Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

# A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles

**Xiao Ma, Trishala Neeraj, Mor Naaman**
Jacobs Institute, Cornell Tech
New York, NY, USA
{xm75, tn338, mor.naaman}@cornell.edu

## Abstract

We developed a novel computational framework to predict the perceived trustworthiness of host profile texts in the context of online lodging marketplaces. To achieve this goal, we developed a dataset of 4,180 Airbnb host profiles annotated with perceived trustworthiness. To the best of our knowledge, the dataset along with our models allow for the first computational evaluation of perceived trustworthiness of textual profiles, which are ubiquitous in online peer-to-peer marketplaces. We provide insights into the linguistic factors that contribute to higher and lower perceived trustworthiness for profiles of different lengths.

## Introduction

Collaborative consumption platforms (e.g., Airbnb, Uber, Lyft, TaskRabbit) have been transforming how economic activities take place, but require high level of trust between users to function. User profiles play a key role in establishing trust between peers in collaborative consumption (Ma et al. 2017), but user-generated profiles could also contain deceptive information. Previous research on online dating profiles shows that users attend to small linguistic cues, such as spelling ability and grammar when assessing potential dates (Ellison, Heino, and Gibbs 2006). However, it is not clear how these linguistic cues impact the assessment of potential exchange partners in sharing economy.

Here, we extend our previous work and dataset on Airbnb host profiles (Ma et al. 2017), by developing a *computational* framework to predict the perceived trustworthiness of host profiles in the context of online lodging marketplaces. We developed a dataset of 4,180 host profiles annotated with perceived trustworthiness scores, the largest such dataset to date. To enable the computational analysis, we developed several models building on various language-based features. Using these features and models, we evaluate a prediction task distinguishing profiles of low and high perceived trustworthiness. In addition, we use Lasso regression to examine the factors that contribute to higher and lower perceived trustworthiness. We discuss the results in relation to previous research on deception (Toma and Hancock 2012) and loan defaults (Netzer, Lemaire, and Herzenstein 2016), showing that the linguistic features contributing to higher perceived trustworthiness may not always align with features that were reported to be associated with other factors that may be indicative of *actual* trustworthiness.

Our work builds on a number of studies using computational approaches to study language and social interactions, enabled by new corpus and techniques from natural language processing and machine learning. For example, Danescu-Niculescu-Mizil et al. (2013) detects politeness computationally from text by constructing new datasets of online requests; Mitra and Gilbert (2014) found that in crowdfunding sites, language that makes direct promises such as "project will be" and "pledgers will receive" is predictive of a project being funded; and finally, Netzer, Lemaire, and Herzenstein (2016) found that language in loan requests can help predict loan defaults.

## Method and Dataset

Our main dependent variable in this work is the *perceived trustworthiness* of host profiles in the context of online lodging marketplaces. Here, trustworthiness is defined as an attribute of a trustee (the host). We measure *perceived* trustworthiness of hosts based on their profile texts alone through a custom scale developed in Ma et al. (2017), which asks potential guests about how confident they are that the host in question is capable, benevolent, and with integrity (Mayer, Davis, and Schoorman 1995)

We used the Airbnb dataset collected by an independent organization, Inside Airbnb[1] on 12 U.S. major cities, and conducted a weighted sample of profiles across all cities for 3,000 unique host profiles. We divided profiles into 150 batches, each containing 20 profiles, and recruited three annotators for each batch from Amazon Mechanical Turk (AMT), paying $1.5 for each task. We required annotators to be based in U.S., adult, and with previous approval rate of at least 90%. We also required that each worker to only perform the task once (i.e. rate exactly 20 profiles). The annotation task is described in more detail by Ma et al. (2017), though for the work here we reduced the number of annotators per-profile from five to three. Finally, in the exit survey, we collected additional information about the annotators that was not collected by Ma et al. (2017), including



[1] http://insideairbnb.com/

demographic information as well as their generalized trust attitude using the scale by Yamagishi and Yamagishi (1994).

We received 450 responses from the AMT workers and applied a series of filtering process to exclude potential "spammers", including checking the answer to a linguistic attentiveness question, the standard deviation and mean of the ratings of the same worker, and task completion time. We filtered out responses that have very low and high (bottom and top 2.5%) standard deviation, mean, and task completion time. We retained the rating of a host profile if it has at least one rating after the filtering. In the end, we retained new annotations of perceived trustworthiness of 2,980 host profiles.

After initial filtering, we performed $z$-score standardization on the scores given by the same annotator, as we expected that each annotator's scores are subjective with different baselines for trust. Indeed, our data shows a significant correlation between an annotator's reported generalized trust attitude, and the average trustworthiness scores the annotator assigned to the 20 profiles [$\beta = .34$, $t(380) = 7.08$, $p < .001$], further justifying the decision to standardize the scores per annotator. After standardization, we took the average of scores given to the same profile by workers to be the perceived trustworthiness score of the profile.

To evaluate the reliability of annotations, we calculated the mean pairwise Pearson correlation for all profiles among three raters, pooling all the data. The average pairwise correlation is $0.49$ ($0.29$ before standardization). Naturally, trustworthiness is a subjective concept. Our data showed patterns similar to data from previous research on politeness (Danescu-Niculescu-Mizil et al. 2013): higher agreement at the extremes and lower agreement in the middle, which motivated our evaluation setup as detailed below.

Since our annotation process is almost exactly the same as Ma et al. (2017), we merged the new dataset with the one reported in previous work in order to boost the amount of data available for training and testing. We performed $z$-score standardization on the previous dataset before merging with the dataset we newly acquired. As a result of this merging process, we now have an extended Airbnb host profile dataset containing a total of 4,180 profiles. The perceived trustworthiness scores in our extended dataset have a mean of zero and standard deviation of $0.8$. We use this extended dataset in subsequent analysis. The extended dataset is available online[2] and contains all profile texts, perceived trustworthiness annotation, as well as the demographic information and generalized trust attitude of annotators.

With the extended dataset, we set up two tasks: prediction and regression. For the first task, our goal is to find the best model that predicts perceived trustworthiness. As trustworthiness is a subjective concept, we set up the task as a binary classification, following the example in Danescu-Niculescu-Mizil et al. (2013). We used logistic-regression classifiers and only top and bottom quartile of the profiles in different profile lengths buckets in terms of perceived trustworthiness score. For the second task, our focus is on understanding, which we address using Lasso regression for feature selection, also using the top and bottom quartile of the data in each length batch.

[2]https://github.com/sTechLab/AirbnbHosts-Extended

## Predicting Perceived Trustworthiness

In this section, we set up the prediction task, and discuss features that we construct from profile text as inputs for different prediction models, as well as model performance.

### Evaluation Setup

We split our data into two parts: a training and cross-validation set (80% of data) that we use during model tuning, and a held-out set (the rest 20%) that is kept separate and reserved for the final test.

We frame the prediction task for profiles of different lengths. We know that length plays a significant role in predicting perceived trustworthiness (Ma et al. 2017), which is again confirmed with our extended dataset [$\beta = .65$, $t(4, 178) = 55.61$, $p < .001$]. To this end, after trimming the outliers, i.e. the shortest and longest profiles (bottom 5% and top 5% in terms of word count), we divided the rest of the profiles into five equal batches based on word count. The batches, from shortest to longest profiles, have the following ranges of word count: 6–19 words, 20–36, 37–58, 59–88, and 89–179.

Within each batch, we calculate the bottom and top quartile of the perceived trustworthiness score. We then use logistic-regression classifiers to predict, for profiles in these two quartiles, whether they will be in the bottom quartile (zero), or top quartile (one), therefore only using 50% of the data in the cross-validation set. We measure the quality of our prediction using the accuracy of the classifiers (we are not using F1 and AUC scores as the labels are balanced).

### Model Features

For our prediction models, we used different combinations of the features described below. We first performed the following data pre-processing. After removing punctuation and numbers using regular expression matching, we converted the remaining letter words into lowercases, and removed stop words using a union of lists of English stop words from NLTK and *scikit-learn* feature extraction module, consisting of 352 stop words. Finally, we lemmatized verbs and nouns using NLTK WordNet lemmatizer.

**LIWC features** We extracted 73 features from raw profile text (before pre-processing) using LIWC (Linguistic Inquiry and Word Count). LIWC is a dictionary-based text analysis tool that counts the percentage of words that reflect linguistic process, psychological process, and personal concerns. LIWC has been shown to predict numerous psychological outcomes (Pennebaker, Francis, and Booth 2001). We used the 2007 version of LIWC and substituted readability in LIWC with Flesch-Kincaid grade level (extracted using the Python package *textstat*).

**Bag-of-Words** We vectorized each of the pre-processed profiles using CountVectorizer from the *scikit-learn* library. We used one-, bi-, and tri-grams and required the grams to have appeared at least 20 times. This process resulted in 1,012 word features, which we use in our baseline model.

| Category | Accuracy | F1-Score | AUC |
|---|---|---|---|
| Interests & Tastes | .89 | .74 | .92 |
| Life Motto & Values | .97 | .18 | .71 |
| Work or Education | .92 | .79 | .93 |
| Relationships | .92 | .52 | .88 |
| Personality | .93 | .52 | .89 |
| Origin or Residence | .89 | .78 | .93 |
| Travel | .93 | .78 | .95 |
| Hospitality | .86 | .72 | .91 |

Table 1: Performance of sentence category classification.

| Features | 6–19 words | 20–36 words | 37–58 words | 59–88 words | 89-179 words |
|---|---|---|---|---|---|
| WC | 57.5% | 53.8% | 46.9% | 43.9% | 50.0% |
| BOW | 60.8% | 58.5% | 62.6% | 59.4% | 58.8% |
| BOW + WC | 59.1% | 57.8% | 63.0% | 59.1% | 60.2% |
| LIWC + WC | **69.1%** | 58.1% | 58.4% | 61.8% | 57.8% |
| Category + WC | 64.0% | 57.1% | 62.3% | 65.2% | 65.3% |
| Category + LIWC + WC | 65.9% | **59.4%** | **65.9%** | **65.9%** | **68.0%** |
| Best Model on Held-Out | 72.4% | 67.1% | 51.2% | 58.2% | 62.5% |

Table 2: Model performance (accuracy) summary by different length batch. Random baseline accuracy is 50%.

**Sentence Categories** In Ma et al. (2017), we manually developed a set of eight sentence categories (shown in the first column in Table 1) that frequently appear in Airbnb host profiles. We created a dataset of 5,248 profile sentences tagged with categories (we used the term "topics" in Ma et al. (2017), but to avoid confusion with the term commonly used in the context of topic modeling in the NLP community, we refer to them as "categories" here).

Here we leverage the sentence level annotation dataset and trained eight binary classifiers to predict whether a sentence belongs to each category. We used the same preprocessing pipeline, and a one-gram bag-of-words model. We set the minimum threshold of token frequency to be 10, resulting in 616 features. We used a Bernoulli naive Bayes classifier, one for each category, and five-fold cross validation to evaluate the performance of the classification. The accuracy, F1-score and AUC for sentence category classification are listed in Table 1. The category *Life Motto & Values* has the worst F1 and AUC performance due to the extremely imbalanced label—there are very few sentences that were tagged to belong to this category.

We applied the trained classifiers on each sentence in the extended dataset, then adding up the classification results for sentences for the same profile into a vector of length eight representing how many sentences in the profile were tagged as belonging to one of the eight categories.

## Models and Evaluation

We combine previously extracted features into different models and evaluate their classification performance using cross-validation. We chose the simple Word Count (WC) and BOW as baseline models. For LIWC and sentence category, we compared the performance of models using each set of features alone, and each plus Word Count, and found that Word Count improves performance; we only include the WC-enhanced models here (WC did not improved on BOW model, but BOW+WC is included here for completeness). We report the performance of all models in Table 2.

The bold numbers in Table 2 indicate the best performing model for each length batch. For the shortest profiles, the LIWC+WC combination achieves the best prediction result, while longer profiles benefit from including sentence category as features.

**Evaluation on Held-Out Set** After picking the best performing model for each length batch, we re-fitted the models on the entire cross-validation dataset to predict data from our completely disjoint held-out set. For the held-out set, we separated the profiles using the word count thresholds as defined in the training stage in to each batch, and obtained the perceived trustworthiness quartile tags using the thresholds obtained from training stage. We report the accuracy of prediction on held-out set in the last line of Table 2. Overall, the performance levels on the held-out set are comparable to other text-classification work (Danescu-Niculescu-Mizil et al. 2013; Tan, Lee, and Pang 2014).

## Factors Contributing to Perceived Trustworthiness

We conduct Lasso regression for profiles of different lengths to uncover factors that contribute to higher and lower perceived trustworthiness. We again use the top and bottom quartile of the data for each length batch in the cross-validation dataset for this analysis, using the $R$ implementation of Lasso logistic regression (`cv.glmnet`) with default 10-fold cross-validation to choose the best parameter ($\lambda$) and using area under curve (AUC) as measure for goodness of fit. We report features that appear in more than 10% of the profiles as well as selected to be non-zero by Lasso in Table 3 and discuss the findings below.

## Discussion and Conclusion

We have developed a computational framework that can distinguish between Airbnb host profiles of low and high perceived trustworthiness. We also uncover features that are most predictive to higher perceived trustworthiness, such as *Hospitality* sentence category, and LIWC features *social* and *work*. However, these features may not always align with features that were reported to be associated with other factors that may be indicative of *actual* trustworthiness.

The key factors contributing to higher perceived trustworthiness are *Hospitality* sentence category, and LIWC features *social* and *work*. The sentence category *Hospitality* contributes to higher perceived trustworthiness for profiles longer than 37 words. The effectiveness of hospitable language strengthens the findings of Ma et al. (2017). The *social* LIWC category also contributes to higher perceived trustworthiness, potentially through the mechanism of uncertainty reduction (Berger and Calabrese 1975). Providing information about one's social relationships can make hosts appear more "real". Finally, LIWC feature *work* predicts higher perceived trustworthiness for all profiles shorter than 59 words, potentially also through the mechanism of uncertainty reduction.

| Profile Length | Positive Features | | Negative Features | |
|---|---|---|---|---|
| | Sentence Category | LIWC | Sentence Category | LIWC |
| 6–19 words | (Not included) | readability, comma, article, conjunction, social, affect, causation, health, work, achieve | (Not included) | word per sentence, exclamation mark, present tense verb, adverb, quantifier, human, tentative, perceptual process, sexual, relativity, motion, space, leisure |
| 20–36 words | Travel | we, social, positive emotion, cognitive process, sexual, work, home | — | parenthesis, adverb, prepositions, perceptual process |
| 37–58 words | Work or Education, Relationships, Hospitality | parenthesis, we, article, auxiliary verb, past tense verb, social, family, friend, certain, work | Interests & Tastes, Personality | comma, dash, exclamation mark, period, negation, quantifier, insight, motion, leisure |
| 59–88 words | Relationships, Hospitality | we, social | — | — |
| 89–179 words | Hospitality | social, inclusive | — | — |

Table 3: Factors contributing to higher and lower perceived trustworthiness by different length batch (note that the "sexual" category contained mostly the word "love").

In contrast, LIWC feature *leisure* and sentence category *Interest & Tastes* are contributing *negatively* to perceived trustworthiness for profiles between 6–19 words and 37–58 words respectively. The negative effect of these features may suggest a separation between the need of sociability and the ability to provide standard goods and services in sharing economy.

Comparing these features that were found to be significant in our work with previous research, we uncover a potential discrepancy between the language that is *perceived* to be trustworthy, and the *actual* trustworthiness of individuals. In terms of perception, as we see in our work, and previous work on crowd funding, LIWC features *social* and *work* contribute to higher perceived trustworthiness or higher likelihood of a project being funded (Mitra and Gilbert 2014). However, in terms of actual trustworthiness, *social* language is found to be associated with higher loan default rates (Netzer, Lemaire, and Herzenstein 2016); and in online dating profiles, online daters used more *work* related words (Toma and Hancock 2012) when their photos are less accurate. Expanding on the discrepancy between perceived and actual trustworthiness would be important future work.

A key question that requires future work is whether our findings are unique to online lodging marketplaces, more specifically to Airbnb, or maybe they apply more generally in peer-to-peer exchange platforms. While we believe some features we identified, for example some LIWC features in the different models, apply more generally, other features are context specific. For example, the sentence category *Hospitality* is specific to the lodging context, though at the same time a version of it can transfer to other domains (e.g., promise of service). Future work can expand our results to other domains.

## References


Berger, C. R., and Calabrese, R. J. 1975. Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human Communication Research* 1(2):99–112.

Danescu-Niculescu-Mizil, C.; Sudhof, M.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. A computational approach to politeness with application to social factors. In *Proceedings of ACL*.

Ellison, N.; Heino, R.; and Gibbs, J. 2006. Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication* 11(2):415–441.

Ma, X.; Hancock, J. T.; Lim Mingjie, K.; and Naaman, M. 2017. Self-disclosure and perceived trustworthiness of airbnb host profiles. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, 2397–2409. New York, NY, USA: ACM.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review* 20(3):709–734.

Mitra, T., and Gilbert, E. 2014. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, CSCW '14, 49–61. New York, NY, USA: ACM.

Netzer, O.; Lemaire, A.; and Herzenstein, M. 2016. When words sweat: Identifying signals for loan default in the text of loan applications (working paper). *Social Science Research Network*.

Pennebaker, J. W.; Francis, M. E.; and Booth, R. J. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71:2001.

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*.

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

Yamagishi, T., and Yamagishi, M. 1994. Trust and commitment in the united states and japan. *Motivation and Emotion* 18(2):129–166.