When Do Words Matter? Understanding the Impact of Lexical Choice on Audience Perception Using Individual Treatment Effect Estimation

Zhao Wang, Aron Culotta
Department of Computer Science
Illinois Institute of Technology, Chicago, IL 60616
zwang185@hawk.iit.edu, aculotta@iit.edu

Abstract

Studies across many disciplines have shown that lexical choice can affect audience perception. For example, how users describe themselves in a social media profile can affect their perceived socio-economic status. However, we lack general methods for estimating the causal effect of lexical choice on the perception of a specific sentence. While randomized controlled trials may provide good estimates, they do not scale to the potentially millions of comparisons necessary to consider all lexical choices. Instead, in this paper, we first offer two classes of methods to estimate the effect on perception of changing one word to another in a given sentence. The first class of algorithms builds upon quasi-experimental designs to estimate individual treatment effects from observational data. The second class treats treatment effect estimation as a classification problem. We conduct experiments with three data sources (Yelp, Twitter, and Airbnb), finding that the algorithmic estimates align well with those produced by randomized-control trials. Additionally, we find that it is possible to transfer treatment effect classifiers across domains and still maintain high accuracy.

1 Introduction[1]

Numerous examples from cognitive science, linguistics, and marketing show that lexical choice can affect audience perception (Danescu-Niculescu-Mizil et al. 2012; Ludwig et al. 2013; Thibodeau and Boroditsky 2013; Riley and Luippold 2015; Reddy and Knight 2016; Preoţiuc-Pietro, Guntuku, and Ungar 2017; Packard and Berger 2017; Nguyen et al. 2017). For example, a social media user who writes *"I'm excited!"* may be more likely to be perceived as female than one who writes *"I'm stoked!"* (Reddy and Knight 2016). Similarly, a book with a review containing the sentence *"I loved this book!"* may be perceived as more desirable than one with a review stating *"An excellent novel."* (Ludwig et al. 2013).

Despite this prior work, we still lack general methods for estimating the causal effect on perception of a single linguistic change in a specific sentence. For example, how much does changing the word *"excited"* to *"stoked"* in the example above increase the chance that a reader will infer the user to be male? Being able to answer such questions has implications not only for marketing and public messaging campaigns, but also for author obfuscation (Hagen, Potthast, and Stein 2017) and stylistic deception detection (Afroz, Brennan, and Greenstadt 2012).

A standard empirical approach is to conduct a Randomized Control Trial (RCT), in which subjects are shown texts that differ only in a single linguistic change, and are subsequently asked to rate their perception with respect to a particular attribute. By controlling for the context, we can then attribute changes in perception to the single linguistic change.

Unfortunately, it is impractical to scale such RCTs to the many possible word substitutions across thousands of sentences, making applications based on such methods infeasible. The goal of this paper is to instead investigate automated methods that estimate how a specific lexical choice affects perception of a single sentence. Our approach builds upon a type of causal inference called *Individual Treatment Effect* (ITE) estimation. An ITE estimation algorithm estimates the effect of an intervention on an individual; e.g., how effective a drug will be for a specific person. Recently, a number of ITE estimators have been proposed that require only observational data, based on Rubin's potential outcome framework (Rubin 1974). In this paper, we formulate our problem as a type of ITE estimation, which we call *Lexical Substitution Effect* (LSE) estimation. We propose two classes of LSE estimators. The first class adapts previous algorithms in ITE estimation to the task of LSE estimation. These methods take as input sentences labeled according to attributes of interest (e.g., a tweet labeled by the gender of the author) and then produces tuples of the form $(w_i, w_j, s, \hat{\tau})$, indicating the estimated LSE ($\hat{\tau}$) of changing the word w_i to w_j for sentence s, with respect to the attribute of interest. The second class of estimator is inspired by recent work that frames causal inference as a classification problem (Lopez-Paz et al. 2015). This approach requires some labeled examples of the form (w_i, w_j, s, τ), where τ is the "true" LSE according to a RCT. It then fits a classifier based on properties of (w_i, w_j, s) to produce LSE estimates for new sentences.

We conduct studies using three data sources: Airbnb listings, Yelp reviews, and Twitter messages. For Airbnb, we

[1] An expanded version of this paper is available at: https://arxiv.org/abs/1811.04890 ; replication files and data are available at: https://github.com/tapilab/aaai-2019-words

consider the perception of the desirability of a rental based on a sentence from the listing description. For Yelp and Twitter, we consider the perception of the gender of the author. We estimate LSE for thousands of word substitutions across millions of sentences, comparing the results of different LSE estimators. For a sample of sentences, we additionally conduct RCTs using Amazon Mechanical Turk to validate the quality of the algorithmic estimates with respect to human judgments. Overall, we find that the algorithmic estimates align well with those produced by RCTs. We also find that it is possible to transfer treatment effect classifiers across domains and still maintain high quality estimates.

2 Related Work

Studies investigating the effect of wording in communication strategies dates back at least 60 years (Hovland, Janis, and Kelley 1953). Recent research has explored the effect of wording on Twitter message propagation (Tan, Lee, and Pang 2014), how word choice and sentence structure affects memorability of movie quotes (Danescu-Niculescu-Mizil et al. 2012), and how characteristics of news articles influence with high story sharing rates (Berger and Milkman 2012). Additionally, there has been recent psycho-linguistic research discovering how to infer user attributes (e.g., gender, age, occupation) based on language styles. Preotiuc-Pietro, Xu, and Ungar (2016) explore a wide set of meaningful stylistic paraphrase pairs and verified a number of psycho-linguistic hypotheses about the effect of stylistic phrase choice on human perception. Preoţiuc-Pietro, Guntuku, and Ungar (2017) further conduct experiment to control human perception of user trait in tweets. Similarly, Reddy and Knight (2016) propose methods to obfuscate gender by lexical substitutions.

As a type of causal inference, individual treatment effect estimation is typically explored in medical trials to estimate effects of drug use on a health outcome. Classical approaches include nearest-neighbor matching, kernel methods and so on (Crump et al. 2008; Lee 2008; Willke et al. 2012). However, the performance of these methods do not scale well with the number of covariates (Wager and Athey 2017). To accommodate a large number of complex covariates, researchers have recently explored techniques such as random forests (Breiman 2001) and causal forests (Wager and Athey 2017). Motivated by their successful applications in the medical domain, we propose to adapt these techniques to the linguistic domain. Specifically, we conduct experiments to algorithmically estimate the causal effect of lexical change on perception for a single sentence.

In summary, while some prior work has studied overall effects of lexical substitution, in this paper we instead propose methods to estimate context-specific effects. That is, we are interested in quantifying the effect on perception caused by a single word change in a specific sentence. The primary contributions are (1) to formalize the LSE problem as a type of ITE; (2) to adapt ITE methods to the text domain; (3) to develop classifier-based estimators that are able to generalize across domains.

3 Individual Treatment Effect Estimation

In this section, we first provide background on Individual Treatment Effect (ITE) estimation, and then in the following section we will adapt ITE to Lexical Substitution Effect (LSE) estimation.

Assume we have dataset D consisting of n observations $D = \{(\mathbf{X}_1, T_1, Y_1), \ldots, (\mathbf{X}_n, T_n, Y_n)\}$, where $\mathbf{X}_i$ is the *covariate vector* for individual i, $T_i \in \{0, 1\}$ is a binary *treatment* indicator representing whether i is in the treatment ($T_i = 1$) or control ($T_i = 0$) group, and Y_i is the observed *outcome* for individual i. For example, in a pharmaceutical setting, i is a patient; $\mathbf{X}_i$ is a vector of the socio-economic variables (e.g., gender, age, height); T_i indicates whether he did ($T_i = 1$) or did not ($T_i = 0$) receive the medication treatment, and $Y_i \in \{0, 1\}$ indicates whether he is healthy ($Y_i = 1$) or sick ($Y_i = 0$).

We are interested in quantifying the causal effect that treatment T has on the outcome Y. The fundamental problem of causal inference is that we can only observe one outcome per individual, either the outcome of an individual receiving a treatment or not. Thus, we do not have direct evidence of what might have happened had we given individual i a different treatment. Rubin's potential outcome framework is a common way to formalize this fundamental problem (Rubin 1974). Let $Y^{(1)}$ indicate the potential outcome an individual would have got had they received treatment ($T = 1$), and similarly let $Y^{(0)}$ indicate the outcome an individual would have got had they received no treatment ($T = 0$). While we cannot observe both $Y^{(1)}$ and $Y^{(0)}$ at the same time, we can now at least formally express several quantities of interest. For example, we are often interested in the *average treatment effect* (τ), which is the expected difference in outcome had one received treatment versus not: $\tau = \mathbb{E}[Y^{(1)}] - \mathbb{E}[Y^{(0)}]$. In this paper, we are interested in the *Individual Treatment Effect* (ITE), which is the expected difference in outcome for a specific type of individual:

$$\tau(\mathbf{x}) = \mathbb{E}[Y^{(1)}|\mathbf{X} = \mathbf{x}] - \mathbb{E}[Y^{(0)}|\mathbf{X} = \mathbf{x}] \quad (1)$$

that is, the treatment effect for individuals where $\mathbf{X} = \mathbf{x}$. For example, if the covariate vector represents the (age, gender, height) of a person, then the ITE will estimate treatment effects for individuals that match along those variables.

Estimating $\tau(\mathbf{x})$ from observational data, in which we have no control over the treatment assignment mechanism, is generally intractable due to the many possible confounds that can exist (e.g., patients receiving the drug may be *a priori* healthier on average than those not receiving the drug). However, numerous algorithms exist to produce estimates of $\tau(\mathbf{x})$ from observational data, for example propensity score matching (Austin 2008). These methods require additional assumptions, primarily the *Strongly Ignorable Treatment Assignment* (SITA) assumption. SITA assumes that the treatment assignment is conditionally independent of the outcome given the covariate variables: $T \perp \{Y^{(0)}, Y^{(1)}\} \mid \mathbf{X}$. While this assumption does not hold generally, methods built on this assumption have often been found to work well.

With SITA, we can estimate ITE using only observational data as follows:

$$\hat{\tau}(\mathbf{x}) = \mathbb{E}[Y|T=1, \mathbf{X}=\mathbf{x}] - \mathbb{E}[Y|T=0, \mathbf{X}=\mathbf{x}] \quad (2)$$

$$= \frac{1}{|S_1(\mathbf{x})|} \sum_{i \in S_1(\mathbf{x})} Y_i - \frac{1}{|S_0(\mathbf{x})|} \sum_{i \in S_0(\mathbf{x})} Y_i \quad (3)$$

where $S_1(\mathbf{x})$ is the set of individuals i such that $\mathbf{X}_i = \mathbf{x}$ and $T_i = 1$, and similarly for $S_0(\mathbf{x})$. In other words, Equation (3) simply computes, for all individuals with covariates equal to $\mathbf{x}$, the difference between the average outcome for individuals in the treatment group and the average outcome for individuals in the control group. For example, if $\mathbf{X} = \mathbf{x}$ indicates individuals with (age=10, gender=male, height=5), $T = 1$ indicates that an individual receives drug treatment and $T = 0$ that they do not, then $\hat{\tau}(\mathrm{x})$ is the difference in average outcome between individuals who receive treatment and those who do not.

A key challenge to using Equation (3) in practice is that $\mathbf{X}$ may be high dimensional, leading to a small sample where $\mathbf{X} = \mathbf{x}$. In the extreme case, there may be exactly one instance where $\mathbf{X} = \mathbf{x}$. Below, we describe several approaches to address this problem, which we will subsequently apply to LSE estimation tasks.

4 Lexical Substitution Effect Estimation

In this section, we apply concepts from §3 to estimate lexical substitution effect on perception. As a motivating example, consider the following two hypothetical sentences describing the neighborhood of an apartment listed on Airbnb:

A: There are plenty of **shops** nearby.

B: There are plenty of **boutiques** nearby.

We are interested in how substituting *shops* with *boutiques* affects the perceived desirability of the rental. E.g., because *boutiques* connotes a high-end shop, the reader may perceive the rental to be in a better neighborhood, and thus more desirable. Critically, we are interested in the effect of this substitution *in one particular sentence*. For example, consider a third sentence:

C: You can take a 10 minute ride to visit some **shops**.

We would expect the effect of substituting *shops* to *boutiques* in sentence **C** to be less than the effect for sentence **A**, since the word *shops* in **C** is less immediately associated with the rental.

First of all, to map the notation of §3 to this problem, we specify a sentence to be our primary unit of analysis (i.e., the "individual"). We make this choice in part for scalability and in part because of our prior expectation on effect sizes — it seems unlikely that a single word change will have much effect on the perception of a 1,000 word document, but it may affect the perception of a single sentence. The covariate vector $\mathbf{X}$ represents the other words in a sentence, excluding the one being substituted. E.g., in example sentence **A**, $\mathbf{X} = \langle$ *There, are, plenty, of, _, nearby* $\rangle$.

Second, we note that there are many possible lexical substitutions to consider for each sentence. If we let p index a substitutable word pair (*control word* → *treatment word*), then we can specify T_i^p to be the lexical substitution assignment variable for sentence i. For example, if p represents the substitutable word pair (*shops*, *boutiques*), then $T_i^p = 0$ indicates that sentence i is in the control group that has the control word *shops* in it, and we call it the control sentence, and $T_i^p = 1$ indicates that sentence i is treated by substituting the control word *shops* to the treatment word *boutiques*.

Third, the outcome variable Y indicates the perception with respect to a particular aspect (i.e., desirability or gender in this paper). For example, in the experiments below, we let $Y \in \{1, 2, 3, 4, 5\}$ be an ordinal variable expressing the perceived desirability level of an apartment rental based on a single sentence.

Finally, with these notations, we can then express the *Lexical Substitution Effect* (LSE), which can be understood as the ITE of performing the word substitution indicated by word pair p on a sentence with context words $\mathbf{X} = \mathbf{x}$:

$$\tau(\mathbf{x}, p) = \mathbb{E}[Y^{p(1)}|\mathbf{X}=\mathbf{x}] - \mathbb{E}[Y^{p(0)}|\mathbf{X}=\mathbf{x}] \quad (4)$$

If we have data of the form $D = \{(\mathbf{X}_1, T_1, Y_1), \ldots, (\mathbf{X}_n, T_n, Y_n)\}$, we can then use the SITA assumption to calculate the LSE:

$$\hat{\tau}(\mathbf{x}, p) = \frac{1}{|S_1^p(\mathbf{x})|} \sum_{i \in S_1^p(\mathbf{x})} Y_i - \frac{1}{|S_0^p(\mathbf{x})|} \sum_{i \in S_0^p(\mathbf{x})} Y_i \quad (5)$$

where $S_1^p(\mathbf{x})$ is the set of sentences i such that $\mathbf{X}_i = \mathbf{x}$ and $T_i^p = 1$, and similarly for $S_0^p(\mathbf{x})$.

As mentioned previously, the high dimensionality of $\mathbf{X}$ is the key problem with using Equation (5). This problem is even more critical in the linguistic domain than in traditional ITE studies in clinical domains — the total number of unique words is likely to be greater than the space of all possible socio-economic variables of a patient. For example, it is entirely possible that exactly one sentence in a dataset has context $\mathbf{X}_i = \mathbf{x}$.

In the subsections that follow, we first describe four algorithms from ITE estimation literature and how we adapt them to LSE estimation. Then, we describe a classifier-based approach that uses a small amount of labeled data to produce LSE estimates. As a running example, we will consider changing *shops* to *boutiques* in the sentence *"There are plenty of shops nearby."* Sentences that contain the control word (e.g., *"shops"*) are called *control samples*, and those containing the treatment word (e.g., *"boutiques"*) are called *treatment samples*. Finally, since we only estimate LSE for one word substitution in one particular sentence each time, we will drop notation p in the following formulas.

4.1 K-Nearest Neighbor (KNN) Matching

KNN is a classical approach for non-parametric treatment effect estimation using nearest neighbor voting. The dimensionality problem is addressed by averaging the outcome variables of K closest neighbors. ITE estimation with KNN computes the difference between the average outcome of K nearest neighbors in treatment samples and control samples:

$$\hat{\tau}_{KNN}(\mathbf{x}) = \Big(\frac{1}{K} \sum_{i \in S_1(\mathbf{x}, K)} Y_i\Big) - \Big(\frac{1}{K} \sum_{i \in S_0(\mathbf{x}, K)} Y_i\Big) \quad (6)$$

For an individual with covariate $X = x$, $S_1(\mathbf{x}, K)$ and $S_0(\mathbf{x}, K)$ are the K nearest neighbors in treatment ($T = 1$) and control ($T = 0$) samples, respectively.

In LSE estimation with KNN matching, we first represent each sentence using standard tf-idf bag-of-words features, then apply cosine similarity to identify the K closest neighbor-sentences. For our running example, we get $S_0(\mathbf{x}, K)$ by selecting the K sentences that have highest cosine similarity with *"There are plenty of __ nearby"* from the control samples, and get set $S_1(\mathbf{x}, K)$ by selecting the K closest sentences to the treatment samples. Then, the KNN estimator calculates LSE by computing the difference between the average label values of K nearest sentences in the treatment samples and control samples.

4.2 Virtual Twins Random-Forest (VT-RF)

The virtual twins approach (Foster, Taylor, and Ruberg 2011) is a two step procedure. First, it fits a random forest with all observational data (including control samples and treatment samples), where each data is represented by inputs $(\mathbf{X}_i, T_i)$ and outcome Y_i. Then, to estimate the ITE, it computes the difference between the predicted values for treatment input $(\mathbf{X}_i, T_i = 1)$ and control input $(\mathbf{X}_i, T_i = 0)$. The name *'virtual twin'* derives from the fact that for the control input $(\mathbf{X}_i, T_i = 0)$, we make a copy $(\mathbf{X}_i, T_i = 1)$ as treatment input that is alike in every way to the control input except for the treatment variable. If $\hat{Y}(\mathbf{x}, 1)$ is the value predicted by the random forest for input $(\mathbf{X} = \mathbf{x}, T = 1)$, then the virtual twin estimate is:

$$\hat{\tau}_{VT}(\mathbf{x}) = \hat{Y}(\mathbf{x}, 1) - \hat{Y}(\mathbf{x}, 0) \quad (7)$$

where $\hat{Y}(\mathbf{x}, 1)$ is the outcome for the *'virtual twin'* (treatment) input and $\hat{Y}(\mathbf{x}, 0)$ is the outcome for control input.

In LSE estimation with VT-RF, we first represent each sentence using binary bag-of-words features (which we found to be more effective than tf-idf). We then fit a random forest to estimate LSE by taking the difference in the posterior probabilities for the virtual twin sentence and the original sentence. For our running example, we fit a random forest classifier using all sentences containing either *shops* or *boutiques* except the current sentence (for out-of-bag estimation). Meanwhile, we generate the virtual twin sentence*"There are plenty of boutiques nearby."* Then the estimated LSE is computed by taking the difference between $P(Y = 1|$*"There are plenty of boutiques nearby"*$)$ and $P(Y = 1|$ *"There are plenty of shops nearby"*$)$.

4.3 Counterfactual Random-Forest (CF-RF)

Counterfactual random forest (Lu et al. 2018) is similar to VT-RF in that they both calculate ITE by taking the difference between predictions of random forest models. However, CF-RF is different from VT-RF by fitting two separate random forests: a control forest fitted with control samples, and a treatment forest fitted with treatment samples. The ITE is then estimated by taking the difference between the prediction (by treatment forest) for a treatment input $(\hat{Y}_1(\mathbf{x}, 1))$ and the prediction (by control forest) for a control input$(\hat{Y}_0(\mathbf{x}, 0))$:

$$\hat{\tau}_{CF}(\mathbf{x}) = \hat{Y}_1(\mathbf{x}, 1) - \hat{Y}_0(\mathbf{x}, 0) \quad (8)$$

In LSE estimation with CF-RF, after representing each sentence with binary bag-of-words features, we first fit a control random forest and a treatment random forest and then estimate LSE by taking probability difference between virtual twin sentence and the control sentence. For example, we fit a control forest with all sentences containing *shops* excluding the current one (for out-of-bag estimation) and a treatment forest with all sentences containing *boutiques*. We then estimate LSE by taking the difference between $P(Y = 1|$*"There are plenty of boutiques nearby"*$)$ predicted by treatment forest and $P(Y = 1|$*"There are plenty of shops nearby"*$)$ predicted by control forest.

4.4 Causal Forest (CSF)

A causal forest (Wager and Athey 2017) is a recently introduced model for causal estimation. While it also uses random forests, it modifies the node splitting rule to consider treatment heterogeneity. Whereas random forests create splits to maximize the purity of Y labels, causal forests instead create splits by maximizing the variance of estimated treatment effects in each leaf. To estimate ITE for an instance i, a causal forest is fit using all treatment and control samples except for instance i. Then for each tree in the fitted forest, instance i is placed into its appropriate leaf node in the tree, and the difference between the treated and control outcomes within that node is used as the ITE estimate of that tree. The final estimate is the average estimate of each tree. Let $L(\mathbf{x})$ be the set of instances in the leaf node to which instance i is assigned, $L_1(\mathbf{x}) \subseteq L(\mathbf{x})$ be the subset of treatment samples, and $L_0(\mathbf{x}) \subseteq L(\mathbf{x})$ be the subset of control samples. Then the estimated causal effect of each tree is:

$$\hat{\tau}_{CSF}(\mathbf{x}) = \frac{1}{|L_1(\mathbf{x})|} \sum_{i \in L_1(\mathbf{x})} Y_i - \frac{1}{|L_0(\mathbf{x})|} \sum_{i \in L_0(\mathbf{x})} Y_i \quad (9)$$

In LSE estimation with CSF, after representing each sentence with binary bag-of-words features, we fit a causal forest model to estimate LSE by aggregating estimations from all trees. For our running example, we fit a causal forest using all sentences containing either *shops* or *boutiques*, excluding *"There are plenty of shops nearby"* and then estimate LSE for the sentence by aggregating estimations from all trees, where estimation by each tree is calculated by taking difference between average label values for treatment samples and control samples inside the leaf where *"There are plenty of shops nearby"* belongs to.

4.5 Causal Perception Classifier

The advantages of the approaches above is that they do not require any randomized control trials to collect human perception judgments of lexical substitutions. However, in some situations it may be feasible to perform a small number of RCTs to get reliable LSE estimates for a limited number of sentences. For example, as detailed in §7.2, we can show subjects two versions of the same sentence, one with w_1 and one with w_2, and elicit perception judgments. We can then aggregate these into LSE estimates. This results in a set of tuples (w_1, w_2, s, τ), where τ is the LSE produced by the randomized control trial. In this section, we develop an approach to fit a classifier on such data, then use it to produce LSE estimates for new sentences.

Our approach is to first implement generic, the non-lexicalized features of each (w_1, w_2, s, τ), then to fit a binary classifier to predict whether a new tuple (w'_1, w'_2, s') has a positive effect on perception or not. This approach is inspired by recent work that frames causal inference as a classification task (Lopez-Paz et al. 2015).

For each training tuple (w_1, w_2, s, τ), we compute three straightforward features inspired by the intuition of the ITE methods described above. Each feature requires a sentence classifier trained on the class labels (e.g., gender or neighborhood desirability). In our experiments below, we use a logistic regression classifier trained on bag-of-words features.

1. Context probability: The motivation for this feature is that we expect the context in which a word appears to influence its LSE. For example, if a sentence has many indicators that the author is male, then changing a single word may have little effect. In contrast, adding a gender-indicative term to a sentence that otherwise has gender-neutral terms may alter the perception more significantly. To capture this notion, this feature is the posterior probability of the positive class produced by the sentence classifier, using the bag-of-words representation of *s after removing word w_1*.

2. Control word probability: This feature is the coefficient for the control word w_1 according to the sentence classifier. The intuition is that if the control word is very indicative of the negative class, then modifying it may alter the perception toward the positive class.

3. Treatment word probability: This feature is the coefficient for the treatment word w_2 according to the sentence classifier. The intuition is that if the treatment word is very indicative of the positive class, then modifying the control word to the treatment word may alter the perception toward the positive class.

We fit a binary classifier using these three features. To convert this into a binary decision problem, we label all tuples where $\tau > 0.5$ as positive examples, and the rest as negative.[2] To compute the LSE estimate for a new tuple (w_1, w_2, s), we use the posterior probability of the positive class according to this classifier. See detailed analysis in §8.

5 Data

This section provides a brief description of experimental datasets (Yelp, Twitter, and Airbnb) for LSE estimation.

A key benefit of our first class of approaches is that it does not require data annotated with human perceptions. Instead, it only requires objective annotations. For example, annotations may indicate the self-reported gender of an author, or an objective measure of the quality of a neighborhood, but we do not require annotations of user perceptions of text in order to produce LSE estimates. While perception and reality are not equivalent, prior work (e.g., in gender perception from text (Flekova et al. 2016)) have found them to be highly correlated. Our results below comparing with human perception measures also support this notion.

Neighborhood Desirability in Airbnb: Airbnb is an online marketplace for short-term rentals, and neighborhood safety is one important factor of desirability that could influence potential guest's decision. Thus, we use crime rate as proxy of neighborhood desirability. We collect neighborhood descriptions[3] from hosts in 1,259 neighborhoods across 16 US cities and collect FBI crime statistics[4] of each city and crime rate of each neighborhood.[5] If a neighborhood has a lower crime rate than its city, we label this neighborhood as desirable; otherwise, undesirable. We get 81,767 neighborhood descriptions from hosts in desirable neighborhoods and 17,853 from undesirable neighborhoods.

Gender in Twitter Message and Yelp Reviews: We choose Twitter and Yelp as representative of different social media writing styles to investigate lexical substitution effect on gender perception. First, we use datasets of tweets and Yelp reviews from (Reddy and Knight 2016), where tweets are geo-located in the US and Yelp reviews are originally derived from the Yelp Dataset Challenge released in 2016.[6] Users in both datasets are pre-annotated with male and female genders. In our sample, we have 47,298 female users with 47,297 male users for Twitter dataset, and 21,650 female users with 21,649 male users for Yelp dataset. Please see Appendix for more details.

6 Generating Candidate Substitutions

Given the combinatorics of generating all possible tuples (w_1, w_2, s) for LSE estimation, we implement several filters to focus our estimates on promising tuples. We summarize these below (see the Appendix for more details):

1. Either w_1 or w_2 must be moderately correlated with the class label (e.g., gender or neighborhood desirability). We implement this by fitting a logistic regression classifier on the labeled data and retaining words whose coefficient has magnitude greater than 0.5.
2. To ensure semantic substitutability, w_2 must be a paraphrase of w_1, according to the Paraphrase Database (PPDB 2.0) (Pavlick et al. 2015).
3. To ensure syntactic substitutability, w_1 and w_2 must have the same part-of-speech tag, as determined by the most frequently occurring tag for that word in the dataset.
4. To ensure substitutability for a specific sentence, we require that the n-grams produced by swapping w_1 with w_2 occur with sufficient frequency in the corpus.

After pruning, for Airbnb we obtained 1,678 substitutable word pairs spanning 224,603 sentences from desirable neighborhoods and 49,866 from undesirable neighborhoods; for Twitter we get 1,876 substitutable word pairs

[2] We use a 1-5 scale in our RCTs, so a treatment effect greater than 0.5 is likely to be significant.

[3] insideairbnb.com

[4] https://ucr.fbi.gov/crime-in-the-u.s/2016

[5] http://www.areavibes.com/

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

spanning 583,982 female sentences and 441,562 male sentences; for Yelp we get 1,648 word pairs spanning 582,792 female sentences and 492,893 male sentences.

7 Experimental Settings

We first carry out experiments to calculate LSE using four estimators and then conduct Randomized Control Trails with Amazon Mechanical Turk (AMT) workers to get human perceived LSE. Next, we fit an out-of-domain causal perception classifier to distinguish LSE directions. Lastly, we evaluate the performance of each method by comparing with human reported values on each dataset separately.

7.1 Calculating LSE Estimates

For experiments with four estimators, we do parameter tuning and algorithm implementation separately. For parameter tuning, we apply the corresponding classification models and do grid search with 5-fold cross validation. For algorithm implementation, we use tuned parameters for each model and follow procedures introduced in §4.

For KNN, we use KNeighborsClassifier in scikit-learn (Pedregosa et al. 2011) for parameter tuning and then select $k = 30$ for estimator implementation. For VT-RF and CF-RF, we use RandomForestClassifier (scikit-learn) for parameter tuning and apply the following values in corresponding estimators: $n\_estimators = 200$, $max\_features =$'$log2$', $min\_samples\_leaf = 10$, $oob\_score = True$. For CausalForest, we use the authors' implementation[7] and experiment with $n\_estimators = 200$ and default values for other parameters as suggested by Wager and Athey (2017).

For the causal perception classifier, our goal is to determine whether the classifier can generalize across domains. Thus, we train the classifier on two datasets and test on the third. We use scikit-learn's logistic regression classifier with the default parameters. To compare this classifier with the results of RCTs, we use the posterior probability of the positive class as the estimated treatment effect, and compute the correlation with RCT estimates.

7.2 Human-Derived LSE Estimates

In order to evaluate the methods, and to train the causal perception classifier, we conducted randomized control trails (RCTs) to directly measure how a specific lexical substitution affects reported perceptions. We do so by eliciting perception judgments from AMT workers.

As it would be impractical to conduct AMT for every tuple (w_1, w_2, s), we instead aim to validate a diverse sample of word substitutions rated highly by at least one of the four LSE estimators. For each dataset, we select the top 10 word substitutions that get the highest LSE according to each estimator. For every selected word substitution (w_1, w_2), we sample three control sentences (sentences containing w_1) with maximum, minimum and median estimated LSE and generate three corresponding treatment sentences by substituting w_1 to w_2 for each control sentence. Thus, we get 120 control sentences and 120 treatment sentences for each dataset. We divide these sentences into batches of size 10; every batch is rated by 10 different AMT workers. The workers are asked to rate each sentence according to its likely perception of an attribute (on a scale from 1 to 5) (e.g., the neighborhood desirability of an Airbnb description sentence, or the gender of author for Twitter and Yelp sentence). Please see Appendix for details on the annotation guidelines.

Increase desirability	**Increase male perception**
store → boutique	gay → homo
famous → grand	yummy → tasty
famous → renowned	happiness → joy
rapidly → quickly	fabulous → impressive
nice → gorgeous	bed → crib
amazing → incredible	amazing → impressive
events → festivals	boyfriends → buddies
cheap → inexpensive	purse → wallet
various → several	precious → valuable
yummy → delicious	sweetheart → girlfriend

Table 1: Samples of substitution words with high LSE

For example, for a tuple (*boyfriend*, *buddy*, *"My boyfriend is super picky"*), we have 10 different workers rate the likely gender of the author for *"My boyfriend is super picky"*, then have 10 *different* workers rate the sentence *"My buddy is super picky"*. The difference in median rating between the second and first sentence is the human perceived effect of changing the word *boyfriend* to *buddy* in this sentence.

Overall, we recruit 720 different AMT workers, 240 for each dataset, and received 237 valid responses for Yelp, 235 for Twitter, and 215 for Airbnb. We compute the Pearson correlation between every two workers who rate the same batch as a measure of inter-annotator agreement as well as the difficulty of LSE tasks for each dataset. These agreement measures, shown in Table 2, suggest that the annotators have moderate agreement (.51-.58 correlation) in line with prior work (Preotiuc-Pietro, Xu, and Ungar 2016). Furthermore, these measures indicate that the Airbnb task is more difficult for humans, which is also expected given that neighborhood desirability is a more subjective concept than gender.

8 Results and Discussion

In this section, we first show a list of substitution words with large LSE estimates, and then provide quantitative and qualitative analysis for different LSE methods.

8.1 Substitution Words with Large LSE

Table 1 shows a sample of 10 substitution words that have large LSE estimates with respect to desirability or gender, based on the automated methods. For example, replacing *shop* with *boutique* increases the perceived desirability of a neighborhood across many sentences. A sentence using the word *tasty* is perceived as more likely to be written by a male than one using *yummy*, and the word *sweetheart* is more often being used by females than *girlfriend*.

[7] https://github.com/swager/grf

	Yelp	Twitter	Airbnb
Agreements-pearson	0.557	0.576	0.513
KNN	0.474	0.291	0.076
VT-RF	0.747	0.333	0.049
CF-RF	0.680	0.279	0.109
CSF	0.645	**0.338**	0.096
Causal perception classifier	**0.783**	0.21	**0.139**

Table 2: Inter-annotator agreement and Pearson correlation between algorithmically estimated LSE and AMT judgment

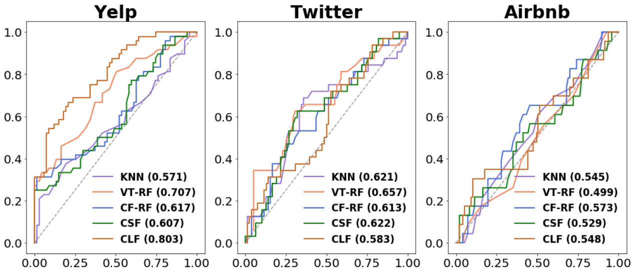

Figure 1: ROC curve for classifying sentences according to AMT perception with estimated LSE as confidence score. *(CLF: causal perception classifier).* Best viewed in color.

8.2 Comparison with RCT Results

To evaluate the performance of LSE estimators, we first compare algorithmically derived LSE with human derived LSE from AMT. Each tuple (w_1, w_2, s) has both an algorithmically estimated LSE $\hat{\tau}$ by each estimator as well as a human derived LSE τ from AMT workers. For 687 annotated tuples, we calculate the Pearson correlation between algorithmic and human-derived LSE estimates. Table 2 shows the results.[8] Additionally, Figure 1 plots ROC curves for classifying sentences as having positive or negative treatment effect, using the LSE estimates as confidence scores for sorting instances.

From these results, we can see that LSE estimators are well aligned with human perception measures, which suggests the suitable proxy of algorithmic estimators with perception measure. There is also considerable variation across datasets, with Yelp having the most agreement and Airbnb the least. Yelp has the most formal writing style among the three datasets, so tree-based estimators (CF-RF, VT-RF, CSF) have competitive performance with humans. Twitter is challenging due to grammatical errors and incomplete sentences. Airbnb has less formal writing style compared with Yelp and contains long sentences with proper nouns (e.g., city names, street names and so on) that lead to the lowest correlation and inter-annotator agreement, suggesting that the more subjective the perceptual attribute is, the lower both human agreement and algorithmic accuracy will be.[9]

	Yelp	Twitter	Airbnb
context pr	-0.348	-0.829	-0.528
control word pr	-0.141	-0.514	-0.367
treatment word pr	0.189	0.401	0.344

Table 3: Logistic regression coefficients for the features of the causal perception classifier

Additionally, we observe that the causal perception classifier outperforms the four other LSE estimators for two of the three datasets. Table 3 shows coefficient values for the classifier when fit on each dataset separately. These coefficients support the notion that certain aspects of LSE are generalizable across domains — in all three datasets, the sign and relative order of the coefficients are the same. Furthermore, the coefficients support the intuition as to what instances have large, positive effect sizes: tuples (w_1, w_2, s) where w_1 is associated with the negative class (control word probability), where w_2 is associated with the positive class (treatment word probability), and where the context is associated with the negative class (context probability).

Finally, we perform an error analysis to identify word pairs for which the sentence context has a meaningful impact on perception estimates. For example, changing the word *boyfriend* to *buddy* in the sentence *"Monday nights are a night of bonding for me and my boyfriend"* is correctly estimated to have a larger effect on gender perception than in the sentence *"If you ask me to hang out with you and your boyfriend I will ... decline."* The reason is that the use of the possessive pronoun "my" reveals more about the possible gender of the author than the pronoun "your." We found similar results on Airbnb for the words *cute* and *attractive* — this change improves perceived desirability more when describing the apartment rather than the owner.

9 Conclusion

This paper quantifies the causal effect of lexical change on perception of a specific sentence by adapting concepts from ITE estimation to LSE estimation. We carry out experiments with four estimators (KNN, VT-RF, CF-RF, and CSF) to algorithmically estimate LSE using datasets from three domains (Airbnb, Twitter and Yelp). Additionally, we select a diverse sample to conduct randomized control trails with AMT and fit causal perception classifiers with domain generalizable features. Experiments comparing Pearson correlation show that causal perception classifiers and algorithmically estimated LSE align well with results reported by AMT, which suggests the possibility of applying LSE methods to customize content to perception goals as well as understand self-presentation strategies in online platforms.

Acknowledgments

This research was funded in part by the National Science Foundation under grants #IIS-1526674 and #IIS-1618244.

[8] Human agreement and algorithmic correlations are calculated differently, so the scores may be in slightly different scales.

[9] Using relative crime rates as a proxy for desirability of Airbnb hosts is a possible limitation.

References

Afroz, S.; Brennan, M.; and Greenstadt, R. 2012. Detecting hoaxes, frauds, and deception in writing style online. In *Security and Privacy, IEEE Symposium on*, 461–475. IEEE.

Austin, P. C. 2008. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in medicine* 27(12):2037–2049.

Berger, J., and Milkman, K. L. 2012. What makes online content viral? *Journal of Marketing Research* 49(2):192–205.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1):5–32.

Crump, R. K.; Hotz, V. J.; Imbens, G. W.; and Mitnik, O. A. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90(3):389–405.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, 892–901. Association for Computational Linguistics.

Flekova, L.; Carpenter, J.; Giorgi, S.; Ungar, L.; and Preoţiuc-Pietro, D. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, 843–854.

Foster, J. C.; Taylor, J. M.; and Ruberg, S. J. 2011. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 30(24):2867–2880.

Hagen, M.; Potthast, M.; and Stein, B. 2017. Overview of the author obfuscation task at pan 2017: safety evaluation revisited. *Working Notes Papers of the CLEF* 33–64.

Hovland, C.; Janis, I.; and Kelley, H. 1953. *Communication and persuasion: psychological studies of opinion change*. Greenwood Press.

Lee, M.-J. 2008. Non parametric tests for distributional treatment effect for randomly censored responses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(1):243–264.

Lopez-Paz, D.; Muandet, K.; Schölkopf, B.; and Tolstikhin, I. 2015. Towards a learning theory of cause-effect inference. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *JMLR Workshop and Conference Proceedings*, 1452–1461. JMLR.

Lu, M.; Sadiq, S.; Feaster, D. J.; and Ishwaran, H. 2018. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics* 27(1):209–219.

Ludwig, S.; De Ruyter, K.; Friedman, M.; Brüggen, E. C.; Wetzels, M.; and Pfann, G. 2013. More than words: The influence of affective content and linguistic style matches in online reviews on conversion rates. *Journal of Marketing* 77(1):87–103.

Nguyen, T. T. D. T.; Garncarz, T.; Ng, F.; Dabbish, L. A.; and Dow, S. P. 2017. Fruitful feedback: Positive affective language and source anonymity improve critique reception and work outcomes. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 1024–1034. ACM.

Packard, G., and Berger, J. 2017. How language shapes word of mouth's impact. *Journal of Marketing Research* 54(4):572–588.

Pavlick, E.; Rastogi, P.; Ganitkevitch, J.; Van Durme, B.; and Callison-Burch, C. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 425–430. Association for Computational Linguistics.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research* 12(Oct):2825–2830.

Preoţiuc-Pietro, D.; Guntuku, S. C.; and Ungar, L. 2017. Controlling human perception of basic user traits. In *Proceedings of the 2017 conference on Empirical Methods in Natural Language Processing*, 2335–2341.

Preotiuc-Pietro, D.; Xu, W.; and Ungar, L. H. 2016. Discovering user attribute stylistic differences via paraphrasing. In *AAAI*, 3030–3037.

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

Riley, T. J., and Luippold, B. L. 2015. Managing investors' perception through strategic word choices in financial narratives. *Journal of Corporate Accounting & Finance* 26(5):57–62.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Tan, C.; Lee, L.; and Pang, B. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. In *Proceedings of ACL*.

Thibodeau, P. H., and Boroditsky, L. 2013. Natural language metaphors covertly influence reasoning. *PloS one* 8(1):e52961.

Wager, S., and Athey, S. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

Willke, R. J.; Zheng, Z.; Subedi, P.; Althin, R.; and Mullins, C. D. 2012. From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Medical Research Methodology* 12(1):185.