [9]. While stylometry was not responsible for the finding, Afroz et al. [2] demonstrated that a close linguistic correlation could be found between the blog and other texts by the same author. In 2013, stylometric analysis performed by Peter Millican and Patrick Juola on the novel *The Cuckoo's Nest* revealed its likely author to be J.K. Rowling under a pseudonym, which she later confirmed.[7] Juola [74] also reports a real-life court case where an asylum-seeker claimed to have written newspaper articles critical of his government, for which he would have faced persecution if not granted the asylum. As evidence, he provided other articles provably written by him, and the court had to evaluate their similarity to the contested articles. In such cases, stylometry can provide assistance for making decisions with large-scale consequences.

A significant problem in the field is the lack of consensus on which features to use [73, 141, 148, 149]. The most prevalent collection argued to be optimal for identifying individual authors even from short texts is the "Writeprints" feature set [1, 183]. It consists of a variety of character-based, lexical, syntactic, and structural features, as partly presented in Table 6. The set was introduced by Zheng et al. [183], who used it to identify authors with 97.69% accuracy from a corpus containing 20 candidate authors, and $30 - 92$ articles of $84 - 346$ words from each candidate. I has since been used in multiple studies [1, 2, 4, 48, 109, 130], and is partially implemented in the JStylo software [109].

While large-scale comparisons between different features applied to the same datasets have been rare, existing comparative studies indicate that *low-level features* like short character n-grams (including unigrams) have a systematically high performance. Grieve [61] applied 39 features prevalent in prior work (before 2007) to a single dataset, using the

[7] For Juola's description of the study, see http://languagelog.ldc.upenn.edu/nll/?p=5315.

Feature types		Example features
Lexical	Character	number of characters, number of letters, number of digits, frequency of letters, frequency of special letters
	Word	number of words, average word length, vocabulary richness, average sentence length
Syntactic		frequency of punctuations, frequency of function words
Structural		number of sentences, number of paragraphs, number of sentences/words/characters in a paragraph, has quotes
Content-specific		frequency of content specific keywords

Table 6. Examples of the Writeprints features (270 altogether) [183]

chi-square test for producing a ranking of the most likely authors. The top-5 feature types with the best performance were word unigrams (including punctuation) and character n-grams in the order $2 > 3 > 4 > 1$ from most to least successful. In contrast, positional features, vocabulary richness, sentence length, and word length had only modest or poor performance. A higher prevalence of function words in comparison to content words further improved the success rate, which is in line with traditional assumptions of style being especially manifested in function words [118].

Juola [73] summarizes over 3 million experiments he and colleagues made on the same datasets comparing combinations of features, pre-processing methods, and classifiers included in the authorship attribution software JGAAP [72]. The datasets were taken from an author attribution competition [71], and are provided with JGAAP. The best results were achieved with punctuation features, using nearest neighbours with Manhattan distance for analysis. According to Juola, a likely explanation of these results is that the corpus exhibited a particularly large variance in quotation marks and other non-alphanumeric notation. Therefore, the results are not applicable to datasets where such features have been normalized.

Potthast et al. [134] evaluate the performance of 15 suggested techniques on three datasets. Their results suggest that using *compression* improves the stability of performance across different corpora. The basic idea behind compression is that single compressed files are produced of the candidate author's texts both alone and together with the unknown author's texts, and divergence is then measured between these files [90, 108]. Potthast et al. further remark that character features were the most effective overall. Similar conclusions regarding compression and character features are reached in a larger comparative study by Neal et al. [123], who evaluate 14 open-source algorithms on a corpus containing 1000 authors. Summarizing their results, the authors note that low-level features like characters fare better especially on smaller samples, where high-level features like syntactic dependencies are sparse.

In addition to the discrepancy between feature sets across different studies, a further problem in stylometry research concerns whether the features are more indicative of *style* or *content*. Evidently, highly content-related features like lexical choices are not applicable across different genres or domains [5, 123]. This is likely among the main reasons for the success of function words [21, 118, 182], which have also been argued to correlate with personality types [28], and

form the basis of the linguistic profiling software LIWC [131, 165]. Small susceptibility to content changes is also a virtue of punctuation features [73] and grammatical structure [65, 133, 139].

With respect to classification algorithms, most stylometry research has focused on traditional supervised machine learning methods, such as SVMs, decision trees, Bayesian classification, or distance metrics [123, 163]. SVMs have been particularly popular due to their strong performance on high-dimensional and sparse data [163]. Deep learning applications have recently become more prominent, with a particular focus on recurrent and convolutional neural networks [7, 56, 164]. Brocordo et al. [16] also experiment with *deep belief networks*, which belong to the class of probabilistic generative models. While deep learning methods have generally demonstrated a strong performance in many NLP tasks [58, 180], their large training data requirements present problems with smaller author corpora [123]. Recent approaches to *transfer learning* in NLP have attempted to improve classifier scalability by first training an initial model to perform some task using a large training set, and subsequently fine-tuning the model for different tasks with smaller additional training sets [40, 67]. However, the transferability of other text classifiers to author identification is yet to be studied.

In order to succeed beyond artificial experimental conditions, author identification should be feasible across a large number of candidates with small example corpora from each. However, as Neal et al. [123] state in their survey, existing techniques face challenges in such settings. To detect potential author groups, they performed graph-based clustering in a large corpus. The number of clusters (16) was much smaller than the number of authors (1000), and there was no clear separation between authors. Neal et al. note the possibility that the clusters represent "meta-classes" characterizing multiple authors, and a single author can belong to many such classes. These results are in line with Labov's dictum that a speaker is never bound to a single style, and vice versa [91]. However, it is worth noting that if these hypothetical "meta-classes" are simply assimilated with the clusters, the claim is difficult to either confirm or falsify.

The largest attempt at author identification so far has been contributed by Narayanan et al. [121], whose data was derived from 100,000 different blogs. The features used were post length, vocabulary richness, word shape (the distribution of lower- and upper-case letters), word length, and the frequencies of letters, digits, punctuation, special characters, function words, and syntactic category pairs. Narayanan et al. correctly predicted the author in over 20% of the cases, which is a significant increase from random chance. Still, from the perspective of deanonymization, the approach cannot be considered successful, as it was far more likely to yield a false prediction than the correct one.

There is no single universally accepted protocol for author identification that could be used directly "out-of-the-box". While significant overlap can be found in the features and classifiers used, most studies have been unique with respect to the particular subset of features, and have not conducted systematic evaluations between different combinations. Software like Signature,[8] JGAAP [72], JStylo [109], and RStylo [43] have been developed to alleviate this problem by allowing researchers to conduct stylometric tests with a simple GUI, selecting from a list of pre-programmed features and classifiers. Some of these systems are very restricted in the range of features they offer, which limits their application potential. For instance, RStylo only uses word- or character n-grams, and does not allow their combination in the same test. The most featurally sophisticated application is JStylo, which contains a subset of the Writeprints feature set [183].

Neal et al. [123] give two plausible reasons for the commonly observed effectiveness of short character n-grams in comparison to high-level properties like abstract grammatical relations. First, the latter are sparse in short texts, whereas all texts contain characters. Second, character-features are less susceptible to noise, such as misspellings or grammatical errors. In addition to these benefits, we believe that character-features can have an exceptionally high correlation

[8] http://www.philocomp.net/humanities/signature.htm

with many other features. For instance, the prevalence of particular function words will impact the frequency of their characters, making character-features indirectly responsive to changes in function word use. Hence, low-level features like character n-grams have the potential to record (partial) information of a large variety of textual properties. They can therefore be expected to fare generally better than high-level features, at least with small corpora. Character-features also have the advantage of being *language-independent* in the sense of requiring no language-specific pre-processing, such as tokenization, POS-tagging, or parsing [123].

3.2 Implications for security and privacy

Author identification and profiling have a multifaceted relation to information security. Forensic studies have been on the forefront in traditional stylometry research, providing assistance in uncovering the identities of criminals [31, 110]. Similar methods can help to unmask *troll users* in online forums. As a case study, Galn-Garcia et al. [54] linked troll profiles to their true profiles, and successfully applied the method to a real-life cyberbullying case. Another important application is the detection of *doppelgängers* or *sockpuppets*, i.e. users with multiple accounts. Solorio et al. [161] used SVMs with 239 linguistic features to detect sockpuppet accounts from Wikipedia user comments, and reached a 68% accuracy. Afroz et al. [3] used stylometric techniques to link doppelgänger users with unsupervised clustering, achieving 85% precision and 82% recall on an underground forum dataset.

In contrast to the assistance that author identification can provide for increasing online security, it also constitutes a *privacy threat* by making it possible to deanonymize authors against their will. Brennan et al. [14] propose an adversarial scenario they call *Alice the anonymous blogger versus Bob the abusive employer*, where an employer uses stylometry to uncover the author of an anonymous complaint. Another potential adversarial purpose of deanonymization is *bullying* or *harassment* [4]. In general, the abusive part can be played by any person or institution, such as a government, corporation, or individual. Narayanan et al. [121] coined the term *deanonymization attack* to denote such scenarios. They further take their empirical results on blog author identification to indicate that the attack is not only a theoretical possibility but a real-life concern.

As we reviewed in Section 3.1, Narayanan et al. were able to detect a blog author from $100,000$ candidates with ca. 20% accuracy [121]. We also noted that, while these results demonstrate a major increase from random chance, they nevertheless most often fail to find the correct author. Still, they are genuinely disconcerting from the perspective of potential victims of a deanonymization attack. In designing secure systems, it is essential to assume a low bar for the attacker and a high bar for the defender. Applying this principle to the deanonymization attack, we conclude that since stylometry can significantly increase the chances of the attacker correctly guessing the author's identity, it constitutes a genuine privacy threat. Additionally, results on smaller author corpora (≤ 20) indicate that high accuracy can be achieved with features contained in Writeprints [1, 183], or similar sets [123, 163]. In many variants of the deanonymization attack, assuming a fairly restricted candidate set is justified: for example, in Brennan et al.'s [14] scenario (see above) the candidates are restricted to Bob's employees.

Motivated by their findings, Narayanan et al. [121] recommend the development of automated tools for transforming writing style while preserving meaning. The field of *adversarial stylometry* [14] involves the study of such countermeasures to deanonymization. In Section 3.3 we review the work conducted in this field, and evaluate whether the deanonymization attack can realistically be mitigated using existing methods.

3.3 Adversarial stylometry

We define *style obfuscation* as any method aimed at fooling stylometric classification. A more restricted variant of obfuscation is *imitation*, where misclassification is intended to target a particular author. Imitation also constitutes an attack, with the original author as the attacker and the imitated author as the victim. We divide obfuscation methods into three basic types: manual, computer-assisted and automatic. The subsections 3.3.1–3.3.3 are divided by these methods, and within each subsection studies are reviewed in the order of publication. If one study has used several methods, its results are divided among the subsections.

3.3.1 Manual obfuscation. Brennan and Greenstadt [15] experimented with two manual methods of masking the original author of a text: *obfuscation* and *imitation*. The former involved conscious altering of a text to avoid displaying properties characteristic of the author, and in the latter authors attempted to mimic the style of another writer. The results form the *Brennan-Greenstadt Corpus*. In a subsequent study, Brennan et al. [14] used Amazon Mechanical Turk to crowdsource the obfuscation task. The results, along with the original corpus, form the *Extended Brennan-Greenstadt corpus*, which is provided with the JStylo software [109]. Brennan et al. evaluated obfuscated and imitated texts with three methods: neural networks with the Basic-9 feature set,[9] a synonym-based classifier [29], and SVMs with the Writeprints features [183]. Both obfuscation and imitation resulted in the success rates of all methods dropping significantly, only the SVM-Writeprints classifier remaining above a chance level. Imitation also succeeded in reaching the correct targets. The SVM method was most resistant against both obfuscation and imitation. The effectiveness of the (original) Brennan-Greenstadt corpus against the authorship attribution program JGAAP [72] was further demonstrated by Juola and Vescovi [76]. Amazon Mechanical Turk was also successfully used by Almishari et al. [4] to reduce automatic author recognition. Both obfuscation and readability evaluation were crowdsourced. On a scale from 1 ("Poor") to 5 ("Excellent"), the average readability score was 4.29, indicating success in retaining the original meaning to a significant degree.

The results reviewed here indicate that writing style can be manually altered to deceive author identification. Contrary to the strong interpretation of the HSH, it thus seems possible to change one's writing style, at least with deliberation. However, manual obfuscation is very time-consuming and laborious. Having to consciously alter the style of everything one wants to write anonymously is not a scalable solution. Crowdsourcing is a possible way to outsource manual obfuscation, but Almishari et al. [4] note, sending your original writings to strangers constitutes a privacy risk. Conceivably, the adversary could even act as a Mechanical Turk worker and see the original text as a job offered by the author. Crowdsourcing is also relatively slow and costly to use.

3.3.2 Computer-assisted obfuscation. The idea of computer-assisted manual style obfuscation was introduced by Kacmarik and Gamon [78], who automatically evaluated the feature changes needed to make classification fail with Koppel and Schler's [89] author identification technique. They present a graph linking the features requiring modification, allowing the user to monitor their success at obfuscation. Anonymouth [109] uses the stylometric framework JStylo to evaluate a text written by the user against reference corpora. Based on this evaluation, it gives the user instructions on modifying the text to evade JStylo. Day et al. [36] developed the concept of Adversarial Authorship, and implemented it as an application called AuthorWeb. It displays a user other texts similar to their current writing in style, allowing them to obfuscate text by controlling which texts their current writing resembles.

[9]The Basic-9 feature set consists of the number of unique words, lexical density, Gunning-Fog readability index, character count (without whitespace), the average number of syllables per word, sentence count, the average sentence length, and the Flesch-Kincaid Readability Test. Like Writeprints [183], Basic-9 is also implemented in JStylo [109].

Computer-assisted methods can be useful in comparison to the fully manual approach, given that they reduce the cognitive load involved in deciding which features to alter. Without external cues the author would need to guess which changes to make, which would be unattainable in practice. However, a problem with automatic evaluators like Anonymouth is their reliance on specific corpora and classifiers, which may be unavailable for the author. Furthermore, while computer-assisted obfuscation is easier than purely manual methods, choosing the candidate remains in the hands of the user. Manual work is still needed, which makes these approaches unfeasible in a large scale.

3.3.3 Automatic obfuscation. Fully automated author masking is, unsurprisingly, significantly more difficult to implement than manual or computer-assisted obfuscation. This section reviews studies conducted in this field, and evaluates the performance of the most prominent methods.

Iterative language translation The easiest and most common method used for automatic style alteration and hence author obfuscation is machine translation (MT) across one or more languages back to the source language. This can also be called Iterative Language Translation (ILT) [105], as its effect can be increased by multiple iterations. It was first suggested by Rao and Rohatgi [142] and has since been applied in multiple studies, with divergent results. Here, we survey and analyze this research. All the studies reviewed have used English as the language to obfuscate. With the exception of Keswani et al. [81], Google Translate has been used for translation, sometimes compared with the Bing translator. It should also be noted that, with one exception [157], all the studies were conducted prior to Google's major update that replaced phrase-based statistical MT with deep neural networks (DNNs) [171].

ILT was attempted by Brennan et al. [14] with both Google and Bing translators. They used German and Japanese as intermediate languages, experimenting with English–German–English, English–Japanese–English, and English–German–Japanese–English (the first two exemplifying *one-step* and the last *two-step* translation). Success of obfuscation was measured with a synonym-based classifier, a neural network with the Writeprints features, and an SVM with the Basic-9 features (see footnote 9). Approximately, in one-step translations the success rate of the synonym-based method declined from 85% to 75%, and the success rates of the neural network and SVM classifiers declined from 65% to 50%. Google obfuscated more with Japanese and Bing with German as the intermediate language. With the exception of Bing translations evaluated with neural networks and SVMs, two-step translation did not markedly improve obfuscation. Overall, the results were deemed underwhelming, and the authors concluded that state-of-the-art MT in 2009 did not provide sufficient means for author obfuscation.

Caliskan and Greenstadt [22] also used Google and Bing's translators with German and Japanese as intermediate languages, but with English–Japanese–German–English as the two-step translation order. The success of obfuscation was measured with JGAAP [75][72] and JStylo [109], using what they call the Translation Feature Set, which was selected via optimization from the Basic-9 and Writeprints feature sets [15, 183].[10] After obfuscation, the average recognition rate remained high at 92%, which accorded with Brennan et al.'s [14] pessimistic conclusions about ILT. Caliskan and Greenstadt further classified translated texts based on the translator (Google or Bing) with an average success rate of 91%, indicating that the translation algorithm itself can be "fingerprinted" if appropriate stylometric features are used.

Using Google Translate, Almishari et al. [4] reduced the linkability between the translated text and the original author by increasing the amount of intermediate languages up to nine, randomly drawn from the 64 languages offered by Google Translate in 2014. They conducted a readability review of 60 translations (produced via nine intermediate

[10]The Translation Feature Set contained the following features: average characters per word, character count, function words, letters, punctuation, special characters, top letter bigrams, top letter trigrams, words, and word lengths [22].

languages) via Amazon Mechanical Turk, receiving an average score of 2.8/5. The readability of a subset of translated texts was further improved manually (also via Mechanical Turk), retaining author anonymity. However, without a comparison to texts produced with other methods, the readability score alone does not tell much about the status of the obfuscated texts.

Mack et al. translated English blog texts back and forth through Arabic, Chinese and Spanish with one to three iterations [105]. The results were evaluated with four Author Identification Systems (AISs): a unigram-based AIS, the O. de Vel et al. AIS [37], a combination of the previous two called Hybrid-I, and Hybrid-I with added syntactic features (Hybrid-II). A genetic algorithm called Genetic and Evolutionary Feature Selection (GEFeS) was further used to mask nonsalient features from each AIS to improve their performance. The addition of GEFeS resulted in Hybrid-II having the best performance overall on recognizing the author from the non-obfuscated test corpus (52%), the other AISs having rates of c.a. 20% – 25%. ILT lowered the identification rate of all AISs with all intermediate languages, Arabic always faring the best. The most significant result was the decrease of Hybrid-II plus GEFeS' identification rate from 51.65 % to c.a. 10 % with all languages on the first iteration, and below 10 % with Arabic. Further iterations did not markedly change the identification rates, irrespective of the language or the AIS.

Day et al. [36] used Google Translate with Spanish and Chinese as intermediate languages. Mack et al.'s Hybrid-II AIS [105] was used to evaluate the success of obfuscation. With the first iteration, accuracy dropped from 54% to 6% with Spanish and to 10% with Chinese. Further iterations did not decrease the identification accuracy, as the second iteration resulted in 6% with Spanish and 11% with Chinese, and the third iteration yielded 7% with Spanish and 11% with Chinese. Day et al. further used Latent Semantic Analysis (LSA) [42] to analyse semantic overlap between the original and obfuscated texts. This algorithm gives a value between -1 (no similarities) and 1 (full similarity). The LSA-values were 0.86 for the Spanish-mediated text and 0.77 for the Chinese-mediated text. Day et al. further fingerprinted the intermediate language with the JGAAP software [72], receiving accuracies of 93% (Spanish) and 90% (Chinese) on the first iteration, 98% (Spanish) and 97% (Chinese) on the second iteration, and 98% (Spanish) and 99% (Chinese) on the third iteration. The number of iterations was also fingerprintable, although less accurately than the translator.

Keswani et al. [81] applied ILT to the author masking task arranged by the PAN 2016 digital forensics event. Using Moses [88], they created their own translation model trained with the Europarl corpus [87]. The text was translated through German and French. Three features were evaluated of the obfuscated texts [135]. *Safety* indicates how well the obfuscated text manages to hide original authorship, and was measured by the obfuscation's impact on classification by various author verification systems from previous PAN tasks. Keswani et al.'s method succeeded in obfuscation 25% – 42% of the time, depending on the dataset. The *sensibility* of the obfuscated text and its *soundness*, i.e. similarity in meaning with the original text, were both manually evaluated from a small subset of texts. Keswani et al.'s text was, in Potthast et al.'s words, considered "impossible to read or understand" by the PAN 2016 evaluator due to the frequency of errors [135].

As a baseline for evaluating their Generative Adversarial Network (GAN) approach called A^{4}NT (discussed below), Shetty et al. [157] applied four variants of ILT with Google Translate, using German, French, Spanish, Finnish, and Armenian as intermediate languages between two and five iterations. None of the variants significantly reduced the classification rate on a word-based Long Short Term Memory (LSTM) network, the largest drop being from 90% to 81% in F1-score. Shetty et al.'s user study also indicated that ILT did not succeed in maintaining semantic similarity.

Based on our review, one reason for the differing outcomes in ILT-obfuscation seems to be the languages used. As summarized in Table 7, studies have generally used different intermediate languages and numbers of iterations. While

Study	Translator(s)	Languages	Iterations	Success
[14]	Google, Bing	German, Japanese	$1-2$	No
[22]	Google, Bing	Japanese, German	$1-2$	No
[4]	Google	(Random)	≤ 9	Yes
[105]	(Not told)	Arabic, Chinese, Spanish	$1-3$	Yes
[36]	Google	Spanish, Chinese	$1-3$	Yes
[81]	Moses	German, French	2	Unclear / No
[157]	Google (DNN)	German, French, Spanish, Finnish, Armenian	$2-5$	No

Table 7. A comparison of studies using ILT for style obfuscation
("Success" = the reported success of the approach in deceiving author identification, based on the source paper.)

small-scale comparisons have been made, the effects of varying the languages have not been systematically evaluated. Results on the effects of iterations are also indecisive. Almishari et al. [4] decreased identification accuracy by adding iterations, whereas Mack et al. [105] and Day et al. [36] did not. More systematic comparative research would be needed to properly evaluate the effects of the languages, the number and direction of iterations, and the translation method. With respect to the last, it is possible that Shetty et al.'s [157] failure to obfuscate with Google Translate even across five intermediate languages was affected by its update from a statistical algorithm to a DNN [171].

It is also likely that ILT will decrease the grammaticality and hence readability of the text, and/or differentiate its semantic interpretation from the original text [123, 142]. Successful change of style would require three properties from the resulting text: (i) grammatically soundness, (ii) retention of the original meaning, and (iii) evasion of author identification. These properties have not been properly measured *together* in the ILT-obfuscation studies reviewed in this section. Additionally, ILT is unreliable due to not taking into account the *direction* of the obfuscation. The changes might take the classification to *any* direction, which may or may not aid obfuscation. As Shetty et al. [157] note, a style obfuscation system should ideally only enact changes that take classification to the desired direction, and no more.

Rule-based substitutions Khosmood and Levinson [82] outline a basic model of rule-based style imitation based on grammatical changes. The purpose of the system is to alter the style of a source text until it is maximally close to that of a target corpus. First, both the source text and the target corpus are analysed based on *style markers*, which are predetermined linguistic features. Next, the styles of the source text and the target corpus are compared, and their stylistic distance is determined based on some metric calculated from the style markers. If the distance is large enough, a predetermined modification rule is applied to the source text that alters some of its style markers. A comparison between the altered source text and the target corpus is made, resulting in a finished transformation if their stylistic distance is close enough, and to another modification otherwise. This *Classification-Transformation Loop* (CTL) [83] is continued until the stylistic distance is sufficiently close or no more transformations are available.

Khosmood and Levinson [82] applied the CTL to a US Department of Justice memorandum excerpt, with a part of Orwell's Animal Farm as the target corpus. They used 10 style markers for analysis and comparison, and modified the source text with three transformation rules of de-hyphenation, lexical substitution and acronym expansion. Stylistic distance was measured with the root-mean-square-error value, which reduced from 5.77 in the source text to 5.63 in the modified text. In another study [84], the same authors divided text into n-grams between one and five units, and searched for synonyms for each from Wordnet [49, 117]. For each word/phrase, one synonym was chosen above others based on its commonness in the target corpus or another database, and the original word/phrase was replaced with it. Using JGAAP for evaluation, Khosmood and Levinson succeeded in obfuscating seven out of thirteen texts. Readability

was manually evaluated by the authors as "correct", "passable" or "incorrect", concluding that the majority of texts were "passable". As no additional evaluators were used, and the authors only provide two example obfuscations, these results are difficult to assess.

Mansoorizadeh et al. [107] employed Wordnet-based lexical substitution for the PAN 2016 Author Masking Task [135]. Synonyms for the top 200 words used by the author of the training text were found from Wordnet via NLTK. Two criteria were used in choosing the replacement synonym from the alternatives provided by Wordnet: Wu et al's semantic similarity metric [172], and the occurrence probability in the original word's context measured with a 4-gram language model trained with the Brown corpus. In their evaluation of the PAN 2016 Author Masking Task, Potthast et al. [135] note that Mansoorizadeh et al.'s algorithm is very conservative, changing at most one word per sentence. While retaining readability well, this also results in reduced safety against author identification. Further, certain replacements resulted in semantic errors, such as *machine* being exchanged for *car*. The system succeeded in obfuscation in 14% – 25% of cases depending on the dataset. In Potthast et al.'s manual sensibility evaluation, the obfuscated text received a grade of 2/5 on a scale from 1 (excellent) to 5 (fail), mainly due to punctuation errors. In a further manual evaluation of similarity to the original text on a three-point scale of "incorrect", "passable" and "correct", all obfuscated texts were graded as "correct" or "passable".

In the same PAN 2016 Author Masking Task, Mihaylova et al. applied various alterations by replacing elements with others, merging, splitting, removal etc. [116]. In Potthast et al.'s evaluation [135], Mihaylova et al. received the best results of all PAN 2016 Author Masking Task contenders in safety (i.e. the success of obfuscation), with an average impact of 36 % – 49 % depending on the dataset. However, in the manual evaluation of sensibility and soundness, the obfuscated texts were deemed practically unreadable and semantically odd. In a subsequent study, the same authors applied similar but improved techniques to the same test setting, shifting stylistic features toward their average distribution in the training set [79]. Using multiple author identification methods from prior PAN competitions, their method achieved an average accuracy drop between 10% and 16%, and maintained a superior readability compared to their prior method.

In terms of retaining the original meaning, rule-based substitution is a more secure obfuscation method than ILT, as it allows deterministic user control of the output. Especially with grammatical changes, transformations can be limited to have only minor semantic impact. However, the scalability of hand-crafted rules across a large variety of datasets is difficult to attain [157]. With lexical replacements semantic retention is harder to control, as the appropriateness of paraphrases can be highly context-dependent. If WordNet is used for synonym replacement, context effects can partly be accounted for by using sense disambiguation techniques, such as the Lesk algorithm [96, 140]. WordNet represents words in the uninflected lemma format, which restricts synonym replacement to contexts where the surface form is identical with the lemma. The Paraphrase Database (PPDB) [55] is the major alternative to WordNet, and links inflected forms directly. Derived from parallel corpora used for MT, PPDB also involves information about the appropriate syntactic environment for the paraphrases. However, since the phrases are represented as raw text, it does not directly allow the use of Lesk or other semantic sense disambiguation algorithms.

MT between styles In addition to translating across different languages, MT can be used within the same language to automatically paraphrase text. Importantly, such a method could allow not only obfuscation of the original author but automatic *imitation* of a predetermined style. MT was used for style transformation by Xu et al. [176], who paraphrased Shakespeare as modern English and evaluated the results both manually and with three automatic methods based on cosine similarity (n-gram overlap), language models with Bayesian probability estimation, and logistic regression. The automatic metrics correlated with human judgement to a significant degree.

In addition to using ILT for obfuscation (see above), Day et al. [36] applied iterative paraphrasing, creating the paraphrase dataset with the online tool Plagiarisma. Paraphrasing decreased the author identification rate with Hybrid-II [105] from 54% to 7% in the first iteration, 1% with the second iteration, and 6% with the third iteration. The LSA-value for paraphrased text was 0.80, indicating relatively high lexical overlap with the original text. Like the MT algorithms, paraphrasing itself was detectable, with fingerprinting accuracies of 86% on the first iteration, 91% on the second iteration, and 95% on the third iteration.

More recently, *neural machine translation* (NMT) techniques [103, 171] have been adopted for automatic style imitation. The input is first mapped to a style-neutral representation, and then a new sentence is generated from this representation while controlling target style. However, the transformations implemented in these studies have often involved semantic changes, as in altering sentiment or political slant [69, 136, 156]. In contrast, the main goal of adversarial stylometry is to retain semantic content to a maximal extent while fooling the author classifier. This is evidently not achieved with examples like Prabhumoye et al.'s political slant transformation from "i thank you, sen. visclosky" to "i'm praying for you sir" [136]. Such examples may deceive a Democrat-Republican classifier, but they also change the original meaning too much to constitute viable forms of transformation for anonymization purposes. Since we are concerned with adversarial stylometry and not content alteration, we do not review research on the latter. An important aspect of future work is a more systematic application of the suggested methods to different kinds of tasks, with a particular focus on their ability to retain content across stylistic changes.

In addition to experimenting on political slant and sentiment, Prabhumoye et al. [136] also tackle the issue of gender profiling, which falls under our scope by being a purely author-related, non-semantic feature. The basis for their method is the notion that translation to another language will remove many style-specific features [137]. First, they train translators between English and French to both directions, and begin the style transformation process by translating the original sentence to French. They then process the French translation with the encoder part of the French-to-English translator. The decoder part of the translator is a generative model that takes the French encoding as a context vector and produces an English target sentence. They split this decoder into different variants, which are trained to produce sentences allocated to particular categories by a CNN classifier. The resulting sentences are thus the combined effect of the original French-English translator and the class-based tuning of the English decoder. Prabhumoye et al. compare their back-translation method with Shen et al.'s [156] cross-aligned autoencoder approach, which is similar but uses a different algorithm for generating the intermediate style-neutral representation. The gender classifier's original accuracy of 82% was reduced to 40% with cross-aligned autoencoders and 43% with back-translation. In a manual fluency evaluation on 60 random sentences, gender imitation by cross-aligned autoencoders received an average rating of 2.42/4, while imitation by back-translation received 2.81/4.

Shetty et al. [157] present a Generative Adversarial Network (GAN) -based approach to style transformation, which they title Adversarial Author Attribute Anonymity Neural Translation (A^4NT). A GAN consists of a classifier trained to discriminate between two or more classes, and a generative model that is trained to fool this classifier [59]. A^4NT is an unsupervised approach where an encoder-decoder network is trained to generate sentences which fool a word-based LSTM author classifier, but also maintain a maximal semantic proximity to the original sentence. Semantic retention was measured as a combination of two components: the probability of reconstructing the original sentence via a reverse A^4NT-transformation, and the distance of sentence embeddings constructed using a pre-trained embedding model [30]. A^4NT was tested across three classification tasks: blog author gender, blog author age, and political speeches by Barack Obama vs. Donald Trump. In all tasks, the method lowered classification accuracy to random chance or below. However,

these results only concerned the same classifier as used in training the GAN. Shetty et al. further show that blog age classification F1-score is dropped from 87% to 62% with the best of 10 alternative classifier candidates. Corresponding results from the two other tasks are not shown. For assessing semantic similarity, they use the MT evaluation metric Meteor [38], which measures n-gram overlap using additional paraphrase tables. They receive scores of 0.69, 0.79, and 0.29 in the gender, age, and Obama/Trump tasks, respectively. Shetty et al. note that these results exceed those received with automatic paraphrasing methods [99], although such comparison is problematic as the studies involve different corpora. Finally, a user-study indicated that human evaluators preferred A^{4}NT to ILT via Google Translate with a similar obfuscation success.

Of the approaches reviewed here, only A^{4}NT has a built-in mechanism for semantic retainment. In spite of this, even its example transformations often include drastic semantic changes, as seen in the following Obama-Trump transformations taken from Shetty et al. [157]:

> "their situation is getting worse." → "their media is getting worse."
> "(...) because i do care" → "(...) because they don't care."
> "that's how our democracy works." → "that's how our horrible horrible trade deals."

A system that cannot secure sufficient semantic retention is unreliable for real-life application, irrespective of its success in fooling author identification. Overall, recent advances in NMT and GANs show promise in generating stylistic transformations, but further research is required to evaluate the feasibility of such methods in more realistic scenarios against a large variety of classifiers. Beyond adversarial stylometry, the transformation of writing style has been studied within *automatic text simplification* [120, 158, 159, 173], which in turn belongs to the broader field of *paraphrase generation* [99, 106]. Effects of these methods on author obfuscation have yet to be investigated, but increasing the interaction between these fields would likely be beneficial to both sides.

4 CONCLUSIONS

Section 1 presented the following three questions concerning deception detection based on writing style:

- Q1 Does deception leave a content-independent stylistic trace?
- Q2 Is the deanonymization attack a realistic privacy concern?
- Q3 Can the deanonymization attack be mitigated with automatic style obfuscation?

Based on the literature review conducted in Section 2, Q1 was answered negatively. We demonstrated that linguistic features that have correlated with deception have been too specific to particular semantic domains to constitute genuine stylistic "deception markers". The practical consequence of this finding is that stylometric analysis has plausible utility for deception detection only if the training and test domain are sufficiently similar. Furthermore, even when successful, stylistic markers of deception are likely to be *content-based* correlates rather than indicators of general psychological mechanisms behind lying.

Our review suggest that alternative approaches to pure stylometry are likely more effective in detecting textual deception. These include, in particular, *content comparison*, *similarity detection*, and using *metadata*. Content comparison allows detecting texts that contain claims with a pre-established truth-value based on an external knowledge base [147]. False claims are not deceptive if they are sincerely believed (see Section 2), but a strong correlation between falsity and deception is nevertheless likely. Surface-level similarity between texts has also proven helpful in finding trolls or spammers, who tend to repeat the same across many discussions [94, 122, 167]. Finally, information beyond linguistic

content has been more effective in detecting fake reviews [119, 144] or trolls [115] than linguistic content. Stylometric classification can assist such techniques, but is severely limited as a stand-alone solution.

With respect to Q2, we argued that stylometry-based deanonymization constitutes a realistic privacy threat, especially if the set of potential authors is small (ca. $\leq$ 20) [4, 14, 15, 183]. Even though author identification has not proven sufficiently scalable to larger sets of authors (ca. $>$ 1000) [123], stylometry can still significantly increase the likelihood of finding the correct author, even among 100000 candidates [121]. From the perspective of an author wishing to retain anonymity, these results are legitimately worrying. With the constant increase in the availability of corpora and computing power, the deanonymization attack will likely continue to be a growing privacy threat. We therefore consider the further development of automatic style obfuscation tools as not merely an academic excercise, but to have important real-life consequences for information security.

Turning to Q3, manual obfuscation remains potentially effective against the deanonymization attack [4, 14, 15], and tools like Anonymouth [109] can help in this task. Fully automatic approaches, in contrast, suffer from the difficulty of balancing sufficient obfuscation success with semantic faithfulness to the original text. So far, only simple rule-based approaches have allowed securing semantic retention, as transformations can be limited to semantically vacuous choices [79, 82]. However, these methods are very limited in application, and have not demonstrated sufficient obfuscation success. Only one of all the studies reviewed in Section 3.3.3 included a semantic similarity measure in the algorithm [157], and even it had trouble with too severe semantic alterations. Approaches have largely relied on *a priori* assumptions about ILT or paraphrase replacement not altering semantics, which has not been sufficiently confirmed. Additionally, while user studies are important for assessing readability and semantic retention, the lack of established baselines makes the results difficult to evaluate. Merely comparative measures between different techniques are also inadequate, as they do not demonstrate whether the transformations are acceptable, but only which are preferred under an obligatory choice.

The detectability of obfuscation methods themselves has not been sufficiently investigated, as only two of the studies we reviewed had conducted such an evaluation [22, 36]. All ILT variants could be detected with a high accuracy in both studies, including even the number of intermediate languages. Day et al. also successfully fingerprinted a paraphrase-based MT-algorithm [36]. These results indicate that even if obfuscation succeeds, obfuscated texts could still be distinguished from original texts. However, it bears emphasis that such classification requires knowledge of the obfuscation algorithm, which may not be available. The general property of being obfuscated with *any* method is unlikely to leave a stylistic trace. The situation is similar to the case of automatically generated fake reviews (Section 2.2.1), where the detection of generated text is possible, but only provided that the generation algorithm is known [77, 178].

Summarizing our discussion on adversarial stylometry, while promising frameworks for automatic text transformation exist especially within NMT, securing semantic retention has not been sufficiently studied or implemented in state-of-the-art style transformation applications. We believe that this constitutes the most important challenge for the field going forward. We further suggest that increased interaction between different fields would likely prove useful. While we have focused on style transformation from the perspective of information security, the field of *automatic paraphrasing* is much broader in scope [99, 106], involving tasks such as automatic text simplification [120, 158, 159, 173], controlling for style in MT [155], politeness transformation [143], or generating exercises for language pedagogy [8]. Systematically examining the effects of methods developed for other purposes on style obfuscation would constitute a valuable addition to the field.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information and System Security*, 26(2):1–29, 2008.

[2] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 461–475, 2012.

[3] Sadia Afroz, Aylin Caliskan-Islam, Ariel Stolerman, Rachel Greenstadt, and Damon McCoy. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*, pages 212–226, 2014.

[4] Mishari Almishari, Ekin Oguz, and Gene Tsudik. Fighting Authorship Linkability with Crowdsourcing. In *Proceedings of the second ACM conference on Online social networks*, pages 69–82, 2014.

[5] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. Automatically profiling the author of an anonymous text. *Communcations of the ACM*, 52(2):119–123, 2009.

[6] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[7] Douglas Bagnall. Author identification using multi-headed recurrent neural networks. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation Forum*, 2015.

[8] Jorge Baptista, Sandra Lourenco, and Nuno Mamede. Automatic generation of exercises on passive transformation in Portuguese. In *IEEE Congress on Evolutionary Computation (CEC)*, pages 4965–4972, 2016.

[9] Daniel Bennett. A 'Gay Girl in Damascus', the Mirage of the 'Authentic Voice' - and the Future of Journalism. In Richard Lance Keeble and John Mair, editors, *Mirage in the Desert? Reporting the Arab Spring*, pages 187–195. Abramis, Bury St. Edmunds, 2011.

[10] Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, Edward Finegan, and Randolph Quirk. *Longman Grammar of Spoken and Written English, volume 2*. Pearson Education, Harlow, 1999.

[11] Steven Bird and Edward Loper. NLTK: the natural language toolkit. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, 2004.

[12] Bernard Bloch. A set of postulates for phonemic analysis. *Language*, 24(1):3–46, 1948.

[13] Charles F. Bond. and Bella M. DePaulo. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234, 2011.

[14] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15(3), 2011.

[15] Michael Brennan and Rachel Greenstadt. Practical Attacks Against Authorship Recognition Techniques. In *Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence*, pages 60–65, 2009.

[16] Marcelo Luiz Brocardo, Issa Traore, Isaac Woungang, and Mohammad S. Obaidat. Authorship verification using deep belief network systems. *Communication Systems*, 30(12), 2017.

[17] David B. Buller and Judee K. Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.

[18] David B. Buller, Judee K. Burgoon, Aileen Buslig, and James Roiger. Testing interpersonal deception theory: The language of interpersonal deception. *Communication theory*, 6(3):268–289, 1996.

[19] Judee K. Burgoon, J. P. Blair, Tiantian Qin, and Jay F. Nunamaker, Jr. Detecting deception through linguistic analysis. In *Proceedings of the 1st NSF/NIJ Conference on Intelligence and Security Informatics*, pages 91–101, Berlin, Heidelberg, 2003. Springer-Verlag.

[20] Pete Burnap and Matthew L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[21] John F. Burrows. Word patterns and story shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing*, 2:61–70, 1987.

[22] Aylin Caliskan and Rachel Greenstadt. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In *IEEE Sixth International Conference on Semantic Computing (ICSC)*, pages 121–125, 2012.

[23] Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. Do not feel the trolls. In *Proceedings of the 3rd International Workshop on Social Data on the Web (SDoW2010)*, 2010.

[24] Erik Cambria and Amir Hussain. *Sentic Computing: A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Springer International Publishing, Cham, 2015.

[25] Antonio Castro and Brian Lindauer. Author Identification on Twitter. In *Third IEEE International Conference on Data Mining*, pages 705–708, 2013.

[26] Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *Transactions on Intelligent Systems and Technology*, 3(2):27, 2011.

[27] Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust and of the 2012 International Conference on Social Computing (PAS-SAT/SocialCom '12)*, pages 71–80, Amsterdam, 2012.

[28] Cindy K. Chung and James W. Pennebaker. The psychological functions of function words. In Klaus Fiedler, editor, *Frontiers of social psychology: Social communication*, pages 343–359. Psychology Press, New York, 2007.

[29] Jonathan H. Clark and Charles J. Hannon. A classifier system for author recognition using synonym-based features. In Alexander Gelbukh and Ángel Fernando Kuri Morales, editors, *Lecture Notes in Computer Science, vol. 4827*, pages 839–849. Springer, 2007.

[30] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680, 2017.

[31] Malcolm Coulthard. Author identification, idiolect and linguistic uniqueness. *Applied Linguistics*, 25(4):431–447, 2004.

[32] Erin Smith Crabb. "Time for some traffic problems": Enhancing e-discovery and big data processing tools with linguistic methods for deception detection. *Journal of Digital Forensics, Security and Law*, 9(2), 2014.

[33] Walter Daelemans. Explanation in computational stylometry. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference (CICLing2013)*, pages 451–462, 2013.

[34] Robert Dale and Ehud Reiter. *Building natural language generation systems*. Cambridge University Press, Cambridge, 2000.

[35] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th Conference on Web and Social Media*, pages 512–515, 2017.

[36] Siobahn Day, James Brown, Zachery Thomas, India Gregory, Lowell Bass, and Gerry Dozier. Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering. In *25th International Conference on Computer Communication and Networks (ICCCN)*, pages 1–6, 2016.

[37] Olivier de Vel, Alison Anderson, Malcolm Corney, and George Mohay. Mining e-mail content for author identification forensics. *ACM Sigmod Rev.*, 30(4):55–64, 2001.

[38] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, 2014.

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

[40] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[41] Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligen Systems*, 2(3):18:1–18:30, 2012.

[42] Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[43] Maciej Eder, Jan Rybicki, and Mike Kestemont. Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8(1):107–121, 2016.

[44] Paul Ekman. *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. Norton, New York, 1985.

[45] Paul Ekman and Wallace V. Friesen. Nonverbal leakage and clues to deception. *Psychiatry*, 32(1):88, 1969.

[46] Paul. Ekman and Maureen O'Sullivan. Who can catch a liar? *American Psychologist*, 46(9):913–920, 1991.

[47] Frank Enos, Elizabeth Shriberg, Martin Graciarena, Julia Hirschberg, and Andreas Stolcke. Detecting deception using critical segments. In *Proceedings of Interspeech*, pages 1621–1624, 2007.

[48] Iqbal Farkhund, Hamad Binsalleeh, Benjamin C.M. Fung, and Mourad Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digital Investigation*, 7(1–2):56–64, 2013.

[49] Christiane Fellbaum (ed.). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, 1998.

[50] 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.

[51] Eileen Fitzpatrick and Joan Bachenko. Building a forensic corpus to test language-based indicators of deception. *Language and Computers*, 71(1):183–196, 2009.

[52] Eileen Fitzpatrick, Joan Bachenko, and Tommaso Fornaciari. *Automatic Detection of Verbal Deception*. Morgan & Claypool, 2015.

[53] Douwe Fokkema and Elrud Ibsch. *Modernist Conjectures. A Mainstream in European Literature*. Hurst, London, 1987.

[54] Patxi Galán-García, José Gaviria de la Puerta, Carlos Laorden Gómez, Igor Santos, and Pablo García Bringas. Supervised machine learning for the detection of troll profiles in Twitter social network: Application to a real case of cyberbullying. In *International Joint Conference of Advances in Intelligent Systems and Computing*, pages 419–428, 2014.

[55] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. PPDB: The Paraphrase Database. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 758–764, 2013.

[56] Zhenhao Ge and Yufang Sun. Domain Specific Author Attribution based on Feedforward Neural Network Language Models. In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 597–604, 2016.

[57] Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230, 2015.

[58] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57(1):345–420, 2016.

[59] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27 (NIPS)*, pages 2672–2680, 2014.

[60] Paul Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge/London, 1989.

[61] Jack Grieve. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 2007.

[62] Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. All you need is "love": Evading hate speech detection. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security (AISec'11)*, pages 2–12, 2018.

[63] Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorhaand, and Michael Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2008.

[64] Jeffrey T. Hancock and David M. Markowitz. Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PLoS ONE*, 9(8), 2014.

[65] Graeme Hirst and Olga Feiguina. Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, 22(4):405–417, 2007.

[66] Dirk Hovy. The enemy in your own camp : how well can we detect statistically-generated fake reviews – an adversarial study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 351–356, 2016.

[67] Jeremy Howard and Sebastian Ruder. Fine-tuned language models for text classification. *CoRR*, abs/1801.06146, 2018.

[68] Nan Hu, Indranil Bose, Noi Sian Koh, and Ling Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Systems*, 52(3):674–684, 2012.

[69] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. Toward controlled generation of text. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1587–1596, 2017.

[70] Timothy Jay and Kirstin Janschewitz. The pragmatics of swearing. *Journal of Politeness Research. Language, Behaviour, Culture*, 4(2):267–288, 2008.

[71] Patrick Juola. Ad-hoc Authorship Attribution Competition. In *Proceedings of the 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH)*, 2004.

[72] Patrick Juola. JGAAP: A System for Comparative Evaluation of Authorship Attribution. *Journal of Digital Humanities and Computer Science*, 1(1), 2009.

[73] Patrick Juola. Large-Scale Experiments in Authorship Attribution. *English Studies*, 93(3):275–283, 2012.

[74] Patrick Juola. Stylometry and immigration: A case study. *Journal of Law and Policy*, 21(2):287–298, 2013.

[75] Patrick Juola, John Sofko, and Patrick Brennan. A prototype for authorship attribution studies. *Literary and Linguistic Computing*, 2(21):169–178, 2006.

[76] Patrick Juola and Darren Vescovi. Empirical evaluation of authorship obfuscation using JGAAP. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security (AISec'10)*, pages 14–18, 2010.

[77] Mika Juuti, Bo Sun, Tatsuya Mori, and N. Asokan. Stay on-topic: Generating context-specific fake restaurant reviews. In *Proceedings of the 23rd European Symposium on Research in Computer Security (ESORICS)*, pages 132–151, 2018.

[78] Gary Kacmarcik and Michael Gamon. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of COLING/ACL: Poster Sessions*, pages 444–451, 2006.

[79] Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. The case for being average: A mediocrity approach to style masking and author obfuscation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–185, 2017.

[80] Parambir S. Keila and David B. Skillicorn. Detecting unusual and deceptive communication in email. In *Centers for Advanced Studies Conference*, pages 17–20, 2005.

[81] Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumde. Author Masking through Translation - Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, 2016.

[82] Foaad Khosmood and Robert Levinson. Automatic natural language style classification and transformation. In *Proceedings of the 2008 BCS-IRSG Conference on Corpus Profiling*, page 3, 2008.

[83] Foaad Khosmood and Robert Levinson. Toward automated stylistic transformation of natural language text. In *Proceedings of the Digital Humanities*, pages 177–181, 2009.

[84] Foaad Khosmood and Robert Levinson. Automatic synonym and phrase replacement show promise for style transformation. In *Proceedings of the Ninth International Conference on Machine Learning and Applications*, pages 958–961, 2010.

[85] Bryan Klimt and Yiming Yang. The Enron Corpus: A New Dataset for Email Classification Research. In *Machine Learning: 15th European Conference on Machine Learning (ECML2004)*, pages 217–226, 2004.

[86] Mark Knapp and Mark Comaden. Telling it like it isn't: A review of theory and research on deceptive communications. *Human Communication Research*, 5(3):270–285, 1979.

[87] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, pages 79–86, 2005.

[88] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, and Richard Zens. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180, 2007.

[89] Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pages 489–495, 2004.

[90] Olga V. Kukushkina, Anatoly A. Polikarpov, and Dimitry V. Khmelev. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184, 2001.

[91] William Labov. Field Methods of the Project in Linguistic Change and Variation. In John Baugh and Joel Sherzer, editors, *Language in Use: Readings in Sociolinguistics*, pages 28–66. Prentice Hall, Englewood Cliffs, 1984.

[92] J. Clayton Lafferty and Patrick M. Eady. *The Desert Survival Problem.* Experimental Learning Methods, Plymouth, Michigan, 1974.

[93] David F. Larcker and Anastasia A. Zakolyukina. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540, 2012.

[94] Raymond Y.K. Lau, S.Y Liao, Ron Chi-Wai Kwok, Kaiquan Xu, Yunqing Xia, and Yuefeng Li. Text mining and probabilistic language modeling for online review spam detecting. *ACM Transactions on Management Information Systems*, 4(2):1–30, 2011.

[95] Chih-Chen Lee, Robert B. Welker, and Marcus D. Odom. Features of computer-mediated, text-based messages that support automatable, linguistics-based indicators for deception detection. *Journal of Information Systems*, 23(1):5–24, 2009.

[96] Michael Lesk. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine code from an ice cream cone. In *Proceedings of the 5th annual international conference on systems documentation*, pages 24–26, 1986.

[97] Jiwei Li, Myle Ott, and Claire Cardie. Identifying manipulated offerings on review portals. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 18–21, 2013.

[98] Jiwei Li, Myle Ott, Claire Cardie, and Eduard Hovy. Towards a General Rule for Identifying Deceptive Opinion Spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1566–1576, 2014.

[99] Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. Paraphrase Generation with Deep Reinforcement Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3865–3878, 2018.

[100] Max Louwerse, K Lin, A Drescher, and Gün Semin. Linguistic cues predict fraudulent events in a corporate social network. In *Proceedings of the 32 Annual Conference of the Cognitive Science Society*, pages 961–966, 2010.

[101] Max M. Louwerse. Semantic variation in idiolect and sociolect: Corpus linguistic evidence from literary texts. *Computers and the Humanities*, 38(2):207–221, 2004.

[102] Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *Proceedings of the Second Conference on Email and Anti-Spam (CEAS)*, 2005.

[103] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1412–1421, 2015.

[104] Wicenty Lutoslawski. *Principes de stylometrie.* E. Leroux, 1898.

[105] Nathan Mack, Jasmine Bowers, Henry Williams, Gerry Dozier, and Joseph Shelton. The Best Way to a Strong Defense is a Strong Offense: Mitigating Deanonymization Attacks via Iterative Language Translation. *International Journal of Machine Learning and Computing*, 5(5):409–413, 2015.

[106] Nitin Madnani and Bonnie Dorr. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Journal of Computational Linguistics*, 36(3):341–387, 2010.

[107] Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. Author Obfuscation using WordNet and Language Models – Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, 2016.

[108] Yuval Marton, Ning Wu, and Lisa Hellerstein. On compression-based text classification. In *Advances in Information Retrieval*, pages 300–314, 2005.

[109] Andrew W.E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. Use fewer instances of the letter i: Toward writing style anonymization. In *Privacy Enhancing Technologies (PETS)*, pages 299–318, 2012.

[110] Gerald R. McMenamin and Dongdoo Choi. *Forensic Linguistics: Advances in Forensic Stylistics.* CRC Press, London, 2002.

[111] Yashar Mehdad and Joel Tetreault. Do characters abuse more than words? In *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303, 2016.

[112] Thomas Corwin Mendenhall. The characteristic curves of composition. *Science*, IX:237–49, 1887.

[113] 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.

[114] Todor Mihaylov, Georgi D. Georgiev, and Preslav Nakov. Finding opinion manipulation trolls in news community forums. In *Proceedings of the 19th Conference on Computational Language Learning*, pages 310–314, 2015.

[115] Todor Mihaylov and Preslav Nakov. Hunting for troll comments in news community forums. In *The 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 399–405, 2016.

[116] Tsvetomila Mihaylova, Georgi Karadjov, Preslav Nakov, Yasen Kiprov, Georgi Georgiev, and Ivan Koychev. SU@PAN'2016: Author Obfuscation – Notebook for PAN at CLEF 2016. In Krisztian Balog, Linda Cappellato, Nicola Ferro, and Craig Macdonald, editors, *CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers*, 2016.

[117] George A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[118] Frederick Mosteller and David L. Wallace. *Inference and disputed authorship: The Federalist.* Addison-Wesley, 1964.

[119] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Nathan S. Glance. What Yelp fake review filter might be doing? In *Proceedings of the Seventh International Conference on Weblogs and Social Media (ICWSM-2013)*, pages 409–418, 2013.

[120] Shashi Narayan and Claire Gardent. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 435–445, 2014.

[121] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*, pages 300–314, 2012.

[122] Kazuyuki Narisawa, Hideo Bannai, Kohei Hatano, and Masayuki Takeda. Unsupervised spam detection based on String Alienness Measures. *Discovery Science*, pages 161–172, 2007.

[123] Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. Surveying stylometry techniques and applications. *ACM Computing Surveys*, 50(6):86:1–86:36, 2017.

[124] Matthew L. Newman, James W. Pennebaker, Diane S. Berry, and Jane M. Richards. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675, 2003.

[125] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153, 2016.

[126] Ray Oshikawa, Jing Qian, and William Yang Wang. A survey on natural language processing for fake news detection. *CoRR*, abs/1811.00770, 2018.

[127] Ricardo Otheguy, Ofelia García, and Wallis Reid. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, 6(3):281–307, 2015.

[128] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*, pages 309–319, 2011.

[129] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Negative deceptive opinion spam. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 497–501, 2013.

[130] Rebekah Overdorf and Rachel Greenstadt. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. In *Proceedings on Privacy Enhancing Technologies*, pages 155–171, 2016.

[131] James W. Pennebaker, Roger J. Booth, and Martha E. Francis. Linguistic Inquiry and Word Count (LIWC): LIWC2007. Technical report, LIWC.net, Austin, Texas, 2007.

[132] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, 2018.

[133] Juan-Pablo Posadas-Duran, Grigori Sidorov, and Ildar Batyrshin. Complete Syntactic N-grams as Style Markers for Authorship Attribution. In Alexander Gelbukh, Félix Castro Espinoza, and Sofía N. Galicia-Haro, editors, *Human-Inspired Computing and Its Applications*, pages 9–17. Springer International Publishing, Cham, 2014.

[134] Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. Who wrote the web? revisiting influential author identification research applicable to information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, pages 393–407. Springer International Publishing, 2016.

[135] Martin Potthast, Matthias Hagen, and Benno Stein. Author obfuscation: Attacking the state of the art in authorship verification. In *CLEF 2016 Working Notes*, 2016.

[136] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style Transfer Through Back-Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 866–876, 2018.

[137] Ella Rabinovich, Shachar Mirkin, Raj Nath Patel, Lucia Specia, and Shuly Wintner. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1074–1084, 2016.

[138] Roshan Ragel, Pramod Herath, and Upul Senanayake. Authorship Detection of SMS Messages Using Unigrams. In *Eighth IEEE International Conference on Industrial and Information Systems*, pages 387–392, 2013.

[139] Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers (ACLShort'10)*, pages 38–42, 2010.

[140] Ganesh Ramakrishnan, B. Prithviraj, and Pushpak Bhattacharyya. A Gloss Centered Algorithm for Word Sense Disambiguation. In *Proceedings of the ACL SENSEVAL*, pages 217–221, 2004.

[141] Congzhou He Ramyaa and Khaled Rasheed. Using machine learning techniques for stylometry. In *Proceedings of the International Conference on Artificial Intelligence (IC-AI'04)*, volume 2, pages 897–903, 2004.

[142] Josyula R. Rao and Pankaj Rohatgi. Can pseudonymity really guarantee privacy? In Steven M. Bellovin and Gregory G., editors, *9th USENIX Security Symposium*, 2000.

[143] Sudha Rao and Joel R. Tetreault. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 129–140, 2018.

[144] Shebuti Rayana and Leman Akoglu. Collective opinion spam detection: Bridging review networks and metadata. In Steven M. Bellovin and Gregory G., editors, *Proceeding of the 21st ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'15)*, 2015.

[145] Paul Rayson, Andrew Wilson, and Geoffrey Leech. Grammatical word class variation within the british national corpus sampler. *Language and Computers*, 36(1):295–306, 2001.

[146] Tapani Rinta-Kahila and Wael Soliman. Understanding crowdturfing: The different ethicallogics behind the clandestine industry of deception. In *Proceedings of the 25th European Conference on Information Systems (ECIS)*, pages 1934–1949, 2017.

[147] Victoria L. Rubin. Deception Detection and Rumor Debunking for Social Media. In Luke Sloan and Anabel Quan-Haase, editors, *The SAGE Handbook of Social Media Research Methods*. SAGE, London, 2017.

[148] Joseph Rudman. The state of authorship attribution studies: some problems and solutions. *Computers and the humanities,*, 31(4):351–365, 1998.

[149] Joseph Rudman. The state of non-traditional authorship studies - 2010: Some problems and solutions. In *Proceedings of the Digital Humanities*, pages 217–219, 2010.

[150] Edward Sapir. Speech as a personality trait. *American Journal of Sociology*, 32(6):892–905, 1927.

[151] Saul Schleimer, Daniel S. Wilkerson, and Alex Aiken. Winnowing: local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*, pages 76–85, 2003.

[152] Anna Schmidt and Michael Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, 2017.

[153] Chun Wei Seah, Hai Leong Chieu, Kian Ming A. Chai, Loo-Nin Teow, and Lee Wei Yeong. Troll Detection by Domain-Adapting Sentiment Analysis. In *Proceedings of the 18th International Conference on Information Fusion*, pages 792–799, 2015.

[154] Gün R. Semin and Klaus Fiedler. The linguistic category model, its bases, applications and range. *European Review of Social Psychology*, 2(1):1–30, 1991.

[155] Rico Sennrich, Barry Haddow, and Alexandra Birch. Controlling politeness in neural machine translation via side constraints. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 35–40, 2016.

[156] Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. Style transfer from non-parallel text by cross-alignment. In *Proceedings of Neural Information Processing Systems NIPS*, 2017.

[157] Rakshith Shetty, Bernt Schiele, and Mario Fritz. A^{4}nt: Author attribute anonymity by adversarial training of neural machine translation. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 1633–1650, 2018.

[158] Advaith Siddharthan. Complex lexico-syntactic reformulation of sentences using typed dependency representations. In *Proceedings of the 6th International Natural Language Generation Conference*, pages 125–133, 2010.

[159] Advaith Siddharthan. Text Simplification using Typed Dependencies: A Comparision of the Robustness of Different Generation Strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11, 2011.

[160] Edgar A. Smith and R.J. Senter. Automated readability index. Technical Report AMRL-TR-66-22, Aerospace Medical Division, Wright-Paterson AFB, Ohio, 1967.

[161] Thamar Solorio, Ragib Hasan, and Mainul Mizan. A Case Study of Sockpuppet Detection in Wikipedia. In *Proceedings of the Workshop on Language in Social Media*, pages 59–68, 2013.

[162] Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition, Second Edition*. Blackwell Publishers, Oxford/Cambridge, 1995.

[163] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[164] K. Surendran, O.P. Harilal, Hrudya Poroli, Prabaharan Poornachandran, and N.K. Suchetha. Stylometry detection using deep learning. In *Computational Intelligence in Data Mining*, pages 749–757, 2017.

[165] 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.

[166] Catalina L. Toma and Jeffrey T. Hancock. What lies beneath: The linguistic traces of deception in online dating profiles. *Journal of Communication*, 62(1):78–97, 2012.

[167] Takashi Uemura, Daisuke Ikeda, Takuya Kida, and Hiroki Arimura. Unsupervised spam detection by document probability estimation with Maximal Overlap Method. *Information and Media Technologies*, 6(1):231–240, 2011.

[168] Hans van Halteren, R. Harald Baayen, Fiona Tweedie, Marco Haverkort, and Anneke Neijt. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics*, 12(1):65–77, 2005.

[169] Gand Wang, Christo Wilson, Xiaohan Zhao, Yibo Zhu, Manish Mohanlal, Haitao Zheng, and Ben Y. Zhao. Serf and turf: crowdturfing for fun and profit. In *Proceedings of the 21st international conference on World Wide Web (WWW)*, pages 679–688, 2012.

[170] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, 2016.

[171] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, ÅĄukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144, 2016.

[172] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual meeting of the Associations for Computational Linguistics (ACL)*, pages 133–138, 1994.

[173] Sander Wubben, Antal van den Bosch, and Emiel Krahmer. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1015–1024, 2012.

[174] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399, 2017.

[175] Jun-Ming Xu, Xiaojin Zhu, and Amy Bellmore. Fast learning for sentiment analysis on bullying. In *Proceedings of the International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM'12)*, pages 1–6, 2012.

[176] Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. Paraphrasing for style. In *Proceedings of COLING*, pages 2899–2914, 2012.

[177] Yinqing Xu, Bei Shi, Wentao Tian, and Wai Lam. A unified model for unsupervised opinion spamming detection incorporating text generality. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 725–731, 2015.

[178] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS'17)*, pages 1143–1158, 2017.

[179] Kyung-Hyan Yoo and Ulrike Gretzel. Comparison of deceptive and truthful travel reviews. In *Information and Communication Technologies in Tourism 2009: Proceedings of the International Conference*, pages 37–47, Vienna, 2009. Springer Verlag.

[180] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709, 2017.

[181] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Proceedings of ESWC*, pages 745–760, 2018.

[182] Ying Zhao and Justin Zobel. Effective and scalable authorship attribution using function words. In *Information Retrieval Technology*, pages 174–189, 2005.

[183] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. A framework of authorship identification for online messages: Writing style features and classification techniques. *Journal American Society for Information Science and Technology*, 57(3):378–393, 2006.

[184] Lina Zhou, Judee K. Burgoon, Jay F. Nunamaker Jr, and Doung P. Twitchell. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communication. *Group Decision and Negotiation*, 13:81–106, 2004.

[185] Lina Zhou, Judee K. Burgoon, Doug P. Twitchell, Tiantian Qin, and Jay F. Nunamaker Jr. A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, 20:139–163, 2004.

[186] Lina Zhou, Judee K. Burgoon, and Douglas P. Twitchell. A longitudinal analysis of language behavior of deception in e-mail. In Hsinchun Chen, Richard Miranda, Daniel D. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, editors, *Intelligence and Security Informatics*, pages 102–110. Springer Verlag, Berlin Heidelberg, 2010.

[187] Lina Zhou, Douglas P. Twitchell, Tiantian Qin, Judee K. Burgoon, and Jay F. Nunamaker Jr. An exploratory study into deception detection in text-based computer mediated communication. In *Proceedings of the 36th Hawaii Intl Conference on Systems Science*, 2003.

[188] Yan Zhou, Zach Jorgensen, and W. Meador Inge. Combating good word attacks on statistical spam filters with multiple instance learning. In *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, pages 298–305, 2007.