How do people interact with biased text prediction models while writing?

Advait Bhat
IIT Bombay
advaitmb@gmail.com

Saaket Agashe
VJTI, Mumbai
saaket.agashe@gmail.com

Anirudha Joshi
IIT Bombay
anirudha@iitb.ac.in

Abstract

Recent studies have shown that a bias in the text suggestions system can percolate in the user's writing. In this pilot study, we ask the question: How do people interact with text prediction models, in an inline next phrase suggestion interface and how does introducing sentiment bias in the text prediction model affect their writing? We present a pilot study as a first step to answer this question.

1 Introduction

Recent years have seen a great improvement in probabilistic language models and with the advent of transformer based language models, text generation has become sophisticated and viable enough to be used in real time writing interfaces with next phrase suggestions.

Each writer carries a certain opinion that they wish to put forth through the write up. For example a movie review writer may have liked or disliked a certain movie with a certain intensity. Subsequently, they would use a specific vocabulary and style that would display their intent to the readers. Any technology that acts as a mediating interface between people and the world, wold have the potential to influence their writing process and therefore possibly even their writing intent. In our example, a text editor equipped with predictive system, would acts as the mediator and may affect the writing process and intent of the user, thereby making the movie seem better or worse than the writer intended; as well as influencing the quality of the writing itself. This, therefore, becomes an interesting area that needs investigation.

While researchers have quantitatively tried to study the effect of biased text prediction systems on user's final written content, qualitative analysis of a writer's writing process in the presence of text suggestions and furthermore, bias, is not pervasive. Along with that, studies so far have used n-gram based Language Models or Recurrent Neural Networks, but not transformers which are capable of generating much more relevant suggestions.

In this study, we try to understand how people interact with text prediction models in an inline next phrase suggestion interface and how does introducing sentiment bias in the text prediction model affect their writing. We ask people to write movie reviews using an intelligent text prediction interface and try to study their writing process. Movie reviews are particularly useful for such a scenario, as every movie review has a star rating associated with it and this acts as a good and reliable proxy to quantify the writer's original intent. Along with this, writers are encouraged to express their personal opinions through their movie reviews while having to follow a sufficiently standard structure.

We present a pilot study as a first step to answer this question. We present a background of the work in this area, our method, which includes the design of the interface, text generation methods and validation of our language models. We then present qualitative analysis of the same.

2 Background

In a recent study (Arnold et al., 2018), introduce sentiment bias in predictive text suggestions on smartphone keyboards for a hotel review writing task. The study demonstrates that if predictive text suggestions are positively sentiment biased for a hotel review writing task, writers tend to write more positive reviews. On the other hand, equivalent effects aren't seen when the system is negatively sentiment biased. Taking these studies as a starting point, we wish to understand how users interact with such a system and how their writing gets affected qualitatively. We also use more updated language models for our study. The larger aim would be to create a model of how users interact with such a system and where this fits into their writing process; similar to cognitive process theory

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 116–121

of writing by (Flower and Hayes, 1981). In this paper, we present a pilot study with six users where we qualitatively analyse their writing process.

3 Method

3.1 Experiment

In this pilot, we ask users to write movie reviews. We also ask them to rate these movies in terms of star rating. In order to compare if the bias in the system affects the user's writing and therefore the perception of the reader who would read it, we randomise these movie reviews and then ask independent readers to read these reviews and guess the rating the movie would've gotten.

Our pilot study had a between and within subject, mixed study design. We assigned two movies to each participant. 3 participants received 2 movies that had low ratings on IMDb and 3 participants received 2 movies that had high ratings on IMDb. We expected that the participants who received movies with low IMDb ratings would also give lower ratings to the movies, as these movies are generally considered bad movies and vice versa for movies with good ratings. We did this to ensure that we get an equal distribution of participants in every condition and so that we covered every condition. Half of the participants were given a pair of below average movies and the other half, above average. Among these individual groups, every participant was given a baseline interface and an interface with suggestions. Within the interfaces with suggestions 2 participants were given a positive biased system, 2 were given a negatively biased system and the remaining 2, neutral.We, therefore had 12 total conditions out of which we could test 6 out leaving out counterbalancing the movie sequence for the pilot.

3.2 Interface Design

The experiment includes a simple text editor with word complete and next phrase suggestion capabilities. Figure 1 shows the interface and the distinction between word and phrase complete. Every time a user presses any key; after a delay of 250ms, a word complete suggestion is displayed. The user can then choose to accept the suggestion by pressing tab or right arrow key or can ignore the suggestion by continuing to type. Every time the user pauses for more than 500ms, the next phrase suggestion is displayed. The user can press tab to accept the subsequent word in the phrase. When the user does that, the word gets selected and added to the input field.

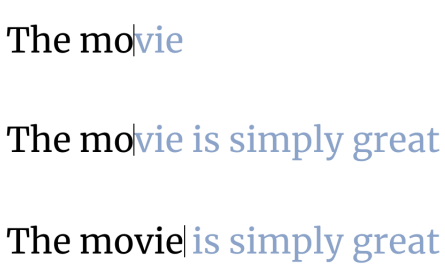

Figure 1: The first line shows a "Word Complete" instance, a user can accept the word as they write. The second line shows a combination of "Word Complete" and "Phrase Complete" instance, user can press tab and accept the word suggestion and then as many words from phrase suggestion that they need. Third is a 'Phrase Complete' Instance

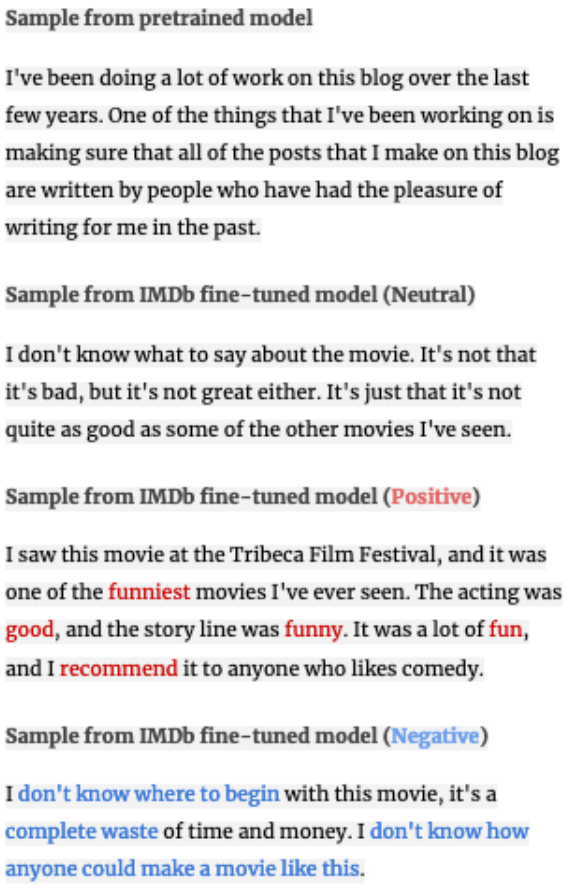

Figure 2: Generated samples from pre-trained, fine-tuned neutral, fine-tuned positive biased and fine-tuned negative biased models

According to the system that users are allotted, users are presented with word complete and next word/phrase suggestions that are either positively or negatively biased or do not have a bias.

3.3 Next Phrase Suggestion Model

We have used the GPT-2 (Radford et al., 2019) pretrained model and fine tuned it on the IMDb movie reviews corpus to create the next phrase suggestion model. We have trained one model on positive reviews, one on negative reviews and another general model on both positive and negative reviews. We then use these models to generate next-phrase and word complete suggestions for users. We fine-tuned each model for 3 training epochs and obtained a test perplexity score of 36.9713 for positive model, 34.6978 for negative model and 32.297 for the neutral model. The figure includes sample texts generated by the various fine-tuned models as well as a sample from the original pretrained model for comparison.

3.4 Sentiment validation

We performed a validation step to ensure that the models trained in such a way are able to generate sentiment controlled suggestions. We used 500 positive and 500 negative test reviews as prompts and used each of the above models to generate text using said prompts. We then pass the generated text (Just the part of the text that was newly generated leaving out the original prompt as that might affect the performance of sentiment validation process) through a BERT sentiment classifier which had been fine-tuned for the IMDb movies reviews sentiment classification task. Here we see that the text generated by the positive model has a mean sentiment logit score of 2.3 and those generated by the negative model have a mean sentiment logit score of -1.6 (The higher the score, the more positive the sentiment).

3.5 Model Usefulness

We describe the model usefulness as the ability of the model to output the same text as the user, given the user's previously typed text. This way of calculating usefulness of the model is inspired by (Arnold et al., 2020). We compute this by running over the sample of user texts which have been written without using text completion. We start with the user's first word and have the model compute the next phrase and we increase the usefulness count by 1 for every consecutive word starting from the first word that the model generates correctly. We then consider both the first and second word typed by the user and generate text using those words, and again update the usefulness count. We continue this process for the entirety of the user's text and finally divide the usefulness count by the total number of words in user input to get the usefulness score. We do this process 5 times for each user and average the scores. We then average the scores for all users to get a usefulness score for the entire model. Higher the score, more relevant suggestions the model gives. We propose this as a way to validate the real-world usefulness of a model in the next phrase suggestion setting. We obtained the following scores from our models and include the score for an AWD-LSTM model fine-tuned on the same IMDb corpus for comparison.

Model	Usefulness
Positive Fine-Tuned GPT-2	0.3
Negative Fine-Tuned GPT-2	0.29
Neutral Fine-Tuned GPT-2	0.29
Neutral Fine-Tuned AWD-LSTM	0.23

Table 1: Usefulness Score Comparison Across different models

3.6 Users

The pilot included 6 participants. Participants had a primary and secondary education in an English medium school. None of the participants were native English speakers and all of them considered English to be a second or third language. All the participants were fairly exposed to western cinema in general. We also made sure that the users had previously used next-phrase / word complete interfaces like Google Smart Compose. Most participants did not have experience writing movie reviews before, but almost all of them had read movie reviews at multiple points of time. Participants were recruited through social media platforms.

3.7 Analysis Methods

3.7.1 Text Analysis

In order to analyse the text qualitatively, we designed a simple tool to simulate and replay the writer's writing process. The tool had play, pause, rewind and fast-forward functionalities which helped us control the writing process like a video. As seen in Figure 3, we use colours to tag Phrase Completes, Word Completes and Suggestions and

replay the entire writing process of the reviewer and analyse the text and user's behaviour. Tagging written content according to user actions gives us an insight into the users writing process and helps us code user behaviour which we can then analyse. We used affinity mapping as a qualitative data analysis approach to analyse these writing processes.

3.7.2 Interviews

After the users wrote their reviews, we conducted a conversational interview. We tried to understand their experience while using the tool and the suggestions, asked them if they remembered any peculiar incidents, and if they had any feedback about the system and the suggestions. We recorded these interview sessions, coded them and analysed them using affinity mapping.

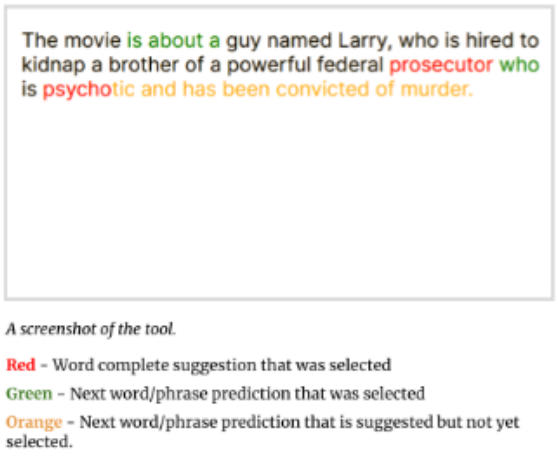

Figure 3: Visualization of the keystroke log with written text tagged with corresponding user actions.

4 Results And Analysis

4.1 Users used suggestions as prompts for writing.

While traditionally, the aim for having text suggestions in a text editor is to save time and effort while writing, users also used these suggestions as prompts for writing. Users reported that they got ideas about sentence structure, vocabulary, and even content from the suggestions. Almost all users reported that suggestions helped them frame their sentences. Some users also stated that writing reviews were easier with suggestions than without any suggestions because of the same.

4.2 Some users considered the suggestions as an authority on how to write movie reviews.

Most users in the study were amateur movie reviewers. Some of them took cues from the suggestions on how to write movie reviews. Users had a notion that the suggestions are 'based on the writing of other (perhaps more experienced) reviewers' while they did not know the exact mechanism through which these suggestions are generated. Users also reported that at times they second-guessed their own writing because of the sentences suggested. As these were inexperienced reviewers, they doubted whether their sentence structure and vocabulary was appropriate for writing a movie review. This could also be a function of users being non-native English speakers.

4.3 Influence on writing

Users reported that they selected a suggestion as long as it was 'more or less' similar to what they wanted to write. While this was true for most users, the amount of deviation (from the sentence they had in mind) that they would allow differed between users. There, at times, seemed to be a compromise between being true to what they wanted to write, on one hand, and saving effort by selecting suggestions, on the other. Some users reported that they wrote the 'key words' of a sentence on their own but used suggestions to enter 'filler' or 'generic' words. That being said, users' vocabulary did get influenced by the suggestions. On multiple instances, users took inspiration from previous suggestions and used the vocabulary appearing in those suggestions for composing new sentences.

4.4 Users sensed the bias in the system.

Most users could sense that the system was biased towards a particular sentiment. When there was a mismatch in user's intent and system bias, a user rightly pointed out that "The suggestions I got were positive, but the review I was writing was on the negative side." Similarly, users could also sense the extent to which the suggestions were biased. One user mentioned that, "The sentences which i was writing were negative and so were the suggestions, but the words chosen were very different. It changed the meaning of what I wanted to say and made it more extreme." Another user perceived the suggestions to have some form of intention. They said, "When I write something negative, the sug-

	Model Bias	Writer rating	Guessed Rating	Word Completes (WC)	Phrase Completes (PC)	Mean Words Accepted per PC	Characters Deleted (CD)	Words Deleted (WD)
0	Pos	7	7.7	8	53	3.857143	152	8
1	Neg	6	6.7	2	17	2.428571	75	8
2	Pos	3	2.3	19	25	1.625000	274	19
3	Neu	10	9.3	1	14	5.000000	157	8
4	Pos	2	1.7	14	33	3.888889	213	20
5	Neg	6	5.3	6	21	4.200000	293	22

Table 2: Writer rating is the star rating that the writers gave after watching the movie. Guessed ratings are star ratings which were given only on the basis of the reviews written by corresponding writers by third party readers. We mention the Word Completes and Phrase Completes that were used by each participant. We also mention the characters and words that were deleted by each user. Furthermore, we also show the average words that the user accepted in every phrase complete instance

gestions try to cover it up, most of the time. Not every time."

4.5 Users often ended up deleting selected phrase suggestions

When asked whether their writing was influenced because of the bias in the suggestions that they had noticed, users often said it wasn't and that they decided what they wanted to write before they started writing. While this was true, users did select a significant amount of next word/phrase predictions, but they also ended up deleting a considerable chunk of this text and replaced it with their own typed text. It might be that the suggestions felt right in the moment, but did not flow well with the argument they were trying to make and hence on a second glance, users often ended up editing these sentences or straight up deleting them.

4.6 Users used the suggestion system creatively with their sentences

Users used the suggestions as tools to anticipate and iterate with different possible sentence constructions to arrive at a desirable final sentence. Similarly, by evaluating how the predicted sentence sounded, users used next phrase prediction to check appropriateness of their sentence's grammar and construction. Users also selected multiple full phrases to see what direction they take and then deleted them but took inspiration from them to write their sentences. Likewise, for word complete, users completed suggested words like 'interesting' and then deleted the last 3 letters to write what the intended to write ie. 'interest'.

4.7 Users sometimes ignored suggestions even when the suggestions matched

There were multiple instances for almost every user where users chose to type the full word out even when word complete or next word suggestions matched precisely to what they were typing. According to what the users reported in the interviews, they did not pay much attention to the suggestions when they were exactly sure about what they wanted to write. This could be one explanation for why users typed the whole word even when the same word was suggested.

4.8 Suggestions disrupted the user's thought process

When a suggestion was not relevant to what users wanted to write, it disrupted their thought process, often causing them to forget what they originally had planned to write. They were particularly disappointed when the irrelevant suggestions were long, since the effort put into reading them felt like a waste.

5 Conclusion

In this pilot study, we qualitatively analysed the writing process of the users while they used an inline next phrase suggestion interface with sentiment bias. This was our first step towards coming up with a theory for how users interact with text prediction models and how introducing sentiment bias in the text prediction model affects users' writing. We conducted the pilot study with 6 users. While we could come up with some interesting themes to describe users' writing process, we are much far away from coming up with a robust theory. We intend to do so in the future.

References

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*, GI '18, page 42–49, Waterloo, CAN. Canadian Human-Computer Communications Society.

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*, IUI '20, page 128–138, New York, NY, USA. Association for Computing Machinery.

Linda Flower and John R. Hayes. 1981. A cognitive process theory of writing. *College Composition and Communication*, 32(4):365–387.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.