AI-based Text Recommendation's Impact on Profile Writing

Ritika Poddar
Cornell Tech
New York, New York, USA
rp477@cornell.edu

Rashmi Sinha
Cornell Tech
New York, New York, USA
rs2584@cornell.edu

ABSTRACT

AI based text recommendations have become increasingly common in online text editors. In this paper we examined the effect of these text recommendations on the content of Airbnb host profiles. Using previously annotated datasets coded with 8 different categories, we fine-tuned two GPT-2 models on category specific Airbnb profile sentences. Using a text editor that we built, we conducted a study where these models generated suggestions for participants as they typed their Airbnb host profiles. We analyzed the sentences in the resulting profiles to evaluate the effect of the model suggestions and whether a nudging effect, in terms of content, was present. Our results found no significant effect on the content produced by the participants based on the category of data the model was fine-tuned with.

CCS CONCEPTS

• **Human-centered computing** → *Human computer interaction (HCI)*; • **Computing methodologies** → **Natural language generation**.

KEYWORDS

machine learning, natural language processing, transformers, nudging

ACM Reference Format:
Ritika Poddar and Rashmi Sinha. 2021. AI-based Text Recommendation's Impact on Profile Writing. In . ACM, New York, NY, USA, 5 pages.

1 INTRODUCTION

AI based text recommendations have become increasingly widespread and have evolved from single word shortcuts and suggestions on texting, to predictive phrases on sites like Gmail, and now to full predictive sentences. These predictive writing assistants are designed to make the writer's life easier, by anticipating the completion of the sentence given a part of it. The creation of new language models like OpenAI's GPT-2 model allows us to build better predictive systems, and use AI to generate fully coherent sentences and paragraphs. GPT-2 is a transformer based language model with 1.5 billion parameters, trained on 8 million online documents[6]. Using this language model we are interested in observing the impact of its predictive suggestions on a specific type of user input: online profiles.

New sharing economy platforms such as Airbnb, Task Rabbit, and Fiverr, allow users to post descriptive profiles of themselves. The profiles are meant to portray different characteristics that ultimately establish trustworthiness. For our study we will be focusing on Airbnb. Airbnb is an online marketplace that connects those looking to rent out their lodging to those seeking for one. It is typically used for travellers and for those looking for unique stays that hotels rarely offer. Ever since its inception in 2008, Airbnb has grown to have around two million people staying with Airbnb each night[3].

Our intention is to fine-tune multiple versions of the GPT-2 model on specific characteristics of these host profiles, based on categories identified by previous studies, and identify the impacts of the models' suggestions on the final content of the participant's host profiles.

In a previous study [4] researchers manually coded about 1,200 Airbnb host profiles with 8 different topics: interests and taste, life mottos and valuess, work and education, relationships, personality, origin and residence, hospitality, and travel. The subsequent study [5] took the same coding scheme and coded 4,180 Airbnb host profiles using a computational model for the classification. We used both these datasets for fine-tuning our GPT-2 model and utilized the coding scheme to evaluate the categories in each of the sentences in our participant's data.

2 RELATED WORK

In the context of Airbnb host profiles, our work draws inspiration from previous work done by Cornell Tech researchers on the perceived trustworthiness of Airbnb host profiles[4], and the subsequent study on a computational approach to perceived trustworthiness of the host profiles[5]. We used topic coded datasets by these previous studies to fine-tune our models, and segmented the participant's results based on the categories described in their previous study.

Previously, research on the impact of text based suggestions was focused on the impact of the suggestions on writing efficiency, such as by measuring the time cost benefit of evaluating the suggestions[7]. However as text generation models have improved, these text suggestion systems have been able to provide longer sentences, that are more coherent and more contextually applicable. The research has therefore evolved more towards evaluating the impact on content.

In 2018 a study conducted by Harvard University and the Draper Laboratory in Cambridge[1] found that predictive text systems could introduce bias into participant's writing, by providing suggestions that were biased towards a specific view over another. The study looked at restaurant reviews and found that participants presented with predictive suggestions that were positive-skewed

, ,

wrote more positive reviews than they did when presented with negative-skewed suggestions.

In a 2020 study conducted at the Harvard John A. Paulson School of Engineering and Applied Sciences (SEAS)[2], researchers examined the effect of how predictive text systems changed characteristics of the participant's writing style. Their study asked participants to write captions for images, and found that those writing with an editor that provided suggestions often wrote with a lower word count, fewer descriptive words, and fewer unexpected words. Their analysis concluded that by adding the suggestions, while increasing efficiency and making writing more concise, also made the writing more predictable and changed the overall content.

3 METHOD

3.1 Experimental Set Up

Our final experiment was an online text editor that required the participants to write a profile for themselves as if they were an Airbnb host. The editor used the models we fine-tuned on specific categories to generate suggestions for the participant as they typed. Our dependent variable in this study was to measure the nudging effects of each of the models on the participant's content. We measured the nudging effects by evaluating whether each sentence in the profile was categorized as *hospitality* related, *interest* related, or *other*. This evaluation metric used sentences instead of words to get rid of the bias that could potentially be introduced by longer sentences. The independent variable was the model used, i.e either *interests* or *hospitality* model. Our experiment setup followed a between-subject setup where each participant was exposed to one set of treatment, i.e either the *interests* model or the *hospitality* model. We would consider the nudging effect to have worked if the number of sentences belonging to the same category as the data the model was fine-tuned on, was higher than the number of sentences belonging to other categories.

For the experiment, we built a web interface hosted on Google Cloud containing two different models. We generated two links, each using one of the models to generate predictive suggestions based on the text entered by the user. The interfaces looked exactly alike, without any indication of which model was present in which interface to prevent any bias from being introduced. The url of the interface just had the IP addresses and did not reveal any information about the model. The prompt on the interface and the instructions provided to the users was also the same in both links (Appendix A). Once the participants were satisfied with their writing, they could submit their text, which was saved as a text file in our Google storage bucket.

Before conducting the full fledged experiment, we conducted a small pilot study with two participants to observe and gather any feedback from their experience. Based on their feedback, we implemented some visual feedback on the interface upon submitting the form in order to indicate to the user that it has been successfully submitted. We also added the ability to save the suggestions accepted by the user in the text file along with their final submission so that we could track how the user used the predictive model.

For our experiment, we recruited participants above the age of 18, and required the participants to have decent vision to be able to use the interface. We also mandated that the participants use only a laptop/desktop as an interface to provide all participants with a uniform experience. As motivation, the participants were given the incentive to randomly win a 20$ Amazon gift card. In total, we were able to recruit a total of 31 participants for the experiment with 18 female participants and 13 male participants, ranging from age 22 to 30. We randomly allocated the models to the participants, ending up with 16 participants using the *hospitality* model and 15 participants using the *interests* model.

We collected the final results using Google Firebase in a text file format containing the text entered by the user, model assigned to the user and the suggestions from the model accepted by the participants. The text file contained no participant information to maintain their privacy.

3.2 Model Data

For our models, we used the Huggingface GPT-2 model as our baseline and went ahead to fine-tune different model versions on different categories. To fine-tune the models, we initially started with the dataset of Airbnb host profiles from the previous study [4], where the categories were manually coded. We segregated each sentence based on its category, and omitted any sentences that belonged to multiple categories. For the *hospitality* model, we extracted hospitality related data from the entire dataset. This gave us around 908 examples, with each example having an average of 13 words present. For the *interests* model, the dataset contained only 556 interests related examples with an average of 13 words per example. As we wanted the users to have uniform experience with both the models to enable us to get relevant results, we went on to get additional interests related data to get a comparable performance. We utilized the dataset generated in second Airbnb study[5] and finally trained our model with 1085 interest related examples and each example having an average of 15 words. As the hospitality data was not very noisy, we used the data as it is without any preprocessing. As the interests related data was noisy, we had to apply data preprocessing which involved collapsing multiple white spaces, line breaks and punctuation marks. We also stripped the sentences of quotation marks and other special characters. After this, we went on to ignore sentences containing less than 3 words as we believe that such sentences just add to noise as most sentences contain a subject-verb-object format. Finally we ended up with 1074 interest related examples to fine-tune our models.

3.3 Training and Evaluation

We fine-tuned the existing GPT-2 models utilizing the Trainer class of Huggingface. We split the data to be fed into each of the models into train and test using a train to test ratio of 0.9. The model parameters used for fine tuning were the number of epochs, evaluation steps, learning rate and warm up steps. Refer to Table 1 for the final parameters utilized for each of the models. The main evaluation metric for our models was the qualitative analysis of the results generated. The secondary evaluation metric was the loss values after training. For the qualitative analysis, we generated a list of ten prompts which were relevant to the model's category, and looped through the model's suggestions given each prompt about 10 to 15 times until we had roughly 200 generated samples from each model. We then developed a three point scale where 3-good, 2-ok

Table 1: Parameters used for final model

Parameters	Hospitality	Interests
number of epochs	5	5
learning rate	5e-3	5e-3
evaluation steps	1e2	1e2
warmup steps	100	100

and 1-bad and manually annotated the suggestions generated by the model. We counted the number of 3’s, 2’s, and 1’s in each model to get a score. We used this score and the general loss value of the model to evaluate models created on different single categories, as well as combinations of categories, in order to determine the models that worked the best. We narrowed down the *hospitality* and *interests* models as the best ones, and then worked to finalize those two models.

4 RESULTS

After all participants had submitted their writing, we collated the results and analyzed the profiles based on the category of each sentence in the profile, in order to evaluate the nudging effects of the models. To analyze the categories, we manually coded each sentence in each profile as either *interest* related, *hospitality* related, or *other*. The reason we chose to evaluate the categories using sentences instead of words is to prevent the results from being biased by the length of the sentences.

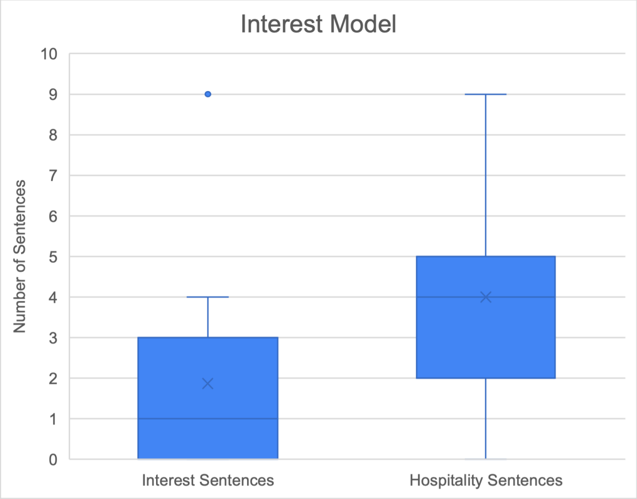

Figure 1: Interest Model Sentence Categories

We observed that both the *hospitality* and the *interests* model have higher mean and median of hospitality sentences (Figure 1, 2). This is probably due to the nature of Airbnb, people tend to talk more about their qualities as a host more than their personal interests. We can, however, see that the 75th percentile and maximum value of hospitality related sentences in the *interests* model (Figure 1) is slightly lesser than that for the *hospitality* model (Figure 2). This can also be seen when looking at the sentences per profile, where the number of hospitality sentences is slightly lower in *interest* model profiles overall than in *hospitality* model profiles. (Figure 4). This potentially suggests that the *interest* model did have some impact on the content for participants using that model, nudging it slightly more towards interests related content. We also observed some outlier presence of 9 interest related sentences in one profile in the *interest* model (Figure 3).

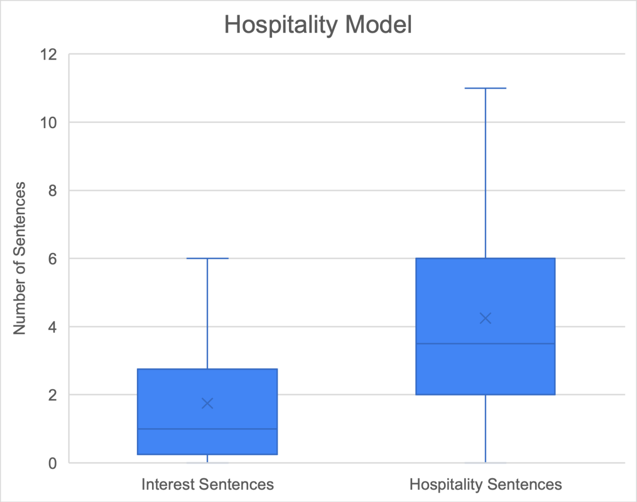

Figure 2: Hospitality Model Sentence Categories

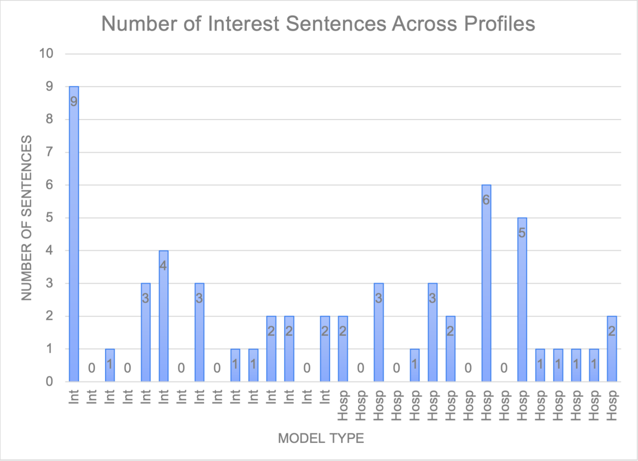

Figure 3: Number of Interest Sentences Per Profile

We also looked at additional metrics including the number of accepted suggestions per profile, the number of sentences per profile, and the number of words per profile, to assess how people used the model and give us insight into the participant’s behavior. Looking at the graphs (Figures 5, 6), we observed that the *hospitality* model has a greater number of people typing more sentences and more words overall compared to the *interests* model, although the mean (represented by the ‘x’ in the figures) number of words per profile and sentences per profile seem to be about the same for both the models. In terms of the number of accepted suggestions, we can

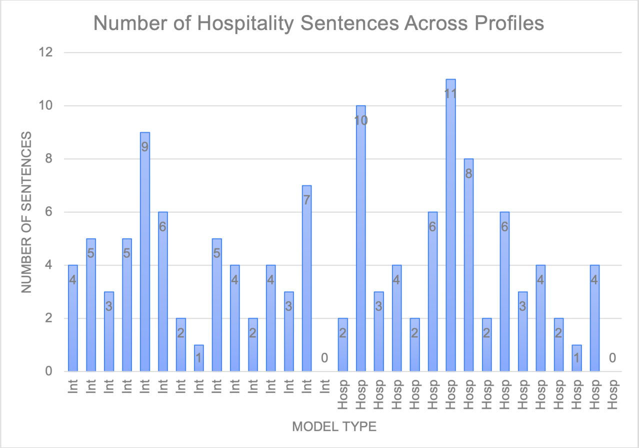

Figure 4: Number of Hospitality Sentences Per Profile

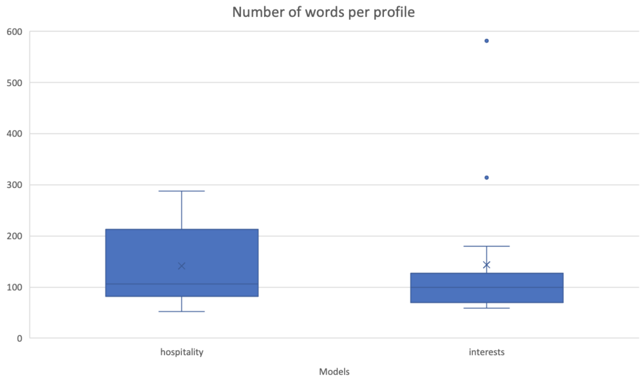

Figure 5: Number of words per profile

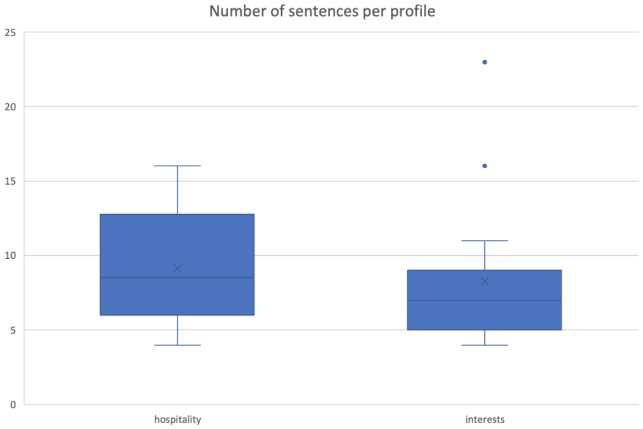

Figure 6: Number of sentences per profile

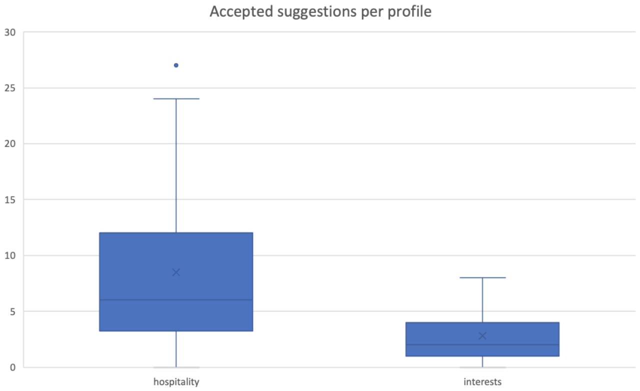

Figure 7: Number of accepted suggestions per profile

see from Figure 7 that the *hospitality* model had a higher number of acceptances than the *interests* model. This again might be because the *hospitality* model produced more contextually relevant suggestions.

Overall from our current results we cannot confidently say that there was an nudging effect from our topic specific models towards a specific category in the content. The hospitality related sentences seemed to be prevalent in both types of models. This might be because the model suggestions in the *interest* model were not as relevant, and therefore were not as effective in nudging people. However, even though the results do not seem to indicate any specific nudging effects of the models, this cannot be directly attributed to model suggestions. The problem statement of being an Airbnb host itself makes people present themselves in a hospitable way to lure people to select their listing. Apart from the quantitative results, we also received positive feedback from more number of users about the *hospitality* model over the *interests* model about how it helped them write their profile and gave relevant suggestions.

5 FUTURE WORK

Transformer models and advanced natural language generation models have only recently been around since the past few years. There is still significant research to be done on the effects of their predictive suggestions on the quality and content of a user's writing. In the future, we plan to explore how we can study the influence of predictive text generation in different online circumstances and contexts, beyond just Airbnb. We theorize that using a different context where the problem statement is more defined and unbiased will help us draw more concrete conclusions about the effects of predictive text on user content. We also plan to explore more robust techniques and better training mechanisms to ensure more relevant suggestions from our models and decrease the randomness and irrelevant suggestions. Also we plan to study more powerful text prediction model architectures such as GPT-3 or GPT-Neo and see how to integrate the same.

REFERENCES

[1] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing. In *Proceedings of the 44th Graphics Interface Conference* (Toronto, Canada) *(GI '18)*. Canadian Human-Computer Communications Society, Waterloo, CAN, 42–49. https://doi.

org/10.20380/GI2018.07

[2] Kenneth C. Arnold, Krysta Chauncey, and Krzysztof Z. Gajos. 2020. Predictive Text Encourages Predictable Writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 128–138. https://doi.org/10.1145/3377325.3377523

[3] Stacey Lastoe. 2019. British couple spends $11,800 on Airbnb rental in Ibiza that doesn't exist. https://www.cnn.com/travel/article/airbnb-ibiza-spain-penthouse-scam-trnd/index.html?fbclid=IwAR1DBqoFEqdw9qVauFNuX0oosbSD8rzXy_MEumXPu4wJN2nT10VFHIC22nE

[4] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. 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* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2397–2409. https://doi.org/10.1145/2998181.2998269

[5] Xiao Ma, Trishala Neeraj, and Mor Naaman. 2017. A Computational Approach to Perceived Trustworthiness of Airbnb Host Profiles. *Proceedings of the International AAAI Conference on Web and Social Media* 11, 1 (May 2017). https://ojs.aaai.org/index.php/ICWSM/article/view/14937

[6] OpenAI. 2019. *Better Language Models and Their Implications*. Retrieved May 17, 2021 from https://openai.com/blog/better-language-models/

[7] Philip Quinn and Shumin Zhai. 2016. A Cost-Benefit Study of Text Entry Suggestion Interaction. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 83–88. https://doi.org/10.1145/2858036.2858305

A STUDY PROMPT

"For the prompt we ask that you pretend you are a host on Airbnb, write a bio that tells potential travelers about you and your background. You want them to get to know you and get a sense for what you are like as a host. (Please think of this as your host bio and not about a specific listing.)The editor will provide suggestions as you type, you can either accept suggested words by pressing 'Tab', accept entire sentence by pressing '->' (right arrow), generate a different suggestion by pressing 'Escape', or ignore them and keep typing. Once you are finished click 'Save and Finish' on the text editor. We require you to use a laptop or computer as the interface for this experiment. Once you are done, please send the screenshot including the timestamp. You will be entered in a chance to win a 20$ Amazon giftcard upon writing a relevant bio."