$9^{th}$ International Conference on Computational Social Science $IC^2S^2$
July 17-20, 2023, Copenhagen, Denmark

# AI Crowdturfing on Yelp: Identifying Patterns in Machine-Generated Fake Reviews

*AI-generated Content, Fake Review Detection, GPT, Social Media, Natural Language Processing*

## Extended Abstract

Online reviews have served as valuable signals about product quality that may bridge information asymmetries between customers and sellers in online marketplaces, which in turn influence customer purchases [1]. With the prevalence of review systems embedded in social media, fake reviews that contain deliberately deceptive information are also witnessed to proliferate on these platforms [4]. Given that the fundamental value of reviews is rooted in their authenticity, fake reviews would harm customers and severely threaten to erode their trust in online reviews and damage the reputation of social media [3]. Major social media platforms propose countermeasures to fight review fraud with both manual analyses by the content moderation team and automated systems. However, they still face significant challenges to counteract machine-generated fake reviews effectively due to advances in generative models.

In this study, we aim to detect patterns in machine-generated fake restaurant reviews that were created using high-quality elite reviews from Yelp. We specifically hypothesized that AI crowdturfing campaigns based on GPT-3 were implemented after its release in 2020, given its easy accessibility and low usage costs through its API. To begin, we collected 447,295 raw reviews associated with 5,959 New York restaurants, including information such as the rating and details about the users and restaurants. We separated the reviews into the elite (verified and authentic) and non-elite categories. Then, we used 12,000 elite reviews to generate fake reviews using OpenAI's GPT-3 model. We conducted a human study, where we asked 80 participants to identify the AI-generated reviews from a set of 40 pairs of reviews containing one authentic and one synthetic review. Our results indicated that humans could only identify fake reviews with an accuracy score of 57.13% (std 13.57%), showing that they often cannot distinguish between real and AI-generated content. Therefore, we employed machine learning techniques to perform the same task. We fine-tuned a pre-trained GPT-2 model to classify fake and real reviews by maximizing Youden's J statistics on the validation set and searching for the best classification threshold. Our model achieved an accuracy score of 94.88% and an F1-score of 94.80% at $J^*$=.5948 out of the sample, indicating that machines significantly outperform humans. Next, we conducted ANOVA tests on 131,266 unverified non-elite reviews posted after 2020. We compared differences across two predicted categories for various variables related to the review, user, and restaurant. The *review*-based variable was the *Rating*, while the *user*-based variables were the number of friends (*#Friends*), the number of previously posted reviews (*#Reviews*), and the number of previously posted photos (*#Photos*). The *restaurant*-based variables included the average restaurant rating (*AvgRating*), price level (*PriceLevel*), chain status (*ChainStatus*), number of restaurant reviews (*#RestReviews*), and the number of visits, both raw (*#Visits*) and normalized (*NormVisits*). We aggregated the last two variables from the NY SafeGraph mobility dataset.

As such, we found that 6.80% of the reviews were identified as fake using the optimized threshold. Such reviews were associated with certain characteristics. The AI-generated reviews had a higher average *Rating* (+.24, $p < .001$), but were connected to users with lower average

*#Friends* (-8.9, $p < .001$), lower average *#Reviews* (-10.66, $p < .001$), and lower average *#Photos* (-32.62, $p < .001$). Additionally, predicted fake reviews were associated with restaurants that had a greater average *#RestReviews* (+32.12, $p < .01$) but had lower average *#Visits* (-207, $p < .01$). Sensitivity analyses were conducted for each category of variables, which are presented in Figures 1 and 2.

Finally, we examined the writing style of the two predicted classes. Writing style refers to how a text is constructed, sentence by sentence and word by word. We evaluated three categories of metrics: *perplexity*-based, *readability*-based, and *sentiment*-based metrics. First, *perplexity*-based metrics include *Perplexity* (*PPL*) and *Textual Coherence* (*TC*). *PPL* was calculated using Equation 1, while *TC* measures the change in *PPL* after randomly shuffling the sentences in each document. We used a zero-shot pre-trained GPT-2 model to calculate both scores. Second, for *readability*-based metrics, we considered the *Automated Readability Index* (*ARI*) as in Equation 2, *Number of Difficult Words* (*#DW*), and *Readability Time* (*RTime*). Finally, the only *sentiment*-based metric was the SiEBERT Sentiment score [2]. Predicted AI-generated fake reviews exhibited a lower average *PPL* (-14.93, $< .001$) and higher *TC* (+.18, $p < .05$), indicating that they were more predictable, repetitive, and grammatically correct. Additionally, AI-generated reviews were easier to comprehend, requiring a lower educational grade level, as measured by *ARI* (-.35, $p < .001$) and *#DW* (-3.07, $p < .001$), and had a more positive tone (+.13, $p < .001$). Figure 3 presents a sensitivity analysis across different thresholds.

In summary, while refraining from making causal claims, we have delineated the accessibility of disseminating fake AI reviews generated on social media platforms. Such accessibility may lead to the proliferation of restaurant crowdturfing campaigns due to the advances in large language models, such as ChatGPT, aimed at distorting user experiences. The study provides evidence that AI-generated fake reviews are becoming more sophisticated and can easily deceive human readers. Therefore, it is imperative for policymakers to develop regulations that require online review platforms to implement tools and processes to detect and remove fake reviews. The study also underscores the need for online review platforms to invest in better detection tools for AI-generated text. As the technology used to generate fake reviews becomes more advanced, review platforms must keep pace with technological advancements to ensure they can detect and remove such content effectively.

# References

[1] Wenjing Duan, Bin Gu, and Andrew B. Whinston. Do online reviews matter? — an empirical investigation of panel data. *Decision Support Systems*, 45(4):1007–1016, 2008.

[2] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 2022.

[3] Yoon Jin Ma and Hyun-Hwa Lee. Consumer responses toward online review manipulation. *Journal of Research in Interactive Marketing*, 8(3):224–244, 2014.

[4] Myle Ott, Claire Cardie, and Jeff Hancock. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, page 201–210, New York, NY, USA, 2012. Association for Computing Machinery.

$$PPL(X) = \exp\left[-\frac{1}{t}\sum_{1}^{t}\log p(x_i|x_{<i})\right] \tag{1}$$

$$ARI = 4.71\frac{\#Chars}{\#Words} + 0.5\frac{\#Words}{\#Sentences} - 21.43 \tag{2}$$

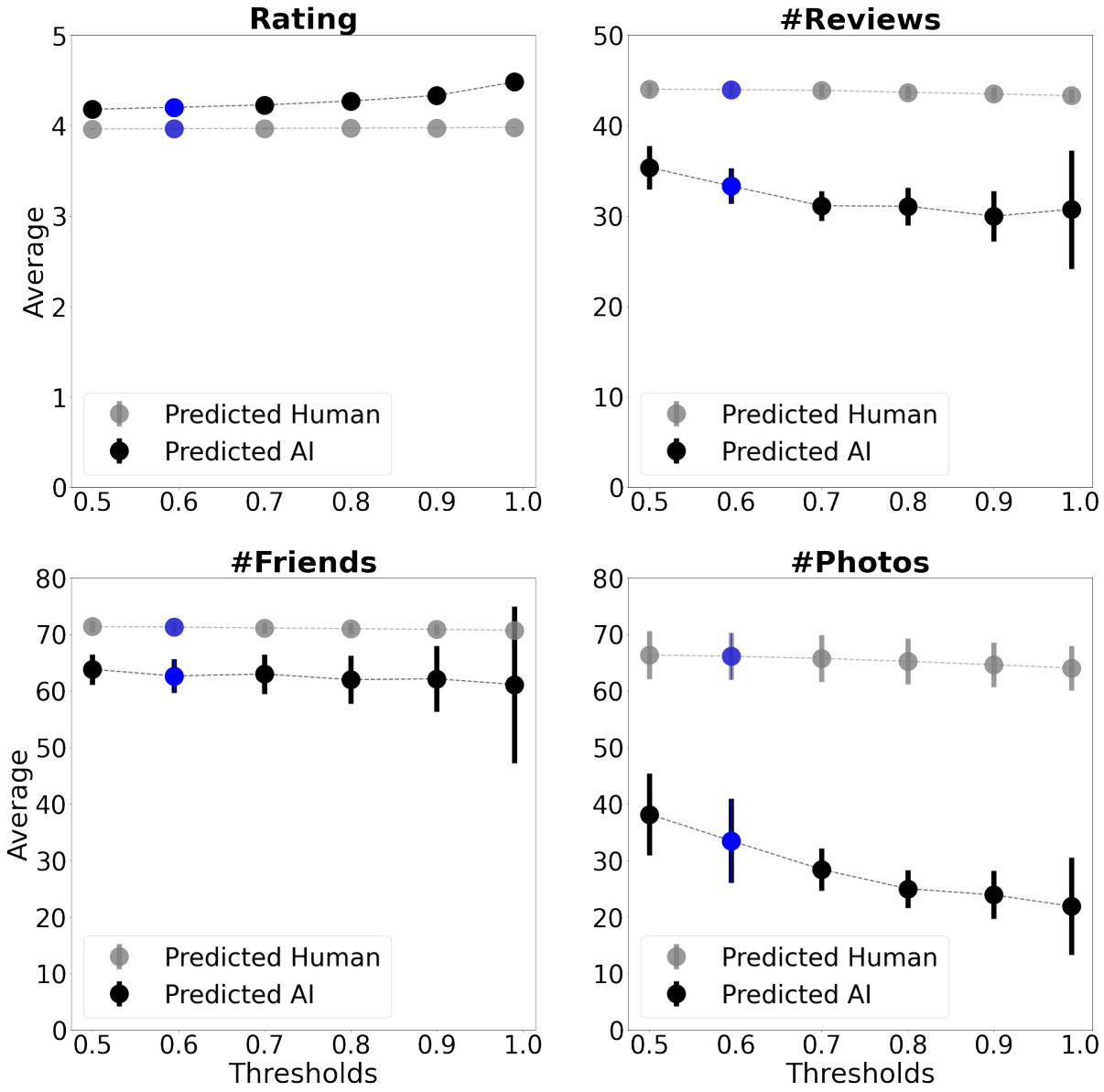


Figure 1: Sensitivity analysis of review-based and user-based variables. In blue, highlighted $J^*$ optimal threshold value. Error bars are the confidence intervals at the .05 significance level.

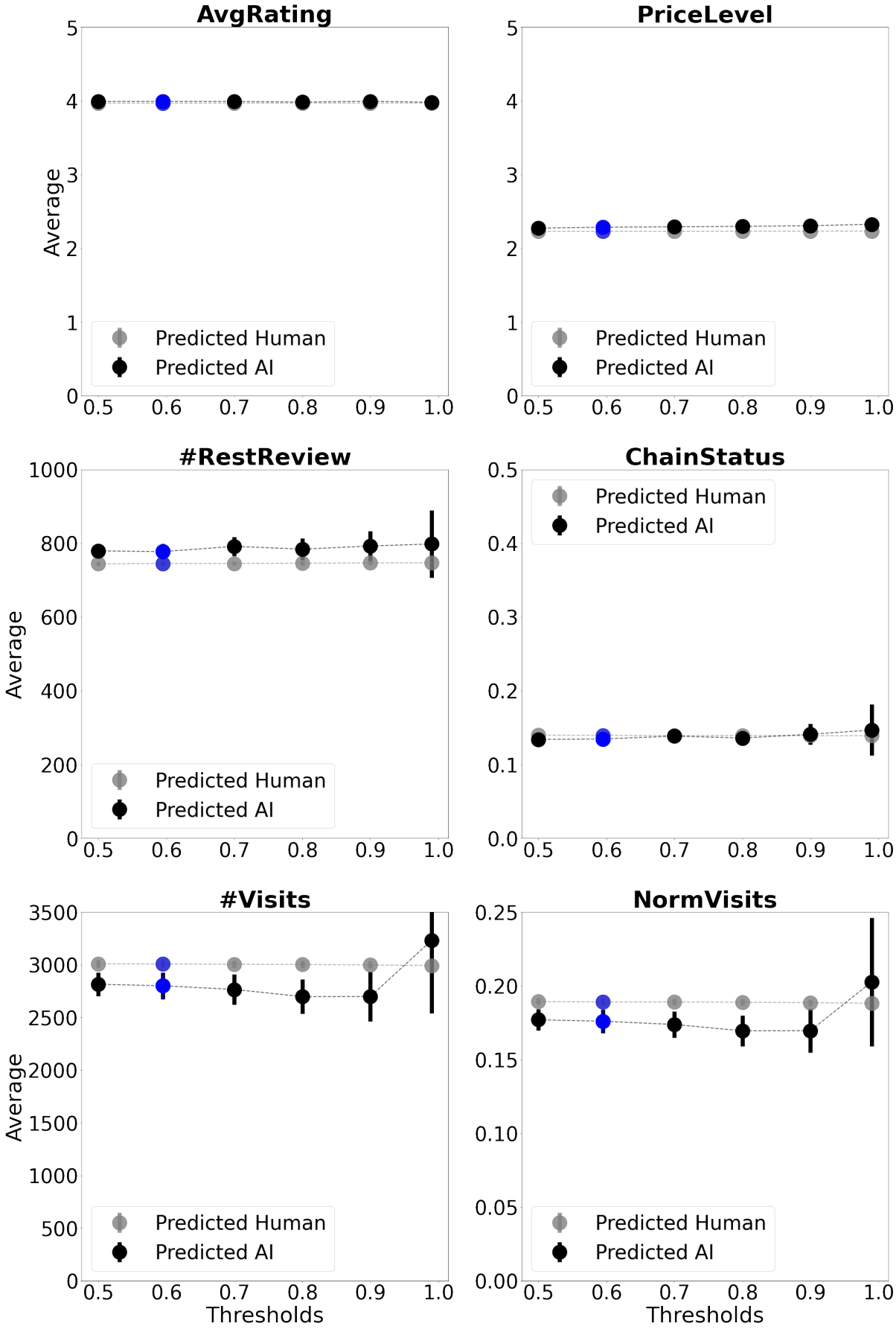


Figure 2: Sensitivity analysis of restaurant-based variables. In blue, highlighted the $J^*$ optimal threshold value. Error bars are the confidence intervals at the .05 significance level.

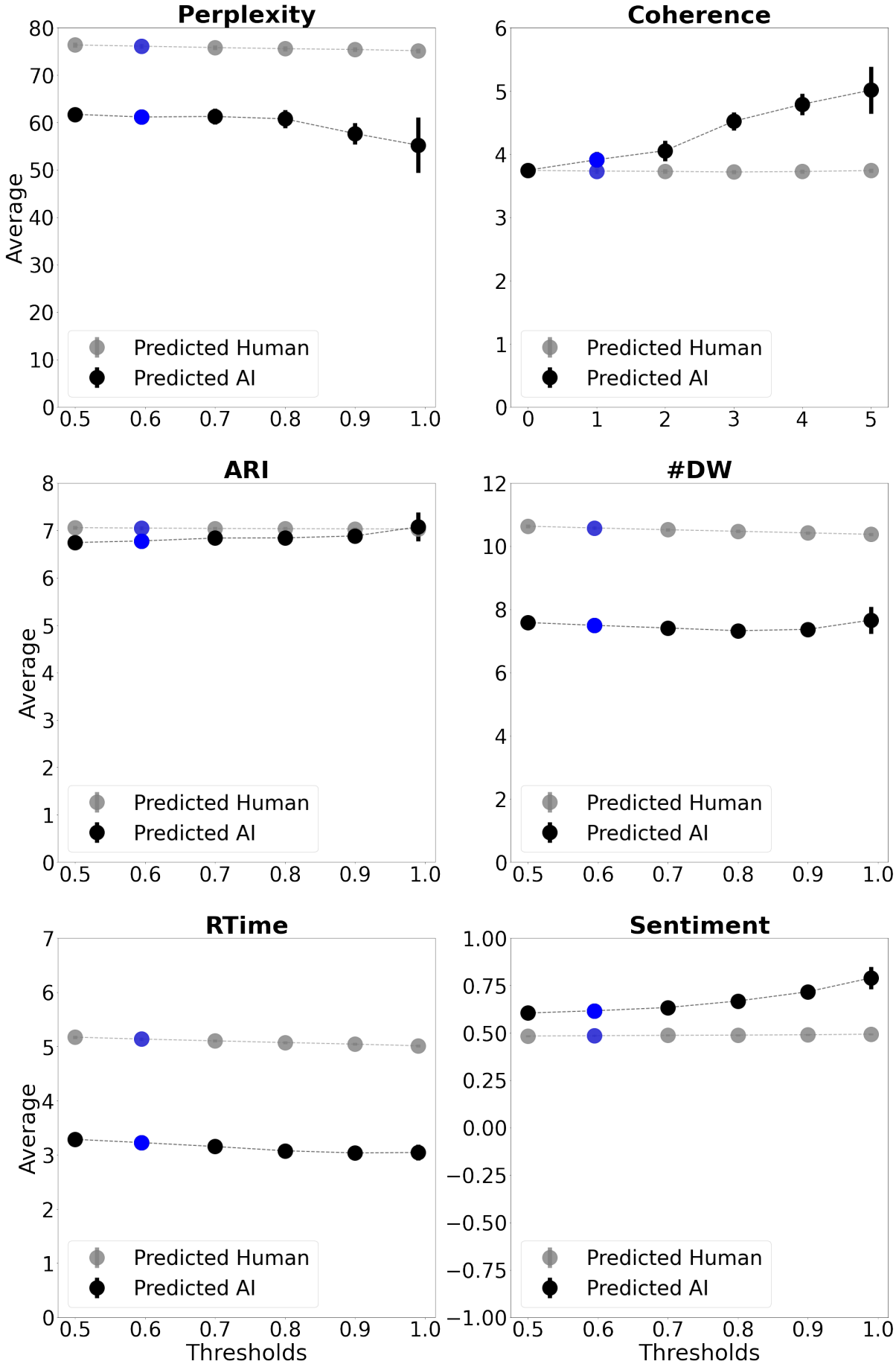


Figure 3: Sensitivity analysis of writing-style variables. In blue, highlighted the $J^*$ optimal threshold value. Error bars are the confidence intervals at the .05 significance level.