Obfuscating Gender in Social Media Writing

Sravana Reddy
Wellesley College
Wellesley, MA
sravana.reddy@wellesley.edu

Kevin Knight
USC Information Sciences Institute
Marina del Rey, CA
knight@isi.edu

Abstract

The vast availability of textual data on social media has led to an interest in algorithms to predict user attributes such as gender based on the user's writing. These methods are valuable for social science research as well as targeted advertising and profiling, but also compromise the privacy of users who may not realize that their personal idiolects can give away their demographic identities. Can we automatically modify a text so that the author is classified as a certain target gender, under limited knowledge of the classifier, while preserving the text's fluency and meaning? We present a basic model to modify a text using lexical substitution, show empirical results with Twitter and Yelp data, and outline ideas for extensions.

1 Introduction

Recent work has demonstrated success in accurately detecting gender or other author attributes such as age, location, and political preferences from textual input, particularly on social media channels like Twitter (Bamman et al., 2014; Burger et al., 2011; Eisenstein et al., 2010; Li et al., 2014; Liu and Ruths, 2013; Pennacchiotti and Popescu, 2013; Rao et al., 2010; Volkova et al., 2015), weblogs (Mukherjee and Liu, 2010; Schler et al., 2006; Yan and Yan, 2006) and user-review sites (Johannsen et al., 2015).

Outside of academic research, detection of author attributes is a major component of "behavioral targeting" which has been instrumental in online advertising and marketing from the early days of the Web. Twitter, for example, uses gender inference over textual and profile features to serve ads (Underwood, 2012) and reports over 90% accuracy. Besides advertising, companies also rely on user profiling to improve personalization, build better recommender systems, and increase consumer retention.

While automatic profiling is undoubtedly valuable, it can also be used in ethically negative ways – the problem of "dual-use" outlined by Hovy and Spruit (2016). Users may wish to mask their demographic attributes for various reasons:

1. A by-product of personalization is inadvertent discrimination: a study (Datta et al., 2015) finds that Google serves fewer ads for high-paying jobs to users profiled as female, and Sweeney (2013) shows that ads for public data about people who are profiled as black are more likely to suggest an arrest record regardless of whether the person had one.

2. Users living under authoritarian governments have the incentive to conceal their identity for personal safety (Jardine, 2016). Even outside of repressive regimes, studies have shown that users value anonymity and are more likely to share controversial content when anonymous (Zhang and Kizilcec, 2014). This is evidenced by the popularity of anonymous-posting networks like Yik Yak and Whisper. Automated demographic profiling on content in these venues compromise this assumption of anonymity.

3. Many web users are concerned about online privacy. A large number choose to opt-out of

Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science, pages 17–26,
Austin, TX, November 5, 2016. ©2016 Association for Computational Linguistics

having their online activities tracked by blocking cookies, or installing blocking tools such as Do Not Track[1] or AdBlock Plus[2].

Turow et al. (2015) argue that the majority of users are not actually willing to compromise their privacy in order to receive benefits – rather, they are resigned to it because they believe they are powerless to limit what companies can learn about them. It is likely that a usable tool that aids in masking their demographic identity would be adopted, at least by privacy-conscious users.

4. Users may wish to conceal aspects of their identity to maintain authority or avoid harassment – some women on online forums will try to come across as male (Luu, 2015), and many female writers in literature have used male pseudonyms for this purpose.

This paper is a study addressing the following question: can we automatically modify an input text to "confound" a demographic classifier? The key challenge here is to transform the text while minimally distorting its meaning and fluency from the perspective of a human reader.

Consider this extract from a tweet:

`OMG I'm sooooo excited!!!`

Most classifiers would infer the author is female due to the use of multiple exclamation marks, the word *omg*, and the lengthening intensifier, features that are particularly gendered. Re-wording the tweet to

`dude I'm so stoked.`

conveys same message, but is more likely to be classified as male due to the words *dude* and *stoked* and the absence of lengthening and exclamation marks.

Although any distortion of text loses information (since word usage and punctuation are signals too), some of these stylistic features may be unintentional on the part of a user who isn't aware that this information can be used to profile or identify them.

2 Related Work

The most relevant existing work is that of Brennan et al. (2012) who explore the related problem of modifying text to defeat authorship detectors. Their program, Anonymouth (McDonald et al., 2012)[3], aids a user who intends to anonymize their writing relative to a reference corpus of writing from the user and other authors. Rather than automatically modifying the text, the program makes suggestions of words to add or remove. However, no substitutions for deleted words or placement positions for added words are suggested, so incorporating or removing specific words without being presented with alternatives requires a great deal of effort on the user's side. They also experiment with foiling the authorship detector with machine translation (by translating the text from English to German or Japanese and back to English), but report that it is not effective. Anonymouth is part of a larger field of research on "privacy enhancing technologies" which are concerned with aiding users in masking or hiding private data such as Google Search histories or network access patterns.

Another closely-related paper is that of Preotiuc-Pietro et al. (2016) who infer various stylistic features that distinguish a given gender, age, or occupational class in tweets. They learn phrases (1-3 grams) from the Paraphrase Database (Ganitkevitch et al., 2013) that are semantically equivalent but used more by one demographic than the other, and combine this with a machine translation model to "translate" tweets between demographic classes. However, since their primary objective is not obfuscation, they do not evaluate whether these generated tweets can defeat a demographic classifier.

Spammers are known to modify their e-mails to foil spam detection algorithms, usually by misspelling words that would be indicative of spam, padding the e-mail with lists of arbitrary words, or embedding text in images. It is unclear whether any of these techniques are automated, or to what extent the spammers desire that the modified e-mail appears fluent.

Biggio et al. (2013) formalize the problem of modifying data to evade classifiers by casting it as an optimization problem – minimize the accuracy of

[1] http://donottrack.us

[2] https://adblockplus.org/features#tracking

[3] https://github.com/psal/anonymouth

the classifier while upper-bounding the deviation of the modified data from the original. They optimize this objective with gradient descent and show examples of the tradeoff between evasion and intelligibility for MNIST digit recognition. They work with models that have perfect information about the classifier, as well as when they only know the type of classifier and an approximation of the training data, which is the assumption we will be operating under as well.

Szegedy et al. (2014) and Goodfellow et al. (2015) show that minor image distortions that are imperceptible to humans can cause neural networks as well linear classifiers to predict completely incorrect labels (such as *ostrich* for an image of a truck) with high confidence, even though the classifier predicts the label of the undistorted images correctly. Nguyen et al. (2015) look at the related problem of synthesizing images that are classified as a certain label with high confidence by deep neural networks, but appear as completely different objects to humans.

A line of work called "adversarial classification" formally addresses the problem from the opposite (i.e. the classifier's) point of view: detecting whether a test sample has been mangled by an adversary. Li and Vorobeychik (2014) describe a model to defeat a limited adversary who has a budget for black box access to the classifier rather than the entire classifier. Dalvi et al. (2004) sketch out an adversary's strategy for evading a Naïve Bayes classifier, and show how to detect if a test sample has been modified according to that strategy. Within the theoretical machine learning community, there is a great deal of interest on learning classifiers that do not adversely affect or discriminate against individuals, by constraining them to satisfy some formal definition of fairness (Zemel et al., 2013).

Our problem can be considered one of paraphrase generation (Madnani and Dorr, 2010) with the objective of defeating a text classifier.

3 Problem Description

The general problem of modifying text to fool a classifier is open-ended; the specific question depends on our goals and assumptions. We consider this (simplified) scenario:

1. We do not have access to the actual classifier or even knowledge of the type of classifier or its training algorithm.

2. However, we do have a corpus of labeled data for the class labels which approximate the actual training data of the classifier, and knowledge about the type of features that it uses, as in Biggio et al. (2013). In this paper, we assume the features are bag-of-word counts.

3. The classifier assigns a categorical label to a user based on a collection of their writing. It does not use auxiliary information such as profile metadata or cues from the social network.

4. The user specifies the target label that they want the classifier to assign to their writing. Some users may want to consistently pass off as another demographic. Some may try to confuse the classifier by having half of their writing be classified as one label and the rest as another. Others may not want to fool the classifier, but rather, wish to amplify their gendered features so they are more likely to be correctly classified.[4]

5. The obfuscated text must be fluent and semantically similar to the original.

We hope to relax assumptions 2 and 3 in future work.

Our experimental setup is as follows:

1. Train a classifier from a corpus

2. Train an obfuscation model from a *separate* but similar corpus

3. Apply the obfuscation model to modify the held-out test sentences towards user-provided target labels. These target labels may be the same as the actual labels or the opposite.

4. Evaluate the accuracy of the classifier relative to the desired target labels, and compare it to the accuracy of the same classifier on the actual labels.

[4] Thus, while we will continue to refer to the problem as "obfuscating" the input, it is more generally interpreted as transforming the text so that it is classified as the target label.

4 Data

While our objective is to confound any user-attribute classification system, we focus on building a program to defeat a gender classifier as a testbed. This is motivated partly by of the easy availability of gender-labeled writing, and partly in light of the current social and political conversations about gender expression and fluidity.

Our data is annotated with two genders, corresponding to biological sex. Even though this binary may not be an accurate reflection of the gender performance of users on social media (Bamman et al., 2014; Nguyen et al., 2014), we operate under the presumption that most demographic classifiers also use two genders.

We use two datasets in our experiments – tweets from Twitter, and reviews from Yelp. Neither of these websites require users to specify their gender, so it's likely that at least some authors may prefer not to be profiled. While gender can be inferred from user names (a fact we exploit to label our corpus), many users do not provide real or gendered names, so a profiler would have to rely on their writing and other information.

We chose these corpora since they are representative of different styles of social media writing. Twitter has become the de facto standard for research on author-attribute classification. The writing tends to be highly colloquial and conversational. Yelp user reviews, on the other hand, are relatively more formal and domain-constrained. Both user-bases lean young and are somewhat gender-balanced.

The data is derived from a random sample from a corpus of tweets geolocated in the US that we mined in July 2013, and a corpus of reviews from the Yelp Dataset Challenge[5] released in 2016. Since gender is not known for users in either dataset, it is inferred from users' first names, an approach commonly employed in research on gender classification (Mislove et al., 2011). We use the Social Security Administration list of baby names[6] from 1990; users whose names are not in the list or are ambiguous are discarded. A name is considered unambiguous if over 80% of babies with the name are one gender rather than the other.

We removed data that is not in English, using Twitter's language identifier for the tweet data, and the language identification algorithm of Lui and Baldwin (2011) for the Yelp reviews.

We also removed Yelp reviews for businesses where the reviewer-base was highly gendered (over 80% male or female for businesses with at least 5 reviews). These reviews tend to contain a disproportionate number of gendered topic words like *pedicure* or *barber*, and attempting to obfuscate them without distorting their message is futile. While tweets also contain gendered topic words, it is not as straightforward to detect them.

Finally, excess data is randomly removed to bring the gender balance to 50%. This results in $432,983$ users in the Yelp corpus and $945,951$ users in the Twitter data. The text is case-folded and tokenized using the Stanford CoreNLP (Manning et al., 2014) and TweetNLP (Gimpel et al., 2011; Kong et al., 2014) tools respectively.

The set of users in each corpus is divided randomly into three parts keeping the gender labels balanced: 45% training data for the classifier, 45% training data for the obfuscator, and 10% test data.

5 Obfuscation by Lexical Substitution

The algorithm takes a target label y specified by the user (i.e., the class label that the user aims to be classified as), and their original input text w. It transforms w to a new text w' that preserves its meaning, so that w' will be classified as y.

Our transformation search space is simple: each word in w can be substituted with another one.

For every token $w_i \in w$

- Compute $\mathrm{Assoc}(w_i, y)$, a measure of association between w_i and y according to the obfuscation training data.

  Positive values indicate that w_i as a unigram feature influences the classifier to label w as y and may therefore be retained (taking a conservative route), while negative values suggest that w_i should be substituted.

- If $\mathrm{Assoc}(w_i, y)$ is negative, consider the set V of all words v such that $\mathrm{SynSem}(w_i, v) >$ some threshold τ and

[5] https://www.yelp.com/dataset_challenge

[6] https://www.ssa.gov/oact/babynames/limits.html

$\mathrm{Assoc}(v, y) > \mathrm{Assoc}(w_i, y)$, where SynSem is a measure of syntactic and semantic similarity between w_i and v. This is the set of candidat words that can be substituted for w_i while retaining semantic and syntactic *and* are more predictive of the target label y.

- Select the candidate in V that is most similar to w_i as well as to the two adjacent words to the left and right under Subst, a measure of substitutability in context. Substitute this candidate for w_i, leaving w_i unchanged if V is empty.

$$\arg\max_{v \in V} \mathrm{Subst}(v, w_i, \{w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}\})$$

τ is a hyperparameter that controls the fidelity between w and w'. Higher values will result in w' being more similar to the original; the trade-off is that the obfuscation may not be strong enough to confound the classifier.

Descriptions of the association, similarity and substitutability functions follow.

5.1 Feature-Label Association (Assoc)

Since we don't have direct access to the classifier, an approximate measure how much a feature (word) contributes to the input being classified as a certain label is needed. For two labels y_1 and y_2, we compute the normalized pointwise mutual information between each word f and each of y_1 and y_2 from the obfuscation training set, and take the difference:

$$\mathrm{nPMI}(f, y_1) = \log \frac{P(f, y_1)}{P(f)P(y_1)} / - \log P(f, y_1)$$

$$\mathrm{Assoc}(f, y_1) = \mathrm{nPMI}(f, y_1) - \mathrm{nPMI}(f, y_2)$$

The words that have the highest associations with each gender are listed in Table 1. While these top items tend to be content/topical words that cannot be easily substituted, adjectives and punctuations that are gender-specific also rank high.

5.2 Syntactic+Semantic Similarity (SynSem)

We considered building the lexical similarity model from databases like PPDB (Ganitkevitch et al., 2013), as in Preotiuc-Pietro et al. (2016), but found that their vocabulary coverage for social media text was insufficient, particularly the words (misspellings, slang terms, etc.) that are most predictive of gender.

Distributional word representations tend to do a good job of capturing word similarity. While methods like the `word2vec` skip-gram neural network model of Mikolov et al. (2013) are effective for word similarities, we need to ensure that the substitutions are also syntactically appropriate for lexical substitution. With a skip-gram context window of 5, the most similar words to *eating* are *eat* and *stomachs*, which cannot substitute for *eating* in a sentence. On the other hand, a short content window of 1 gives high similarities to words like *staying* or *experiencing*, which are syntactically good but semantically weak substitutes.

In order to capture syntactic as well as semantic similarities, we employ dependency parses as contexts, using the `word2vec` extension of Levy and Goldberg (2014). Larger corpora of 2.2 million Yelp reviews and 280 million tweets, parsed with Stanford CoreNLP and TweetNLP, are used to train the word vectors. (According to these vectors, the most similar words to *eating* are *devouring* and *consuming*.)

The lexical similarity function $\mathrm{SynSem}(a, b)$ is defined as the cosine similarity between the dependency-parse-based word vectors corresponding to the words a and b.

5.3 Substitutability (Subst)

This determines which of the lexically similar candidates are most appropriate in a given context. We use the measure below, adapted from Melamud et al. (2015), giving the substitutability of a for b in the context of a list of tokens C by averaging over b and the context:

$$\mathrm{Subst}(a, b, C) = \frac{\mathrm{SynSem}(a, b) + \sum_{c \in C} \mathrm{Sem}(a, c)}{|C| + 1}$$

Unlike Melamud et al. (2015) who rely on the dependency-parse-based system throughout, we take $\mathrm{Sem}(a, c)$ to be the cosine similarity between the regular window 5 skip-gram vectors Mikolov et al. (2013), and use the two adjacent words on either side of b as the context C. We found this works

Table 1: Words having the highest associations with each gender

Twitter	
Male	bro, bruh, game, man, team, steady, drinking, dude, brotha, lol
Female	my, you, me, love, omg, boyfriend, miss, mom, hair, retail
Yelp	
Male	wifey, wifes, bachelor, girlfriend, proposition, urinal, oem corvette, wager, fairways, urinals, firearms, diane, barbers
Female	hubby, boyfriend, hubs, bf, husbands, dh, mani/pedi, boyfriends bachelorette, leggings, aveda, looooove, yummy, xoxo, pedi, bestie

better, probably because social media text is syntactically noisier than their datasets.

6 Results

We train L2-regularized logistic regression classification models with bag-of-words counts for the two corpora on their classification training sets. Table 2 shows the prediction accuracies on the unmodified test data as a baseline. (Performance is lower for Twitter than Yelp, probably because of the latter's smaller vocabulary.)

The same classifiers are run on the obfuscated texts generated by the algorithm described above in §5, with target labels set to be (1) the same as the true labels, corresponding to when the test users want to amplify their actual genders, and (2) opposite to the true labels, simulating the case when all test users intend to pass off as the opposite gender. Table 2 shows the accuracy of the classifier at recovering the intended target labels, as well as the relative number of tokens changed from the original text.

The modified texts are significantly better at getting the classifier to meet the intended targets – in both directions – than the unmodified baseline. As expected, lower thresholds for semantic similarity (τ) result in better classification with respect to the target labels, since the resulting text contains more words that are correlated with the target labels.

The more important question is: do the obfuscated inputs retain the meanings of the original, and would they be considered grammatically fluent by a human reader? Future work must obtain participant judgments for a more rigorous evaluation. Examples of the modified texts are shown in Table 3, including some good outputs as well as unacceptable ones. We find that $\tau = 0.8$ is a good balance between semantic similarity of the modified texts with the original and prediction accuracy towards the target label.

Substitutions that don't change the meaning significantly tend to be adjectives and adverbs, spelling variants (like *goood* for *good*), and punctuation marks and other words – generally slang terms – that substitute well in context (like *buddy* for *friend*). Interestingly, spelling errors are sometimes introduced when the error is gendered (like *awsome* or *tommorrow*). Unfortunately, our association and similarity measures also hypothesize substitutions that significantly alter meaning, such as *Plano* for *Lafayette* or *paninis* for *burgers*. However, on the whole, topical nouns tend to be retained, and a perfunctory qualitative examination shows that most of the substitutions don't significantly alter the text's overall meaning or fluency.

7 Discussion

This paper raises the question of how to automatically modify text to defeat classifiers (with limited knowledge of the classifier) while preserving meaning. We presented a preliminary model using lexical substitution that works against classifiers with bag-of-word count features. As far as we are aware, no previous work has tackled this problem, and as such, several directions lie ahead.

Improvements A major shortcoming of our algorithm is that it does not explicitly distinguish content words that salient to the sentence meaning from stylistic features that can be substituted, as long the words are highly gendered. It may help to either restrict substitutions to adjectives, adverbs, punctuation, etc. or come up with a statistical corpus-based

Table 2: Gender identification performance of a logistic regression classifier with bag-of-words features on the original texts from the test sets and the modified texts generated by our algorithm. Performance is measured relative to the target gender label: does every user want the classifier to predict their actual gender correctly, or have it predict the *opposite* gender? Chance is 50% in all cases; higher prediction accuracies are better. Better classifier performance indicates that the texts that are successfully modified towards the users' target labels, which may be to pass off as another gender *or* to reinforce their actual gender. τ controls the trade-off between semantic similarity to the original and association to the target label.

Target		τ	Twitter		Yelp	
			Tokens Changed	Accuracy	Tokens Changed	Accuracy
Reinforce Gender	Original Text	-	0%	69.67%	0%	74.72%
	Modified Text	0.9	2.17%	74.49%	0.38%	76.56%
		0.8	4.45%	80.32%	3.42%	88.17%
		0.5	11.01%	88.73%	9.53%	96.93%
Present as Opposite Gender	Original Text	-	0%	30.33%	0%	25.28%
	Modified Text	0.9	2.61%	37.93%	0.61%	61.19%
		0.8	5.94%	51.58%	4.62%	65.27%
		0.5	15.23%	77.82%	12.74%	91.87%

Table 3: Samples where the classifier predicts the *target* gender correctly on the modified text ($\tau = 0.8$) of the user but incorrectly on the original. Predictions are shown in parentheses.

Yelp		
Original	Modified	Similar meaning/ fluency?
Took my friend here (F)	*Took my buddy here* (M)	Yes
and food still outstanding (M)	*and food still amazing* (F)	Yes
Exceptional view, excellent service, great quality (M)	*Impeccable view, amazing service, wonderful quality* (F)	Yes
the drinks are great, too! (M)	*the drinks are wonderful, too!!* (F)	Yes
tasted as amazing as the first sip I took! Definitely would recommend (F)	*tasted as awsome as the first sip I took; certainly would recommend* (M)	Yes
My wife and I can't wait to go back. (M)	*My husband and I can't wait to go back!* (F)	Somewhat
the creamy rice side dish - delish. (F)	*the succulent rice side dish; unreal.* (M)	Somewhat
I like burgers a lot (M)	*I like paninis a lot* (F)	No
PK was our server (F)	*PK was our salesperson* (M)	No
and I was impressed (M)	*and I is impressed* (F)	No
The girls who work there are wonderful (F)	*The dudes who work there are sublime* (M)	No
Twitter		
Original	Modified	Similar meaning/ fluency?
Yeah.. it's gonna be a good day (M)	*Yeaaah.. it's gonna be a goood day* (F)	Yes
who's up? (M)	*who's up?!* (F)	Yes
I'm so excited about tomorrow (F)	*I'm so pumped about tommorrow* (M)	Yes
I will never get tired of this #beachday (F)	*I will never get tired of this #chillin* (M)	Somewhat
all my niggas look rich as fuck (M)	*all my bitchess look rich as eff* (F)	Somewhat
people from Lafayette on twitter (M)	*people from Plano on tumblr* (F)	No
#TheConjuring (F)	*#pacificrim* (M)	No

measure of whether a word carries meaning in context.

A practical program should handle more complex features that are commonly used in stylometric classification, such as bigrams, word categories, length distributions, and syntactic patterns, as well as non-linear classification models like neural networks. Such a program will necessitate more sophisticated paraphrasing methods than lexical substitution. It would also help to combine word vector based similarity measures with other existing data-driven paraphrase extraction methods (Ganitkevitch et al., 2013; Xu et al., 2014; Xu et al., 2015).

Paraphrasing algorithms benefit from parallel data: texts expressing the same message written by users from different demographic groups. While such parallel data isn't readily available for longer-form text like blogs or reviews, it may be possible to extract it from Twitter by making certain assumptions – for instance, URLs in tweets could serve as a proxy for common meaning (Danescu-Niculescu-Mizil et al., 2012). We would also like to evaluate how well the machine translation/paraphrasing approach proposed by Preotiuc-Pietro et al. (2016) performs at defeating classifiers.

We plan to extensively test our model on different corpora and demographic attributes besides gender such as location and age, as well as author identity for anonymization, and evaluate the quality of the obfuscated text according to human judgments.

Our model assumes that the attribute we're trying to conceal is independent of other personal attributes and a priori uniformly distributed, whereas in practice, attributes like gender may be skewed or correlated with age or race in social media channels. As a result, text that has been obfuscated against a gender classifier may inadvertently be obfuscated against an age predictor even if that wasn't the user's intent. Future work should model the interactions between major demographic attributes, and also account for attributes that are continuous rather than categorical variables.

Other paradigms The setup in Sec. 3 is one of many possible scenarios. What if the user wanted the classifier to be uncertain of its predictions in either direction, rather than steering it one of the labels? In such a case, rather than aiming for a high classification accuracy with respect to the target label, we would want the accuracy to approach 50%. What if our obfuscation program had no side-information about feature types, but instead had some other advantage like black-box access to the classifier? In ongoing work, we're looking at leveraging algorithms to explain classifier predictions (Ribeiro et al., 2016) for the second problem.

Security and adversarial classification Note that we have not shown any statistical guarantees about our method – a challenge from the *opposite* point of view is to detect that a text has been modified with the intent of concealing a demographic attribute, and even build a classifier that is resilient to such obfuscation.

We also hope that this work motivates research that explores provably secure ways of defeating text classifiers.

Practical implementation Eventually, we would like to implement such a program as a website or application that suggests lexical substitutions for different web domains. This would also help us evaluate the quality of our obfuscation program in terms of (1) preserving semantic similarity and (2) its effectiveness against real classifiers. The first can be measured by the number of re-wording suggestions that the user chooses to keep. The second may be evaluated by checking the site's inferred profile of the user, either directly if available, or by the types of targeted ads that are displayed. Further, while our objective in this paper is to defeat automatic classification algorithms, we would like to evaluate to what extent the obfuscated text fools human readers as well.

Acknowledgments

We thank the anonymous reviewers for their suggestions. The first author was partly supported by a Neukom Fellowship at Dartmouth College.

References

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18:135–160.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndic̀, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Proceedings of ECMLPKDD*.

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*.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of EMNLP*.

Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of KDD*.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: how phrasing affects memorability. In *Proceedings of ACL*.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1).

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of ACL*.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of ACL*.

Eric Jardine. 2016. Tor, what is it good for? Political repression and the use of online anonymity-granting technologies. *New Media & Society*.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of CoNLL*.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of EMNLP*.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*.

Bo Li and Yevgeniy Vorobeychik. 2014. Feature cross-substitution in adversarial classification. In *Proceedings of NIPS*.

Jiwei Li, Alan Ritter, and Eduard Hovy. 2014. Weakly supervised user profile extraction from Twitter. In *Proceedings of ACL*.

Wendy Liu and Derek Ruths. 2013. What's in a name? Using first names as features for gender inference in Twitter. In *Proceedings of AAAI Spring Symposium on Analyzing Microtext*.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of IJCNLP*.

Chi Luu. 2015. How to disappear completely: linguistic anonymity on the Internet. JSTOR Daily: http://daily.jstor.org/disappear-completely-linguistic-anonymity-internet/.

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

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL (System Demonstrations)*.

Andrew W. E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*.

Oren Melamud, Omer Levy, and Ido Dagan. 2015. A simple word embedding model for lexical substitution. In *Proceedings of the Workshop on Vector Space Modeling for NLP (VSM)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of NIPS*.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist. 2011. Understanding the demographics of Twitter users. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of EMNLP*.

Dong Nguyen, Dolf Trieschnigg, A. Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska de Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING*.

Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of CVPR*.

Marco Pennacchiotti and Ana-Maria Popescu. 2013. A machine learning approach to Twitter user classification. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.

Daniel Preotiuc-Pietro, Wei Xu, and Lyle Ungar. 2016. Discovering user attribute stylistic differences via paraphrasing. In *Proceedings of AAAI*.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in Twitter. In *Proceedings of SMUC*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. ”Why should I trust you?” Explaining the predictions of any classifier. In *Proceedings of KDD*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Spring Symposium on Analyzing Microtext*.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*.

Joseph Turow, Michael Hennessy, and Nora Draper. 2015. The tradeoff fallacy. Technical report, Annenberg School for Communication, University of Pennsylvania.

April Underwood. 2012. Gender targeting for promoted products now available. https://blog.twitter.com/2012/gender-targeting-for-promoted-products-now-available.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Proceedings of AAAI*.

Wei Xu, Alan Ritter, Chris Callison-Burch, William B. Dolan, and Yangfeng Ji. 2014. Extracting lexically divergent paraphrases from Twitter. *Transactions of the Association for Computational Linguistics*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. Semeval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of SemEval*.

Xiang Yan and Ling Yan. 2006. Gender classification of weblog authors. In *Proceedings of the AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *Proceedings of ICML*.

Kaiping Zhang and René F. Kizilcec. 2014. Anonymity in social media: Effects of content controversiality and social endorsement on sharing behavior. In *Proceedings of the AAAI Conference on Weblogs and Social Media (ICWSM)*.