Adversarial Stylometry in the Wild:
Transferable Lexical Substitution Attacks on Author Profiling

Chris Emmery
CSAI, Tilburg University
CLiPS, University of Antwerp
cmry@pm.me

Ákos Kádár
Borealis AI
akos.kadar@borealisai.com

Grzegorz Chrupała
CSAI, Tilburg University
g.a.chrupala@uvt.nl

Abstract

Written language contains stylistic cues that can be exploited to automatically infer a variety of potentially sensitive author information. Adversarial stylometry intends to attack such models by rewriting an author's text. Our research proposes several components to facilitate deployment of these adversarial attacks in the wild, where neither data nor target models are accessible. We introduce a transformer-based extension of a lexical replacement attack, and show it achieves high transferability when trained on a weakly labeled corpus—decreasing target model performance below chance. While not completely inconspicuous, our more successful attacks also prove notably less detectable by humans. Our framework therefore provides a promising direction for future privacy-preserving adversarial attacks.

1 Introduction

The widespread use of machine learning on consumer devices and its application to their data has sparked investigation of security and privacy researchers alike in correctly handling sensitive information (Edwards and Storkey, 2016; Abadi et al., 2016b). Natural Language Processing (NLP) is no exception (Fernandes et al., 2019; Li et al., 2018); written text can contain a plethora of author information—either consciously shared or inferable through stylometric analysis (Rao et al., 2000; Adams, 2006). This characteristic is fundamental to author profiling (Koppel et al., 2002), and while the field's main interest pertains to the study of sociolinguistic and stylometric features that underpin our language use (Daelemans, 2013), herein simultaneously lie its dual-use problems. Author profiling can, often with high accuracy, infer an extensive set of (sensitive) personal information, such as age, gender, education, socio-economic status, and mental health issues (Eisenstein et al., 2011; Alowibdi et al., 2013; Volkova et al., 2014; Plank and Hovy, 2015; Volkova and Bachrach, 2016). It therefore potentially exposes anyone sharing written online content to unauthorized information collection through their writing style. This can prove particularly harmful to individuals in a vulnerable position regarding e.g., race, political affiliation, or mental health.

Privacy-preserving defenses against such inferences can be found in the field of adversarial[1] stylometry. Our research[2] concerns the obfuscation subtask, where the aim is to rewrite an input text such that the style changes, and stylometric predictions fail. It is part of a growing body of research into adversarial attacks on NLP (Smith, 2012), which various modern models have proven vulnerable to; e.g., in neural machine translation (Ebrahimi et al., 2018), summarization (Cheng et al., 2020), and text classification (Liang et al., 2018).

Adversarial attacks on NLP are predominantly aimed at demonstrating vulnerabilities in existing algorithms or models, such that they might be fixed, or explicitly improved through adversarial training. Consequently, most related work focuses on white or black-box settings, where all or part of the target model is accessible (e.g., its predictions, data, parameters, gradients, or probability distribution) to fit an attack. The current research, however, does not intend to improve the targeted models; rather, we want to provide the attacks as tools to protect online privacy. This introduces several constraints over other NLP-based adversarial attacks, as it calls for a realistic, in-the-wild scenario of application.

Firstly, authors seeking to protect themselves from stylometric analysis cannot be assumed to be

[1] These are adversarial attacks on models making stylometric predictions, not to be confused with adversarial learning.

[2] All code, data, and materials to fully reproduce the experiments are openly available at https://github.com/cmry/reap.

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 2388–2402
April 19 - 23, 2021. ©2021 Association for Computational Linguistics

knowledgeable about the target architecture, nor to have access to suitable training data (as the target could have been trained on any domain). Hence, we cannot optimally tailor attacks to the target, and need an accessible method of mimicking it to evaluate the obfuscation success. To facilitate this, we use a so-called substitute model, which for our purposes is an author profiling classifier trained in isolation (with its own data and architecture) that informs our attacks. Attacks fitted on substitute models have been shown to transfer their success when targeting models with different architectures, or trained on other data, in a variety of machine learning tasks (Papernot et al., 2016). The effectiveness of an attack fitted on a substitute model when targeting a 'real' model is then referred to as *transferability*, which we will measure for the obfuscation methods proposed in the current research.

Secondly, for an obfuscation attack to work in practice (e.g., given a limited post history), it should suggest relevant changes –to– the author's writing *on a domain of their choice*. This implies the substitute models should be fitted locally, and therefore need to meet two criteria: reliable access to labeled data, and being relatively fast and easy to train. To meet the first criterion, the current research focuses on gender prediction, as: i) Twitter corpora annotated with this variable are by far the largest (and most common), ii) author profiling methods typically use similar architectures for different attributes; therefore, the generalization of attacks to other author attributes can be assumed to a large extent, and, most importantly, iii) Beller et al. (2014) and Emmery et al. (2017) have shown that through distant labeling, a representative corpus for this task can be collected in under a day. This allows us to measure transferability of attacks fitted using realistically collected distant corpora to models using high-quality hand labeled corpora.

As for the attacks, we focus on lexical substitution of content words strongly related to a given label, as those have been shown to explain a significant portion of the accuracy of stylometric models (see e.g., Rao et al., 2000; Burger et al., 2011; Sap et al., 2014; Rangel et al., 2016). To that effect, we extend the substitution attack of Jin et al. (2020) and apply it to author attribute obfuscation. Specifically, we explore the potential of training a simple (as to meet the speed criterion), non-neural substitute model f' to indicate relevant words to perturb, where retaining the original meaning is prioritized.

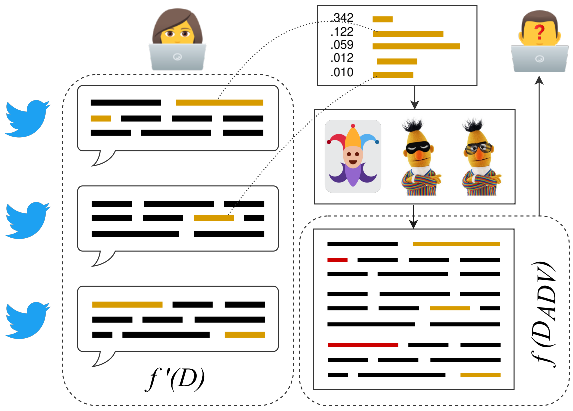

Figure 1: Obfuscation scenario: model f' trains on tweet batches, an omission score is used to determine and rank the words according to their classification contribution. These are then passed to either TextFooler, Masked BERT, or Dropout BERT to suggest top-k replacement candidates. From these, a selection is made based on their class probability change on $f'(D)$. Finally, f is evaluated on the perturbed tweets D_{ADV}.

Two transformer-based models are introduced to the framework to propose and rank lexical substitutions towards a change in the predictions of f'. We evaluate if the attacks on f' transfer across corpora, architectures, and a separately trained target model f (see Figure 1). Finally, we measure the quality of changes using automatic evaluation metrics, and conduct an human evaluation that focuses on detection accuracy of the attacks.

2 Related Work

Stylometry, the study of (predominantly) writing style, dates back several decades (Mosteller and Wallace, 1963), and has seen increased accessibility through the introduction of statistical models (see surveys by Holmes, 1998; Neal et al., 2017) and machine learning (e.g., Matthews and Merriam, 1993; Merriam and Matthews, 1994). Computational stylometry distinguishes several subtasks such as determining (Baayen et al., 2002) and verifying author identity (Koppel and Schler, 2004), and author profiling (Argamon et al., 2005); e.g., predicting demographic attributes. Adversarial stylometry (as conceptualized by Brennan et al., 2012) intends to subvert these inferences by changing an author's text through imitation, or, as pertains to our research, the obfuscation of writing style (Kacmarcik and Gamon, 2006; Caliskan et al., 2018; Le et al., 2015; Xu et al., 2019).

These changes, or perturbations, can be produced in several ways, and the task is therefore often conflated with paraphrasing (Reddy and Knight, 2016), style transfer (Kabbara and Cheung, 2016), and generating adversarial samples or triggers (Zhang et al., 2020b). Regardless of the employed method, the main challenge of obfuscation lies in retaining the original meaning of an input text; its written language medium limits any perturbations to discrete outputs, and unnatural discrepancies are significantly better discernible by humans than, say, a few pixel changes in an image. An additional, persistent limitation is the absence of evaluation metrics that guarantee complete preservation of the original meaning of the input whilst changes remain unnoticed (Potthast et al., 2016). This not only inhibits automatic evaluation of obfuscation, but all natural language generation research (Novikova et al., 2017)—placing an emphasis on human evaluation (van der Lee et al., 2019).

It is perhaps for this reason that most obfuscation work uses heuristically-driven, controlled changes such as splitting or merging words or sentences, removing stop words, changing spelling, punctuation, or casing (see e.g., Karadzhov et al., 2017; Eger et al., 2019). These specific attacks are typically easier to mitigate through preprocessing (Juola and Vescovi, 2011). Obfuscation through lexical substitution (Mansoorizadeh et al., 2016; Bevendorff et al., 2019, 2020) provides a middle ground of control, semantic preservation and attack effectiveness; however, they might prove less effective against models relying on deeper stylistic features (e.g. word order, part-of-speech (POS) tags, or reading complexity scores). End-to-end systems have been employed for similar purposes (Shetty et al., 2018; Saedi and Dras, 2020), or to rewrite entire phrases (Emmery et al., 2018; Bo et al., 2019) using (adversarially-driven) autoencoders. Such attacks seem less common, and provide less control over the perturbations and semantic consistency.

Our work does not assume the attacks to run end-to-end, but with a hypothetical human in the loop. We further opt for techniques that are more likely to find strong semantic mirrors to the original text while making minimal changes. A substitute model (the algorithm, hyper-parameters, and output of which an author can manipulate as desired) is employed to indicate candidate replacement words, and our attacks suggest and rank those against this substitute. Moreover, prior work typically attacks adversaries trained on the same data, whereas we add a transferability measure. Lastly, while author identification has been investigated in the wild (Stolerman et al., 2013), our work is, to our knowledge, the first to make a conscious effort towards realistic applicability of obfuscation techniques.

3 Method

Our attack framework extends TextFooler (TF, Jin et al., 2020) in several ways. First, a substitute gender classifier is trained, from which the logit output given a document is used to rank words by their prediction importance through an omission score (Section 3.1). For the top most important words, substitute candidates are proposed, for which we add two additional techniques (Section 3.2). These candidates can be checked and filtered on consistency with the original words (by their POS tags, for example), accepted as-is, or re-ranked (Section 3.3). For the latter, we add a scoring method. Finally, the remaining candidates are used for iterative substitution until TF's stopping criterion is met (i.e., the prediction changes, or candidates run out).

3.1 Target Word Importance

We are given a target classifier f, substitute classifier f', a document D consisting of tokens D_i, and a target label y. Here, f' is trained on some corpus X, and receives an author's new input text D, where the author provides label y. We denote a class label as $\bar{y}$ if $f'(D)$ predicts anything but y. Our perturbations form adversarial input D_{ADV}, that intends to produce $f'(D_{\text{ADV}}) = \bar{y}$, and thereby implicitly $f(D_{\text{ADV}}) = \bar{y}$. Note that we only submit D to f for evaluating the attack effectiveness, and it is never used to fit the attack itself.

To create D_{ADV}, a minimum number of edits is preferred, and thus we rank all words in D by their omission score (similar to e.g., Kádár et al., 2017) according to f' (`omission_score` in Algorithm 1). Let $D_{\setminus i}$ denote the document after deleting D_i, and $o_y(D)$ the logit score by f'. The omission score is then given by $o_y(D) - o_y(D_{\setminus i})$, and used in an importance score I of token D_i, as:

$$I_{D_i} = \begin{cases} o_y(D) - o_y(D_{\setminus i}), \\ \quad \text{if } f'(D) = f'(D_{\setminus i}) = y. \\ o_y(D) - o_y(D_{\setminus i}) + o_{\bar{y}}(D) - o_{\bar{y}}(D_{\setminus i}), \\ \quad \text{if } f'(D) = y, f'(D_{\setminus i}) = \bar{y}, y \neq \bar{y}. \end{cases} \tag{1}$$

With I_{D_i} calculated for all words in D, the top k ranked tokens are chosen as target words T.

ALGORITHM 1: Obfuscation by lexical replacement.

```
Input  :  $f'$ – substitute model
          $D = \{w_0, w_1, \ldots, w_n\}$ – document
          $y$ – target label
          checks – apply checks (bool)
          k – target max $k$-amount words
Output :  $D_{\text{ADV}}$ – obfuscated document
for $D_i \in D$ do
    // via Equation 1
    $I_{D_i} \leftarrow$ omission_score($f', y$)
$T \leftarrow$ top_k(argsort_desc($D, I_{D_i}$ scores), k)
$D_{\text{ADV}} = D$
for $t \in T$ do
    // substitution attack on $t$
    $C_t \leftarrow$candidates($t$)
    $A = (D_{\text{ADV}\,1:i-1}, C_{t,j}, D_{\text{ADV}\,i+1:n})_{1 \leq j \leq |C_t|}$
    $\bar{A} =$ filter/rank($D, A; t$, checks)
    // test attack success on $f'$
    for $D' \in \bar{A}$ do
        if arg max $o_y(D') \neq y$ then
            return $D_{\text{ADV}} = D'$
        else if $o_y(D') < o_y(D_{\text{ADV}})$ then
            $t$ in $D_{\text{ADV}}$ is replaced with $c$ from $D'$
return $D_{\text{ADV}}$
```

3.2 Lexical Substitution Attacks

Four approaches to perturb a target word $t \in T$ are considered in our experiments. These operations are referred to as `candidates` in Algorithm 1.

Synonym Substitution (WS) This TF-based substitution embeds t as $\boldsymbol{t}$ using a pre-trained embedding matrix $\boldsymbol{V}$. C_t is selected by computing the cosine similarity between $\boldsymbol{t}$ and all available word-embeddings $\boldsymbol{w} \in \boldsymbol{V}$. We denote cosine similarity with $\Lambda(\boldsymbol{t}, \boldsymbol{w})$. A threshold δ is used to keep only reliable candidates $\Lambda(\boldsymbol{t}, \boldsymbol{w}) > \delta$.

Masked Substitution (MB) The embedding-based substitutions can be replaced by a language model predicting the contextually most likely token. BERT (Devlin et al., 2019)—a bi-directional encoder (Vaswani et al., 2017) trained through masked language modeling and next-sentence prediction—makes this fairly trivial. By replacing t with a mask, BERT produces a top-k most likely C_t for that position. Implementing this in TF does imply each previous substitution of t might be included in the context of the current one. This method of contextual replacement has two drawbacks: i) semantic consistency with the *original* word is not guaranteed (as the model has no knowledge of t), and ii) the replaced context means semantic drift can occur, as all subsequent substitutions follow the new, possibly incorrect context.

Dropout Substitution (DB) A method to circumvent the former (i.e., BERT's masked prediction limitations for lexical substitution), was presented by Zhou et al. (2019). They apply dropout (Srivastava et al., 2014) to BERT's internal embedding of target word t before it is passed to the transformer—zeroing part of the weights with some probability. The assumption is that C_t (BERT's top-k) will contain candidates closer to the original t than the masked suggestions.

Heuristic Substitution To evaluate the relative performance of the techniques we described before, we employ several heuristic attacks as baselines. In the order of Table 3: 1337-speak: converts characters to their leetspeak variants, in a similar vein to e.g. diacritic conversion (Belinkov and Bisk, 2018). Character flip: inverts two characters in the middle of a word, which was shown to least affect readability (Rayner et al., 2006). Random spaces: splits a token into two at a random position.

3.3 Candidate Filtering and Re-ranking

Given C_t, either all, or only the highest ranked candidate can be accepted as-is. Alternatively, all D' can be filtered by submitting them to checks, or re-ranked based on their semantic consistency with D. These operations are referred to as `rank/filter` in Algorithm 1—both of which can be executed.

Part-of-Speech and Document Encoding TF employs two checking components: first, it removes any c that has a different POS tag than t. If multiple D' exist so that $f'(D') = \bar{y}$, it selects the document D' which has the highest cosine similarity to the Universal Sentence Encoder (USE) embedding (Cer et al., 2018) of the original document D. If not, the D' with the *lowest* target word omission score is chosen (as per TF's method).

BERT Similarity Zhou et al. (2019) use the concatenation of the last four layers in BERT as a sentence's contextualized representation $\boldsymbol{h}$. We apply this in both Masked (MB) and Dropout (DB) BERT to re-rank all possible D' by embedding them. Given document D, target t, and perturbation candidate document D', C_t would be ranked via an embedding similarity score:

$$\textsc{sim}\left(D, D'; t\right) = \sum_{i}^{n} w_{i,t} \times \Lambda\left(\boldsymbol{h}\left(D_i|D\right), \boldsymbol{h}(D'_i|D')\right) \qquad (2)$$

	AUTHORS	TWEETS	FEMALE	MALE	TRAIN	TEST	TOKENS	TYPES	AVG SIZE
Huang et al.	37,929	47,211	26,758	20,453	30,602	7,651	935,062	46,600	28
Emmery et al.	6,610	16,788,612	61,736	32,900	75,918	18,718	146,736,657	9,942,399	301
Volkova et al.	4,620	12,226,859	32,376	26,708	47,298	11,777	67,186,535	7,836,539	269

Table 1: Corpus statistics indicating the number of authors, tweets, female and male labels, the size of the train and test splits, number of types (unique words) and tokens (total words), and average tokens per document (avg size).

where $\boldsymbol{h}\left(D_i|D\right)$ is BERT's contextualized representation of the i^{th} token in D, and $w_{i,t}$ is the average self-attention score of all heads in all layers ranging from the i^{th} token with respect to t in D.[3]

4 Experiment

4.1 Data

We use three author profiling sets (see Table 1 for statistics) that are annotated for binary gender classification (male or female): first, that of Volkova et al. (2015) which was collected through annotating 5,000[4] English Twitter profiles by crowdsourcing via Mechanical Turk. This can be considered a 'random' sample of Twitter profiles, and is therefore the most unbiased set of the three. Hence, we consider it the most representative of an author profiling set, and employ this as training split (80%) for f, and test split for our attacks (20%).

The second is the English portion of the Multilingual Hate Speech Fairness corpus of Huang et al. (2020), which was collected with a different objective than author profiling. It was aggregated from existing hate speech corpora (by Waseem and Hovy, 2016; Waseem, 2016; Founta et al., 2018)—which were largely bootstrapped with lookup terms, selection of frequently abusive users, etc.—and annotated post-hoc with demographic information. The collection did not focus on profiles, and most authors are only associated with a single tweet. This can cause a significant domain shift compared to general author profiling. However, it can be seen as freely available (noisy) data.

Lastly, we include a weakly labeled author profiling corpus by Emmery et al. (2017), collected through English keyword look-up for self-reports—similar to Beller et al. (2014). This corpus likely includes incorrect labels, but was collected in less than a day, making it an ideal candidate for realistic access to (new) data to fit the substitute model.

Preprocessing & Sampling All three corpora were tokenized using spaCy[5] (Honnibal and Montani, 2017). Other than lowercasing, allocating special tokens to user mentions and hashtags (# and text were split), and URL removal, no additional preprocessing steps were applied. Every author timeline was divided into chunks for a maximum of 100 tweets (i.e., some contain less) to form our documents, implying a maximum of 25 instances per author (some contain one, 2,500 is the API history limit). From the test set, the last[6] 200 instances were sampled for the attack (110 male, 90 female). While fairly small, this sample does reflect a realistic attack duration and timeline size, as they would be executed for a single profile.

4.2 Attacks

For the extension of TF, we re-implemented code[7] by Jin et al. (2020) to work with Scikit-learn[8] (Pedregosa et al., 2011). For their synonym substitution component, we similarly used counter-fitted embeddings by Mrkšić et al. (2016) trained on Simlex-999 (Hill et al., 2015). The USE (Cer et al., 2018) implementation uses TensorFlow[9] (Abadi et al., 2016a) as back-end, and all BERT-variants were implemented in Hugging Face's[10] Transformers library (Wolf et al., 2020) with PyTorch[11] (Paszke et al., 2019) as back-end.

We adopt the same parameter settings as Jin et al. (2020) throughout our TF experiments: they set N (considered synonyms) and δ (cosine similarity minimum) empirically to 50 and 0.7 respectively. For MB and DB, we capped T at 50 and top-k at 10 (to improve speed). For DB, we follow Zhou et al. (2019) and set the dropout probability to 0.3.

[3] Zhou et al. (2019) additionally use a proposal score for finding T that we replaced with the omission score.

[4] Profile counts in the current work differ due to collection limitations (e.g., removed accounts).

[5] https://spacy.io

[6] As the datasets are not shuffled to avoid overfitting on author-specific features, a few documents of the same author might spill from the train into the test split; this avoids incorporating those in our attack sample.

[7] https://github.com/jind11/TextFooler

[8] https://scikit-learn.org/

[9] https://tensorflow.org/

[10] https://huggingface.co/

[11] https://pytorch.org/

data	Huang, Emmery, Volkova
importance	Omission score
attack	Heuristics, TextFooler, Masked BERT, Dropout BERT
model	Logistic Regression, N-GrAM
ranking	None, POS + USE, BERT Sim

Table 2: Grid of possible experimental configurations.

4.3 Models

For f and f' we require (preferably fast) pipelines that achieve high accuracy on author profiling tasks, and are sufficiently distinct to gauge how well our attacks transfer across architectures, rather than solely across corpora. As state-of-the-art algorithms have not yet proven to be sufficiently effective for author profiling (Joo et al., 2019) we opt for common n-gram features and linear models.

Logistic Regression Logistic Regression (LR) trained on tf·idf using uni and bi-gram features proved a strong baseline in author profiling in prior work. The simplicity of this classifier also makes it a substitute model that can realistically be run by an author. No tuning was performed: C is set to 1.

N-GrAM The New Groningen Author-profiling Model (N-GrAM) from Basile et al. (2018), was proposed as a highly effective—simple—model that outperforms more complex (neural) alternatives on author profiling with little to no tuning. It uses tf·idf-weighted uni and bi-gram token features, character hexa-grams, and sublinearly scaled tf $(1 + \log(\text{tf}))$. These features are then passed to a Linear Support Vector Machine (Cortes and Vapnik, 1995; Fan et al., 2008), where $C = 1$.

4.4 Experimental Setup

To summarize (and see Table 2), the experiment is conducted as follows: the substitute target model (f')—LR for all experiments—is fit on a given corpus. The real target model (f, either LR or N-GrAM) is always fit on the corpus of Volkova et al. (2015). To evaluate the attacks, a 200-instance sample is used. Target words are ranked via omission scores from f', fed to either our Heuristics, TF, MB, or DB attacks. The heuristics directly change the target words, while the rest outputs a ranked set of replacement candidates. The latter can either be evaluated against f' through the TF pipeline, or the Top-1 candidate is returned. Filtering can be applied through POS/USE for semantic similarity and POS compatibility checks (Check), or not (~~Check~~).

		test = Volkova et al.					
	LR $f' \rightarrow$	Huang et al.		Emmery et al.		Volkova et al.	
	$f \rightarrow$	LR	NG	LR	NG	LR	NG
	none	.885	.940	.885	.940	.885	.940
Heuristic	1337	.770	.850	.775	.835	.715	.860
	flip	.900	.950	.885	.905	.840	.905
	space	.845	.925	.760	.870	.720	.850
Top-1	WS	.825	.930	.805	.890	.750	.915
	MB	.655	.905	.595	.785	.145	.410
	DB	.625	.895	.575	.785	.210	.530
~~Check~~	WS	.540	.855	.355	.670	**.000**	**.009**
	MB	**.415**	.790	**.120**	**.420**	**.000**	.085
	DB	.430	**.775**	.175	.430	**.000**	.085
Check	TF	.705	.920	.780	.910	.375	.700
	TF + MB	.640	.880	.760	.890	.380	.725
	TF + DB	.650	.885	.755	.890	.435	.715

Table 3: Post-attack accuracy scores (below chance (55%) = better) of f on a test sample from the Volkova et al. corpus. Left, the attack conditions: heuristics, top-1 synonym, applying POS and USE similarity checks, or not applying those checks (~~Check~~). Splits per training corpus are noted for f' (always Logistic Regression (LR)). As target model, either LR, or N-GrAM (NG) was used. The substitution attacks are TextFooler (TF), Masked (MB) and Dropout BERT (DB). If TF's stopping criterion was used, TF + is noted. Word Similarity (WS), reflects the TF pipeline without checks.

Note that we are predominantly interested in transferability, and would therefore like to test as many combinations of data and architecture access limitations as possible. If we assume an author does not have access to the data, the substitute classifier is trained on any other data than the Volkova et al. corpus. If we assume the author does not know the target model architecture, the target model is N-GrAM (rather than LR). A full model transfer setting (in both data and architecture) will therefore be, e.g.: data f' = Emmery et al., data f = Volkova et al., f' = LR, and f = NGrAM. Finally, for comparison to an optimal situation, we test a setting where we do have access to the adversary's data.

4.5 Evaluation

Metrics The obfuscation success is measured as any accuracy score below chance level performance, which given our test sample is 55%. We would argue that random performance is preferred in scenarios where the prediction of the opposite label is undesired. For the current task, however, any accuracy drop to around or lower than chance level satisfies the conditions for successful obfus-

	Volkova et al. →	TRAIN	TEST
TRAIN	Huang et al.	0.640	0.620
	Emmery et al.	0.725	0.890

Table 4: Gender prediction accuracies of the substitute models f' on train and test splits of f.

cation.[12] To evaluate the semantic preservation of the attacked sentences, we calculate both METEOR (Banerjee and Lavie, 2005; Lavie and Denkowski, 2009) using `nltk`,[13] and BERTScore (Zhang et al., 2020a) between D and D_{ADV}. METEOR captures flexible uni-gram token overlap including morphological variants, and BERTScore calculates similarities with respect to the sentence context.

Human Evaluation For the human evaluation, we sampled 20 document pieces (one or more tweets) for each attack type in the best performing experimental configuration. A piece was chosen if it satisfied these criteria: i) contains changes for all three attacks, ii) consists of at least 15 words (excluding emojis and tags), and iii) does not contain obvious profanity.[14] All 60 document pieces of the three models were shuffled, and the 20 original versions were appended at the end (so that 'correct' pieces were seen last). Each substitute model therefore has 80 items for evaluation.

While in prior work it is common to rate semantic consistency, fluency, and label a text (see e.g., Potthast et al., 2016; Jin et al., 2020), our Twitter data are too noisy (including many spelling and grammar errors in the originals), and document batches too long to make this a feasible task. Instead, our six participants (three per substitute) were asked to indicate if: a) a sentence was artificially changed, and if so, b) indicate one word that raised their suspicion. This way, we can evaluate which attack produces the most natural sentences, and the least obvious changes to the input.

The items were rated individually; the human evaluators did not know beforehand that different versions of the same sentences were repeated, nor that the originals were shown at the end. All participants have a university-level education, a high English proficiency, and are familiar with the domain of the data. Several example ratings of the same sentence can be found in Table 6.

5 Results

5.1 Domain Shift

As we alluded to in Section 4.1, both corpora used to train our substitute models were in fact not reference corpora for author profiling, and can therefore be considered as suboptimal, disjoint domains. The Huang et al. corpus in particular shows a strong domain shift (see Table 4) for both training and test sets. The distantly labeled Emmery et al. corpus achieves 7.5% more accuracy on the train split of Volkova et al., and test performance is significantly higher (27%). We might therefore expect better obfuscation performance from the latter.

5.2 Baselines

The results for all attacks are shown in Table 3. Note that these are performances for f; therefore, when no attacks are applied (none), the performance for both substitute corpora stays the same (as those only influence the attacks). For the heuristic attacks, 1337 seems to make the more robust baseline; outperforming some of the other settings—even on transferability. A surface-level advantage is that this attack has a minor impact on readability (when applied conservatively) and does not change semantics; however, the heuristic attacks are fairly simple to mitigate in preprocessing (Juola and Vescovi, 2011) and through character features (as shown by the performance of the N-GrAM model). For transferability, we evidently need to do more than simply trying to convert words to be out-of-vocabulary (OOV) with noise. While it can be argued the heuristics could change all words, shifting everything OOV would not be robust; the target model side could easily spot the anomalous input and might act (e.g., reject) accordingly.

5.3 Attack Transferability

Transferability can be assessed by comparing the LR and N-GrAM (NG) columns. Globally it can be observed that the substitute models trained on the Emmery et al. corpus systematically outperform those trained on Huang et al.; both for the settings where the adversary's architecture is known (LR), and where it is unknown (NG). This matches our

[12] If an attack drops accuracy to 0%, this effectively flips (in case of a binary label) the label. This label might *also* be undesired by the author (e.g., being classified as having polar opposite political views). This implies the target model being maximally unsure about the classification is desirable.

[13] `https://www.nltk.org/_modules/nltk/translate/meteor_score.html`

[14] To avoid exposing the raters to overly toxic content, blatant examples were filtered using a keyword list. Some minor examples remained, for which we added a disclaimer.

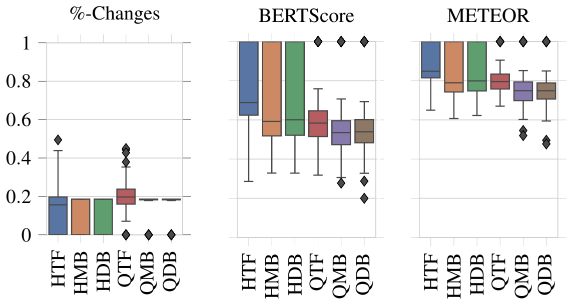

Figure 2: Rewrite metrics for the three different attacks: TextFooler (TF), Masked (MB) and Dropout BERT (DB)—split by corpora Huang et al. (H), Emmery et al. (Q). Shown are the relative number of changes (by document length), BERTScore (F_1), and METEOR (MET) with respect to the original document.

expectations from the observed domain shift. Our results also show that a noticeable decrease in obfuscation performance occurs (10-30% increased target model performance) when the attacks are transferred to different data and another model. In contrast, as can be observed from the last two columns in Table 3, in a practically unrealistic setting where the model and data are available, the obfuscation is specifically tailored to known weaknesses and therefore highly effective.[15]

5.4 Transformer Performance

Looking at the Top-1, ~~Check~~ and Check brackets (Table 3), other than the BERT-based models having higher success of transferability than TF, they also retain obfuscation success; deteriorating the target model's performance to lower than chance level (55%) for the settings not using additional checks. This comparison also demonstrates the synonym ranking to work (Top-1 vs. ~~Check~~ and Check), and the Check condition to be too restrictive; attaining lower attack power, and low transferability. This is further illustrated by the %-changes shown in Figure 2. Comparing the MB and DB variants, their performance seems almost identical, with masking having a slight advantage. As Zhou et al. (2019) argued, applying dropout should produce words that are closer to the original (compared to MB), which might affect obfuscation performance. Additionally, the BERT similarity ranking (described in Section 3.3) applied to the Masked substitution candidates could have some beneficial effect. This will have to be studied in more detail using the output evaluations.

		Huang et al.			Emmery et al.		
	ORG	TF	MB	DB	TF	MB	DB
ALTERED	.888	.967	.633	.783	.950	**.617**	.633
WORD	-	.950	.583	.700	.867	**.433**	**.433**

Table 5: Human accuracy scores of predicting if a text was altered, and guessing the attacked word (lower is better). All substitute models are those with the ~~Check~~ setting, trained on different corpora (i.e., different words are attacked per training corpus). ORG indicates correct prediction of the originals.

Rewrite Metrics The metrics in Figure 2 show a common initial limitation in their application to this task: the more frequent an attack makes no changes, the higher the automatic evaluation metrics (BERTScore, METEOR). Hence, to compare models, these scores need to be considered in light of the obfuscation performance, and related work. It can be observed that with consistently higher changes, MB and DB score lower on semantic consistency than TF. However, between MB and DB, and TF for the Emmery et al. corpus, these differences are minor. Furthermore, despite being fit on a different domain, these scores are comparable to prior obfuscation work (e.g., Shetty et al. (2018) show METEOR scores between 0.69 and 0.79).

Human Evaluation The results in Table 5 reflect the same trend that can be observed in Table 3; high obfuscation success seems to result in higher human error when predicting if a sentence was obfuscated. Conversely, it seems that despite higher semantic consistency scores, the original TF pipeline is easier to detect. This can be attributed to the number of spelling and grammar errors the model makes without its additional checks. Furthermore, the 11% error in identifying the original sentences also reflects some expected margin of error in this task, as our Twitter data is inherently noisy. Finally, while these results are in line with the obfuscation success, and are lower than detectability scores in related work (Mahmood et al., 2020), they also indicate that the models are still detectable above chance-level. Given three alternatives (including the original), performance should be 25% or lower to indicate no intrusive changes are made to text (that are not semantically coherent or not inconspicuous enough—both metrics used by Potthast et al., 2016). Therefore, while the presented approaches are effective, and realistically transferable, there is room for improvement for practical applicability.

[15] Jin et al. (2020) found similar drops to 0% accuracy with a comparable percentage of changes for word-level models.

ORG	ready to go home already . a better relationship with god 💙🙌🏼 . i need another job asap .
HTF	**loan** to go **houses** already . a **improved relations** with **jesus** 💙🙌🏼 . i **should** another **labour** asap .
HMB	ready to go **on** already . a better relationship with god 💙🙌🏼 . i need another **guy man** .
HDB	ready to go **somewhere** already . a better relationship with god 💙🙌🏼 . i need another **position vs** .
ORG	trump criticizes kim jong un after missile launch : ' does this guy have anything better to do ? ' .
HTF	**tramp criticized kam yung jt** after **rocket start** : ' does this **boyfriend** have anything **best** to do ? ' .
HMB	trump criticizes **ha woman congressman** after **campaign** launch : ' does this **book** have anything **else** to do ? ' .
HDB	trump criticizes **in at sin** after **bomb** launch : ' does this **kid** have anything **less** to do ? ' .

Table 6: Example ratings of different attacks (not shown together to the human evaluators) on two sentences with varying semantic consitency and human detection accuracy. In the first example, HMB was marked unaltered by all raters, HDB by the majority, and HTF by none. In the second, only HDB was marked unaltered, by only one rater. Attacked words are marked in bold, guessing any one of these would count as correctly identifying the attack.

6 Discussion and Future Work

We have demonstrated the performance of author attribute obfuscation under a realistic setting. Using a simple Logistic Regression model for candidate suggestion, trained on a weakly labeled corpus collected in a day, the attacks successfully transferred to different data and architectures. This is a promising result for future adversarial work on this task, and its practical implementation.

It remains challenging to automatically evaluate how invasive the required number of changes are for successful obfuscation—particularly to an author's message consistency as a whole. However, in practice such considerations could be left up to the author. In this human-in-the-loop scenario, a more extensive set of candidates could be suggested, and their effect on the substitute model shown interactively. This way, the attacks can be manually tuned to find a balance of effectiveness, inconspicuousness, and to guarantee semantic consistency. It would also show the author how their writing style affects potential future inferences.

Regarding the performance of the attacks: we demonstrated the general effectiveness of contextual language models in retrieving candidate suggestions. However, the quality of those candidates might be improved with more extensive rule-based checks; e.g., through deeper analyses using parsing. Nevertheless, such venues leave us with a core limitation of rewriting language, and therefore more broadly NLP: while the Masked attacks seemed more successful in our experiments, after manual inspection of the perturbations Dropout was found to often be semantically closer (see also Table 6)—which was not reflected in the human evaluation. This begs the question if *any* automated approach, evaluated under the current limitations of semantic consistency metrics, could realistically optimize for both obfuscation and inconspicuousness.

As such, we would argue that future work should focus on making as few perturbations as possible, retaining only the minimum amount of required obfuscation success. Given this, the other constraints become less relevant; one could generate short sentences (e.g., a single tweet) that might be semantically or contextually incorrect, but if it is a message in a long post history, it will hardly be detectable or intrusive. This would require certain triggers (as demonstrated by Wallace et al. (2019) for example), and ascertaining how well they transfer.

7 Conclusion

In our work, we argued realistic adversarial stylometry should be tested on transferability in settings where there is no access to the target model's data or architecture. We extended previous adversarial text classification work with two transformer-based models, and studied their obfuscation success in such a setting. We showed them to reliably drop target model performance below chance, though human detectability of the attacks remained above chance. Future work could focus on further minimizing this detection under our realistic constraints.

Acknowledgments

Our research strongly relied on openly available resources. We thank all whose work we could use. We would also like to thank the anonymous reviewers, Bertrand Higy, Bram Willemsen, and Chris van der Lee for their valuable feedback. Ákos Kádár contributed to this research independently. This work does not reflect Borealis AI's views nor any information Ákos Kádár may have learned while employed by Borealis AI.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016a. Tensorflow: a system for large-scale machine learning. In *Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation*, pages 265–283. USENIX Association.

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016b. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318.

Carlisle Adams. 2006. A classification for privacy techniques. *U. Ottawa L. & Tech. J.*, 3:35.

Jalal S Alowibdi, Ugo A Buy, and Philip Yu. 2013. Empirical evaluation of profile characteristics for gender classification on twitter. In *Machine Learning and Applications (ICMLA), 2013 12th International Conference on*, volume 1, pages 365–369. IEEE.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*, pages 1–16.

Harald Baayen, Hans van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6th JADT*, volume 1, pages 69–75.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Angelo Basile, Gareth Dwyer, Maria Medvedeva, Josine Rawee, Hessel Haagsma, and Malvina Nissim. 2018. Simply the best: minimalist system trumps complex models in author profiling. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 143–156. Springer.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*.

Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'ma belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 181–186.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108.

Janek Bevendorff, Tobias Wenzel, Martin Potthast, Matthias Hagen, and Benno Stein. 2020. On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification. *it-Information Technology*, 62(2):99–115.

Haohan Bo, Steven HH Ding, Benjamin Fung, and Farkhund Iqbal. 2019. Er-ae: differentially-private text generation for authorship anonymization. *arXiv preprint arXiv:1907.08736*.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Aylin Caliskan, Fabian Yamaguchi, Edwin Dauber, Richard E. Harang, Konrad Rieck, Rachel Greenstadt, and Arvind Narayanan. 2018. When coding style survives compilation: De-anonymizing programmers from executable binaries. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3601–3608. AAAI Press.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Walter Daelemans. 2013. Explanation in computational stylometry. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 451–462. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Javid Ebrahimi, Daniel Lowd, and Dejing Dou. 2018. On adversarial examples for character-level neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 653–663.

Harrison Edwards and Amos J. Storkey. 2016. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. In *Proceedings of NAACL-HLT*, pages 1634–1647.

Jacob Eisenstein, Noah A Smith, and Eric P Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 1365–1374. Association for Computational Linguistics.

Chris Emmery, Grzegorz Chrupała, and Walter Daelemans. 2017. Simple queries as distant labels for predicting gender on twitter. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 50–55.

Chris Emmery, Enrique Manjavacas, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 984–996.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing*, 13(3):111–117.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7.

Xiaolei Huang, Xing Linzi, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the Twelveth International Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, France. European Language Resources Association (ELRA).

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Youngjun Joo, Inchon Hwang, L Cappellato, N Ferro, D Losada, and H Müller. 2019. Author profiling on social media: An ensemble learning model using various features. *Notebook for PAN at CLEF*.

Patrick Juola and Darren Vescovi. 2011. Analyzing stylometric approaches to author obfuscation. In *IFIP International Conference on Digital Forensics*, pages 115–125. Springer.

Jad Kabbara and Jackie Chi Kit Cheung. 2016. Stylistic transfer in natural language generation systems using recurrent neural networks. In *Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods*, pages 43–47.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics.

Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780.

Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. 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. Springer.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the twenty-first international conference on Machine learning*, page 62.

Alon Lavie and Michael J Denkowski. 2009. The meteor metric for automatic evaluation of machine translation. *Machine translation*, 23(2-3):105–115.

Hoi Le, Reihaneh Safavi-Naini, and Asadullah Galib. 2015. Secure obfuscation of authoring style. In *IFIP International Conference on Information Security Theory and Practice*, pages 88–103. Springer.

Chris van der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4208–4215.

Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan. 2020. A girl has A name: Detecting authorship obfuscation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2235–2245. Association for Computational Linguistics.

Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. 2016. Author obfuscation using wordnet and language models—notebook for pan at clef 2016. In *CLEF 2016 Evaluation Labs and Workshop–Working Notes Papers*, pages 5–8.

Robert AJ Matthews and Thomas VN Merriam. 1993. Neural computation in stylometry i: An application to the works of shakespeare and fletcher. *Literary and Linguistic computing*, 8(4):203–209.

Thomas VN Merriam and Robert AJ Matthews. 1994. Neural computation in stylometry ii: An application to the works of shakespeare and marlowe. *Literary and Linguistic Computing*, 9(1):1–6.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–148.

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

Jekaterina Novikova, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *2017 Conference on Empirical Methods in Natural Language Processing*, pages 2231–2242. Association for Computational Linguistics.

Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Barbara Plank and Dirk Hovy. 2015. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98.

Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. In *CLEF (Working Notes)*, pages 716–749.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. *Working Notes Papers of the CLEF*.

Josyula R Rao, Pankaj Rohatgi, et al. 2000. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, pages 85–96.

K Rayner, SJ White, RL Johnson, and SP Liversedge. 2006. Raeding wrods with jubmled lettres: there is a cost. *Psychological science*, 17(3):192.

Sravana Reddy and Kevin Knight. 2016. Obfuscating gender in social media writing. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 17–26.

Chakaveh Saedi and Mark Dras. 2020. Large scale author obfuscation using Siamese variational auto-encoder: The SiamAO system. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 179–189, Barcelona, Spain (Online). Association for Computational Linguistics.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, David Stillwell, Michal Kosinski, Lyle Ungar, and H Andrew Schwartz. 2014. Developing age and gender predictive lexica over social media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1146–1151.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4nt: author attribute anonymity by adversarial training of neural machine translation. In *Proceedings of the 27th USENIX Conference on Security Symposium*, pages 1633–1650.

Noah A Smith. 2012. Adversarial evaluation for models of natural language. *arXiv preprint arXiv:1207.0245*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2013. Classify, but verify: Breaking the closed-world assumption in stylometric authorship attribution. In *IFIP Working Group*, volume 11, page 64.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010. Curran Associates Inc.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring perceived demographics from user emotional tone and user-environment emotional contrast. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring user political preferences from streaming communications. In *ACL (1)*, pages 186–196.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2153–2162.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Qiongkai Xu, Chenchen Xu, and Lizhen Qu. 2019. ALTER: Auxiliary text rewriting tool for natural language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 13–18, Hong Kong, China. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li. 2020b. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(3):1–41.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373.