A Appendix A: Additional Results

We provide supplemental information and detailed analysis in this section.

A.1 Details for Datasets

Airbnb We collect neighborhood descriptions from hosts in 1,259 neighborhoods across 16 US cities from insideairbnb.com by May 2017. Table 4 shows the 16 cities and corresponding number of neighborhoods we collect in each city.

City	Number of Neighborhoods
LA	248
NY	219
Oakland	108
SanDiego	96
Portland	92
Seattle	87
Denver	74
Chicago	74
NewOrleans	69
Austin	43
WDC	39
SanFrancisco	37
Nashville	35
Boston	25
Asheville	8
SantaCruz	5

Table 4: Cities and number of neighborhoods in each city

Neighborhood Desirability Desirability for a rental in Airbnb varies from person to person. A rental that is attractive to one person might not be attractive to another. The desirability of a rental could be measured by multiple factors such as safety, convenience, surroundings, traffic and so on. Since we want to get an objective measure that could be applied to rentals anywhere instead of subjective measures and we only consider Airbnb rentals inside USA, where safety is a very important factor that could influence potential guest's decision, so we decide to use crime rate as proxy of neighborhood desirability.

We collect crime rate of cities and neighborhoods separately from two sources. For crime rate of cities, we collect from FBI crime statistics[10]. For crime rate of each neighborhood, we collect from areavibes[11]. Considering that crime rate varies for different cities, it is unfair to directly compare neighborhoods in different cities, we decide to make comparisons city by city. We conduct the labeling process in following procedure:

- Label a neighborhood: if a neighborhood has lower crime rate than the city it locates, we label this neighborhood as desirable; otherwise, undesirable.
- Label a host: we assign the same label for hosts in each neighborhood and get 81,767 neighborhood descriptions from hosts in desirable neighborhoods and 17,853 from undesirable neighborhoods. The data imbalance might due to the fact that low-crime areas are more desirable to potential guests, so more Airbnb rentals are listed in low-crime areas than in high-crime areas.
- Label a sentence: we label each neighborhood description sentence with the same label of the neighborhood, which means all desirable neighborhood description sentences are labeled as desirable, otherwise undesirable.

Twitter and Yelp We use tweets and Yelp reviews from datasets introduced in (Reddy et al. 2016). According to (Reddy et al. 2016), tweets are collected in July 2013 and only consider those geolocated in US; the corpus of Yelp reviews is obtained from 2016 Yelp Dataset Challenge[12]. Then they infer the gender of users by mapping users' first names with Social Security Administration list of baby names from 1990[13].

The two datasets are annotated with two genders: male and female, which is suggested as an accurate reflection of social media users. We consider non-binary gender labels as an important area of future work.

After a series of processes including removing users with ambiguous names, dropping non-English and highly gendered texts, they get 432,983 user corpus for Yelp and 945,951 for Twitter. Sampling from their datasets, we get Twitter corpus from 47,298 female users and 47,297 male users, and Yelp corpus from 21,650 female users and 21,649 male users.

Please refer to (Reddy et al. 2016) for more details about Twitter and Yelp datasets.

A.2 Identify Qualified Lexical Substitutions

To generate triples $< w_1, w_2, sentence >$ for LSE estimation tasks, we first search for substitutable word pairs (w_1, w_2) and then select the set of sentences that are qualified for substituting w_1 to w_2.

Considering the large number of possible lexical substitutions, we first apply several criteria to select the most representative words and then match them with the most appropriate substitutions.

Select Representative Words Since we are exploring the subtle effect of a single word change on perceived perception of the corresponding sentence, we want to find words that are representative of attributes we are exploring and thus substituting them might cause effects large enough to be captured. For example, given a sentence: *"I had lunch with my boyfriend"* written by female, *boyfriend* is the most representative words with regard to gender of female, substituting *boyfriend* to *girlfriend* will change the perceived gender of the author from female to male, while substituting the word *had* to *took* does not change the perceived perception. To select representative words, we fit a binary Logistic Regression classifier for each dataset separately.

[10] https://www.freep.com/story/news/2017/09/25/database-2016-fbi-crime-statistics-u-s-city/701445001/

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

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

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

	Airbnb	**Twitter / Yelp**
Q&A	Rate the desirability of a short-term apartment rental based on a single sentence.	Rate how likely you think this tweet / Yelp review sentence is written by male or female.
5	Very desirable	Very likely male
4	Somewhat desirable	Somewhat likely male
3	Neither desirable nor undesirable	Neutral, neither male nor female
2	Somewhat undesirable	Somewhat likely female
1	Very undesirable	Very likely female

Table 5: Amazon Mechanical Turk annotation guidelines

For Airbnb dataset, we fit a classifier with 81,767 desirable and 17,853 undesirable neighborhood descriptions. Considering that description sentences contain lots of proper nouns like street names, famous place names, neighborhood names and city names, we limit the vocabulary to common words that appear at least 8 times in 6 cities and thus eliminating the classifier bias towards proper nouns. By doing so, we get 1,549 common words as representative words of desirable and undesirable classes.

For Twitter and Yelp datasets, after marking proper nouns with NLTK toolkit[14], we fit a binary classifier for Twitter with tweets from 47,298 female users and 47,297 male users; and for Yelp with reviews from 21,650 female users and 21,649 male users. Using coefficient thresholds greater than 0.5 or smaller than -0.5, we select 4,087 gender representative words for Twitter and 2,264 for Yelp .

After selecting representative words, we search for semantically and syntactically qualified substitutions for them.

Semantically Qualified Substitutions (Reddy et al. 2016) apply word2vec extensions of Yelp reviews and tweets parsed with CoreNLP and TweetNLP to capture semantically similar words, and (Preotiuc-Pietro, Xu, and Ungar 2016) use Paraphrase Database (PPDB) to get stylistic paraphrases with equivalence probability greater than 0.2. In our case, we have three corpus with different writing styles and our goal is to find single word substitutions that express the same meaning, so we choose PPDB as our source to get paraphrases in this paper and will consider word2vec extensions of Airbnb corpus in future work. PPDB((Pavlick et al. 2015a)) is a collection high precision paraphrases extracted from bilingual parallel corpora with each paraphrase be assigned with probability and similarity scores according to Google ngrams and Gigaword corpus, and later extended with equivalent scores that interpret semantic relationship between paraphrase pairs. We search for paraphrase pairs with equivalence probability of at least 0.15 ((Preotiuc-Pietro, Xu, and Ungar 2016) use 0.2, we decide to use 0.15 as a relative loose criteria).

Syntactically Qualified substitutions Despite of checking semantics of substitution words, we need to make sure the substitutions are also syntactically qualified. For example, we want to make sure that substitution words have same singular or plural forms. To do so, we first do POS tagging[15] for all sentences in three corpus and store the annotated POS tags of each word, and then check the most common POS tag of each paraphrase pair and only retain paraphrase pairs that have the same most common POS tags.

After limiting substitutions of representative words to semantically and syntactically suitable ones, we search for the set of sentences that are qualified for each specific word substitution.

Check Word Substitutability in Specific Sentences We first build a bi-gram vocabulary using three datasets. Then, for each pair of substitution words (w_1, w_2), we search for sentences containing w_1 and check for every sentence that if substituting w_1 to w_2 produces valid bi-grams by looking up the bi-gram vocabulary. For example, to check the substitutability of $(perced, drilled)$ in *"I'm having my ears pierced on Saturday"*, we decide the grammatically correctness of the sentence after substitution *"I'm having my ears drilled on Saturday"* by checking if *"ears drilled"* and *"drilled on"* exist in our bi-gram vocabulary. If yes, we will keep the current sentence as a qualified sentence for this substitution, otherwise, remove the sentence.

Overall, after pruning with the above criteria, we obtained 1,678 substitutable word pairs spanning 224,603 sentences from desirable neighborhoods and 49,866 from undesirable neighborhoods; and 1,876 substitutable word pairs spanning 583,982 female sentences and 441,562 male sentences for Twitter dataset; and 1,648 word pairs spanning 582,792 female sentences and 492,893 male sentences for Yelp dataset.

A.3 Crowd-sourcing Experiments with Amazon Mechanical Turk

We take a triple $< w_1, w_2, sentence >$ as the unit of analysis in LSE tasks. Despite of algorithmically calculate how much does substituting w_1 to w_2 for the $sentence$ affects its perceived perception, we conduct Randomized Control Trails to directly measure LSE by eliciting judgments from Amazon Mechanical Turk (AMT) workers. Detailed procedures are as follows:

- **Select word pairs with highest LSE** Among all substitution word pairs, we first select those rated highly by at least one of the four LSE estimators (KNN, VT-RF, CT-RF, CSF). Specifically, for each dataset, we get top-10 word substitutions according to each of the four estimators. If a substitution word pair is rated as top-10 with more than one estimators, then we only keep this word pair for the estimator that gives the highest rank (e.g., for

[14] https://www.nltk.org/

[15] We use NLTK (http://nltk.sourceforge.net/) for POS tagging.

a substitution word pair (w_1, w_2), if KNN estimator rank it as the second and VT-RF estimator ranks it as the fifth, then we keep (w_1, w_2) for KNN estimator). Thus, we get 40 substitution word pairs with 10 for each estimator.

- **Select sentences with maximum, minimum and median LSE for each word pair** For each selected word substitution (w_1, w_2), we rank all sentences containing w_1 (and these sentences are qualified for this substitution) according to LSE calculated by the corresponding estimator and sample three sentences with maximum, minimum and median estimated LSE. Meanwhile, we generate corresponding treatment sentences using the given substitution word. Thus, we get 120 control sentences and 120 treatment sentences for each dataset.
- **Create AMT tasks** We divide 120 control sentences into 12 batches with each batch has 10 different sentences, and the same process for 120 treatment sentences. We take each batch as a HIT task in AMT, and for each HIT task, we recruit 10 different workers to pick a scale (ranges from 1 to 5) for each sentence according to its likely perception of an attribute. Table 5 shows the annotation guidelines for three datasets.
- **Quality control of AMT tasks** To eliminate possible biases, we limit that each worker only have access to one batch of either control sentences or treatment sentences. If a worker rates a batch of control sentences, then he won't be able to see the corresponding treatment sentences, so that his decision is not affected by knowing which word is being substituted. For quality control, we require workers to be graduates of U.S. high schools, and we include attentiveness checks using manually created *'dummy'* sentences. For example, a *'dummy'* sentence for gender perception, *'I am the son of my father'*, should be rated as written by a male. We remove responses from workers who provide incorrect answers for dummy questions.

A.4 Experiments with four estimators

We first conduct parameter tuning to select the most suitable parameters for each estimator and then implement four estimators following procedures introduced in §4.

Parameter Tuning As we are estimating LSE on sentence level, we do parameter tuning with all labeled sentences of each dataset.

- **Feature Representation** We try both bag-of-words and tf-idf feature representation techniques for each method.
- **KNN tuning** We use scikit-learn implementation of KNeighborsClassifier, and do grid search for $n\_neighbors$ (since we only need the number of neighbors in KNN estimator implementation, so we don't consider parameters such as min_df and max_df) and get the best 5fold cross validation score with $n\_neighbors = 30$.
- **Random-Forest tuning** We use scikit-learn implementation of Random Forest classifier, and do grid search for a set of parameters and get the best 5-fold cross validation score with $n\_estimators = 200$, $max\_features =$*'log2'*, $min\_samples\_leaf = 10$ and $oob\_score = True$.

Yelp	**Label**
My wife likes this place.	Male
I like coming here with my fraternity brothers.	Male
My brother and I come here for guys night out.	Male
My husband likes this place.	Female
I like coming here with my sorority sisters.	Female
My sister and I come here for girl's night out.	Female
Twitter	
I love playing football and video games.	Male
My wife is waiting on me.	Male
I am my father's son.	Male
I love getting a pedicure at girls night out.	Female
My husband says I smile too much.	Female
I am my mom's daughter.	Female
Airbnb	
This is by far the best neighborhood in the city.	Desirable
This neighborhood is amazing in every way.	Desirable
What a world-class neighborhood this is!	Desirable
This neighborhood is not so great.	Undesirable
Yes, there is a lot of crime in this neighborhood.	Undesirable
Lots of shootings in this neighborhood.	Undesirable

Table 6: Dummy sentences for Yelp, Twitter and Airbnb

Estimator Implementation For estimator implementation, we follow the process introduced in §4 and use the best parameters reported by the above tuning process for KNN VT-RF, CF-RF. For Causal Forest, we try $n\_estimators = 200$ and default values for other parameters.

A.5 Causal Perception Classifier

We fit two classifiers for this task. First, we fit three dataset-classifiers with each fitted with all sentences of one dataset. Dataset classifiers help calculate the proposed features: posterior probability of a context, coefficient of substitution words, and the number of positively and negatively related words. After representing each triple $< w_1, w_2, sentence >$ with these features, we fit three causal perception classifiers only using samples labeled by Amazon Mechanical Turks. Specifically, we fit one causal perception classifier using samples of two datasets and make out-of-domain prediction for samples of the third dataset.

A.6 Results and Analysis

In this section, we provide both qualitative and quantitative analysis from the following aspects:

- First, we present a sample of substitution words estimated to have large LSE.
- Second, we compare the performance of four LSE estimators.
- Third, we evaluate the agreement of each estimator with human perception RCTs using Amazon Mechanical Turk.
- Fourth, we assess the causal perception classifier and interpret feature importance with experimental findings.

- Finally, we provide a preliminary analysis of how this approach may be used to characterize communication strategies online.

Substitution Words with Large Estimated Effects We first display a sample of substitution words estimated to have large LSE by at least one estimator. Table 7 shows a sample of 10 substitutable word pairs for each dataset.

For Airbnb, the substitution words are reported to increase the perceived desirability of a rental. For example, since *boutique* often related with nice neighborhoods, substituting *shop* to *boutique* helps increase the neighborhood desirability. For Twitter and Yelp, the substitution words are reported to increase male perception or decrease female perception of the author. For example, a sentence using *tasty* is more likely to be written by a male than using *yummy*, and chances are high that *sweetheart* would appear in a female sentence while *girlfriend* in a male sentence.

Additionally, to assess the quality of substitution word pairs, we select the top 20 word pairs with largest LSE reported by each estimator and manually check if these word pairs are both syntactically and semantically qualified substitutions. As indicated by Table 8, we find that KNN estimator is somewhat more likely to assign large LSE for qualified substitutions. Unsuitable word pairs are often generated due to the fact that the paraphrase database (PPDB) was trained on general texts, but the validity of a substitution can depend on domain. For example, *gross* and *overall* are potential paraphrases according to PPDB due to one sense of *gross*, but in the Twitter data *gross* is instead more commonly used as a synonym for *disgusting*. More conservative pruning using language models trained on the in-domain data may reduce the frequency of such occurrences.

	Yelp	Twitter	Airbnb	Mean
KNN	100%	85%	90%	**91.67%**
VT-RF	100%	65%	90%	85%
CF-RF	85%	75%	75%	78.33%
CSF	80%	70%	50%	66.67%

Table 8: Fraction of top 20 word substitutions that are judged to be acceptable by manual review

Quantitative Analysis of four LSE Estimators In this section, we quantitatively compare the similarities and differences of each LSE estimator. We expect there to be differences between KNN and the forest-based methods, since their underlying classification functions are different: KNN estimator directly search from all training instances to identify k nearest neighbors in control and treatment group. In contrast, VT-RF, CF-RF and CSF are all tree-based methods, which attempt to place instances in the same leaf if they are homogeneous with resepect to the covariate vector $\mathbf{X}$.

To quantitatively compare the performance of four estimators, we first generate the entire ranked list of $(w_1, w_2, sentence)$ triples according to each estimator and then compute Spearman's rank correlation for ranked list of every two estimators. According to results shown in Table 14, we observe that:

- Forest based methods (VT-RF, CF-RF, CSF) perform more similar than KNN.
- Four estimators have less agreement on Airbnb dataset than on Twitter and Yelp, which suggests that estimating LSE on Airbnb is harder, because hosts are incentivized to highlight desirable aspects of the neighborhood.

	Yelp	Twitter	Airbnb
KNN	90.5%	70.5%	86.6%
VT-RF	93.9%	71.3%	64.9%
CF-RF	**96.6%**	**77.3%**	**87.7%**
CSF	95.9%	71%	84.4%

Table 9: Percentage of negative sentences in top 1000 highly ranked instances with respect to LSE

Then, we calculate the percentage of sentences labeled as negative (refers to undesirable for Airbnb and female for Yelp and Twitter) among top 1000 sentences with large LSEs. Results in Table 9 shows that:

- All of the four LSE estimators tend to pick negative instances for large LSE. Since we rank sentences in descending order of estimated LSE, more negative sentences in top 1000 instances means more effective.
- CF-RF estimator picks the most negative instances for large LSE.
- VT-RF estimator performs differently with other estimators, and especially for Airbnb dataset. The reason may lie in the fact that we label each description sentences as desirable or undesirable according to crime rate of a neighborhood, which means all sentences describing low-crime neighborhoods are labeled as desirable and vice versa. However, sentences describing low-crime neighborhoods are not guaranteed to disclose desirability. Thus, sentences describing undesirability of low-crime neighborhoods are mislabeled as desirable, which misleads VT-RF estimator and explains the difference of this estimator.

Qualitative Analysis of four Estimators To qualitatively assess the performance four estimators, we first show examples to get a better understanding of how do four estimators perform differently in recommending substitution words for one sentence. As shown in Table 16:

- For Yelp, we pick a sentence labeled as male, and find substitution words to make it more likely a sentence written by female. Four estimators give same recommendations for this instance.
- For Twitter, we pick a sentence labeled as female, and four methods recommend substitution words to make it more likely a male sentence. E.g., as *boyfriend* is most likely to be used by females while *buddy* by males, substituting *boyfriend* to *buddy* makes the sentence more likely to be perceived as written by male.
- For Airbnb, we pick a neighborhood description sentence labeled as undesirable, and four estimators make recommendations to improve its desirability. CF-RF and CSF agree on recommendations for this sentence.

Yelp	Twitter	Airbnb
lovely → delightful	gay → homo	store → boutique
cute → attractive	yummy → tasty	famous → grand
helpful → useful	happiness → joy	famous → renowned
fabulous → terrific	fabulous → impressive	rapidly → quickly
gorgeous → outstanding	bed → crib	nice → gorgeous
salesperson → dealer	amazing → impressive	amazing → incredible
belongings → properties	boyfriends → buddies	events → festivals
thorough → meticulous	purse → wallet	cheap → inexpensive
happily → fortunately	precious → valuable	various → several
dirty → shitty	sweetheart → girlfriend	yummy → delicious
Increase male perception or decrease female perception		Increase desirability

Table 7: Substitution words with large LSE

Additionally, we show an example to see how does LSE of a same substitution word vary for different sentences in Table 10. We randomly pick one word substitution in each dataset, and get its highest and lowest LSE sentence according to CSF estimator.

- For Airbnb, substituting *shop* to *boutique* gives lowest LSE on the sentence that is less immediately associated with rental, because it is *"located a mile away"*.
- For Twitter, substituting *boyfriend* to *buddy* gives highest treatment effect for the sentence talking about *"my boyfriend"*, which the word *"my"* is directly associated with the writer of this sentence, so substituting it to *"my buddy"* makes a big change on the writer's gender. But for the lowest treatment sentence, the substitution makes a small change because *"your boyfriend"* and *"your buddy"* do not refer to the writer's gender.

Performance of Causal Perception Classifier Table 11 shows performance of causal perception classifier.

Performance	Yelp	Twitter	Airbnb
AUC	0.803	0.583	0.548
Precision	0.80	0.70	0.65
Recall	0.69	0.72	0.81
F1	0.63	0.62	0.72

Table 11: Performance of causal perception classifier

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

Comparing LSE Reported by Different Estimators with Human Judgments In this section, we evaluate the agreement of four estimators with human perception RCTs using Amazon Mechanical Turk. To do this, we first calculate inner-annotator agreement using both pearson and Spearman's rank and take it as a measure of the difficulty of LSE task with each dataset, and then compute Pearson correlation between LSE reported by each estimator. For the RCTs, we compute human perceived LSE as the difference between median ratings for treatment sentence and control sentence.

Table 12 shows the Pearson correlation between each LSE estimator and AMT reported LSE:

- LSE estimated by four estimators are well aligned with AMT perceived results, which suggests the suitable proxy of objectives measures we use with perception measure. Specially treatment effect for Yelp dataset calculated by CF-RF method has the highest correlation 0.57.
- LSE task for Yelp has the highest correlation with AMT perceived results and three tree-based methods (CF-RF, VT-RF, CSF) have competing performance with innter-annotator agreement.
- LSE task for Airbnb has the lowest correlation, and inter-annotator agreement by pearson and Spearman's rank also be the lowest, which suggests the difficulty for LSE estimation on Airbnb dataset.
- LSE estimators' performance is affected by data source. Yelp has the most formal writing style among the three datasets, so LSE estimators perform as good as humans. Airbnb has less formal writing style compared with Yelp, and there are many long sentences contain proper nouns (e.g., city names, street names, park names and so on.). Finally, hosts always talk attractiveness of a neighborhood even it is not, so the difference between texts describing undesirable neighborhoods and those describing desirable neighborhoods are subtle, which is hard even for humans.

In addition to correlation, we also evaluate whether the direction of algorithmically estimated LSE agree with AMT perceived LSE. To do so, we code estimated LSE as positive or negative, and compute ROC curves for each estimator shown in Table 13.

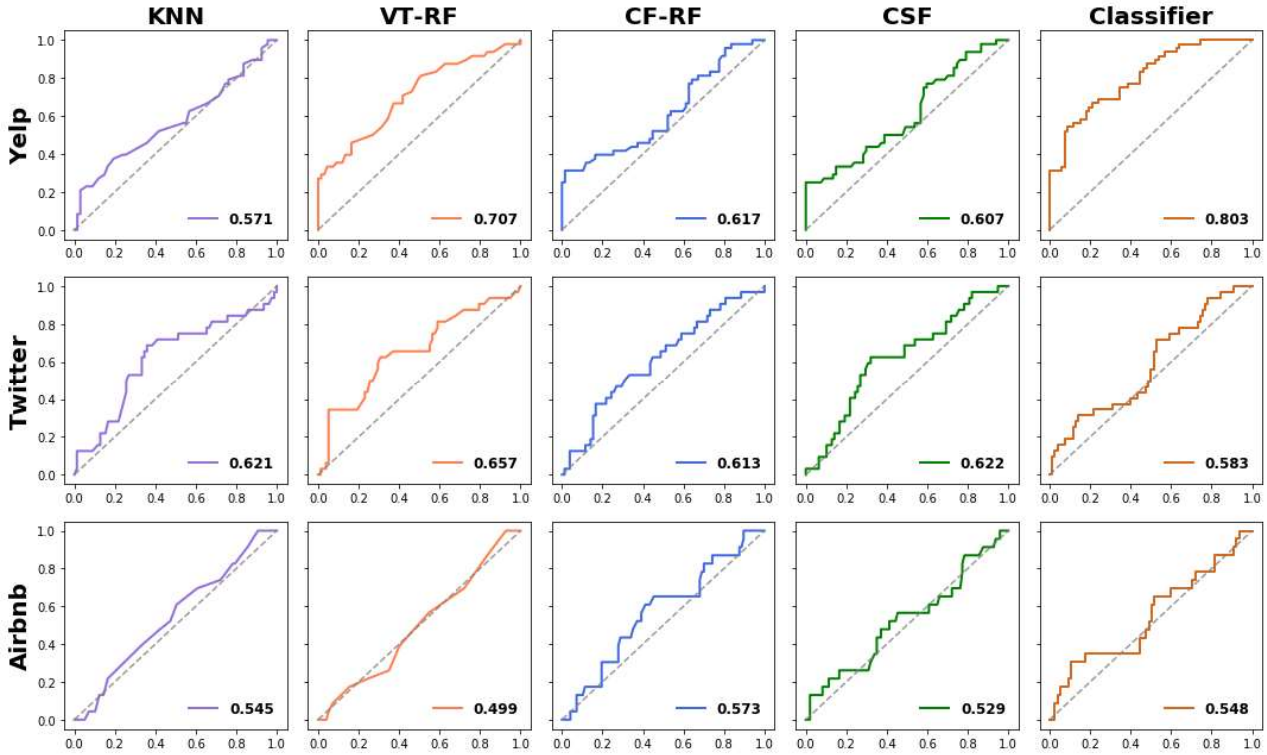

Figure 2: ROC curve for classifying sentences according to AMT perception with estimated LSE as confidence score

Yelp (cute → attractive)	
Largest effect	The joint is *cute* and clean and parking is a breeze.
Smallest effect	Our *cute* Long Island native , Mary suggested the best things on the menu - even telling us what was off and on from the specials board that would work or not.
Twitter (boyfriend → buddy)	
Largest effect	Monday nights are a night of bonding for me and my *boyfriend* ! We both LOVE #TeenWolf user user.
Smallest effect	If you ask me to hang out with you and your *boyfriend* I will look at you like you're stupid then impolitely decline.
Airbnb (store → boutique)	
Largest effect	Check: Andersonville, in particular, has attracted many gay residents (who have remade the upper reaches of Clark Street into a hot design-*store* destination).
Smallest effect	Beachwood Village grocery store and coffee *shop* conveniently located a mile away.

Table 10: Sentences that get the largest and smallest treatment effects for a same word pair

	Yelp	Twitter	Airbnb
KNN	0.571	0.621	0.545
VT-RF	0.707	**0.657**	0.499
CF-RF	0.617	0.613	**0.573**
CSF	0.607	0.622	0.529
Classifier	**0.803**	0.583	0.548

Table 13: Area under ROC curve

Preliminary Analysis Using LSE in Online Communication Strategy In this section, we provide a preliminary analysis of how LSE estimators may be used to characterize communication strategies online. We show potential com-

	Yelp				Twitter				Airbnb			
	KNN	VT-RF	CF-RF	CSF	KNN	VT-RF	CF-RF	CSF	KNN	VT-RF	CF-RF	CSF
KNN	1.0	0.674	0.715	0.655	1.0	0.699	0.729	0.668	1.0	0.469	0.561	0.455
VT-RF		1.0	934	**0.945**		1.0	0.932	**0.935**		1.0	**0.822**	0.773
CF-RF			1.0	0.899			1.0	0.883			1.0	0.733
CSF				1.0				1.0				1.0

Table 14: Spearmanr correlation between ranked sentences of four estimators

	Airbnb	**Twitter**	**Yelp**
Increase desirability or male perception	closest → best stores → boutiques famous → old plaza → place	okay → good sweatheart → girlfriend purse → wallet precious → rare	gorgeous → super yummy → tasty fabulous → excellent hunt → search
Decrease desirability or male perception	excellent → safe best → hottest gorgeous → great boutiques → stores	ma → mom crib → bed impressive → wonderful buddy → boyfriend	tasty → yummy excellent → cute good → yummy attractive → cute

Table 15: Word substitutions with high LSE used most frequently by authors of the opposite class (e.g., "male" words used by female users, and visa versa.)

Yelp (make it more likely a female sentence)	
Original	Very fresh , and *tasty* herbs and spring rolls as well !
KNN/VT-RF/ CF-RF/CSF	Very fresh , and *yummy* herbs and spring rolls as well !
Twitter (make it more likely a male sentence)	
Original	Every girl I know is with it and makes *nice* dinners for their *boyfriends* while I just order pizza and drink *beer* with mine #sorrybabe.
KNN/CF-RF	Every girl I know is with it and makes *good* dinners for their *buddies* while I just order pizza and drink *beer* with mine #sorrybabe.
VT-RF/CSF	Every girl I know is with it and makes *nice* dinners for their *buddies* while I just order pizza and drink *brew* with mine #sorrybabe.
Airbnb (increase desirability)	
Original	I don't suggest long walks after dark, but I would *definitely* not let this neighborhood discourage your stay, it's *vibrant*, fun and *exciting*.
KNN	I don't suggest long walks after dark, but I would *truely* not let this neighborhood discourage your stay, it's *dynamic*, fun and *interesting*.
VT-RF	I don't suggest long walks after dark, but I would *really* not let this neighborhood discourage your stay, it's *dynamic*, fun and *stunning*.
CF-RF/CSF	I don't suggest long walks after dark, but I would *absolutely* not let this neighborhood discourage your stay, it's *dynamic*, fun and *spectacular*.

Table 16: Different recommendations of substitution words for one sentence

munication strategies people use for perception management (try to improve positive perception and reduce negative perception, or to change female style to male or vice versa) according to results suggested by current datasets.

To do this, we first select top 20 highest and lowest ranked word substitutions according to each LSE estimator. Then, for the 20 highest ranked word substitutions, we sort them according to the frequency of positive treatment words used in negative sentence; for the 20 lowest treatment word pairs, we sort them according to the frequency of negative treatment words used in positive sentence. Table 15 shows a list of highly ranked word pair selected according to each estimator:

- For Airbnb, hosts in undesirable neighborhoods use words *best* instead of *closest* and *boutiques* instead of *shops* more often, which are signs of improving desirability perception. While for hosts in desirable neighborhoods, the estimators suggest them to use words *excellent* instead of *safe* because *safe* reduces positive perception compared with *excellent* (according to LSE recommendations); this makes sense because hosts located in safe neighborhoods would not emphasize safety.
- For gender perception of Twitter and Yelp, LSE estimators recommend that if you want to write sentence like a female, then use *sweatheart* instead of *girlfriend* and use *yummy* instead of *tasty*. Otherwise, if you want to write sentences like a male, use *buddy* instead of *boyfriend* and use *attractive* instead of *cute*. Additionally, LSE estimators recommend to use more emotional words for female sentence than for male.
- zw: Add explanations from reviews' comments