

Reiss, Michael V.; Roggenkamp, Hauke

**Working Paper**

## A Comment on "Negativity Drives Online News Consumption"

I4R Discussion Paper Series, No. 199

**Provided in Cooperation with:**

The Institute for Replication (I4R)

*Suggested Citation:* Reiss, Michael V.; Roggenkamp, Hauke (2025) : A Comment on "Negativity Drives Online News Consumption", I4R Discussion Paper Series, No. 199, Institute for Replication (I4R), s.l.

This Version is available at:

<https://hdl.handle.net/10419/311304>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*

No. 199

DISCUSSION PAPER SERIES

# A Comment on “Negativity Drives Online News Consumption”

Michael V. Reiss

Hauke Roggenkamp

**This paper received a response:**

Robertson, C.E., N. Pröllochs, P. Pärnamets, J.J. Van Bavel, and S. Feuerriegel. 2025. Response to Replication Report of “Negativity Drives Online News Consumption”. *I4R Discussion Paper Series* No. 200. Institute for Replication

**February 2025**

## I4R DISCUSSION PAPER SERIES

I4R DP No. 199

# **A Comment on “Negativity Drives Online News Consumption”**

**Michael V. Reiss<sup>1</sup>, Hauke Roggenkamp<sup>2</sup>**

*<sup>1</sup>Leibniz-Institute for Media Research, Hamburg/Germany*

*<sup>2</sup>University of St. Gallen/Switzerland*

FEBRUARY 2025

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

### **Editors**

**Abel Brodeur**  
*University of Ottawa*

**Anna Dreber**  
*Stockholm School of Economics*

**Jörg Ankel-Peters**  
*RWI – Leibniz Institute for Economic Research*

# A comment on “Negativity drives online news consumption”

Michael V. Reiss\*      Hauke Roggenkamp\*\*


December 30, 2024


We examine the reproducibility and robustness of the central claims from Robertson et al. (2023) who investigate the impact of negative language on online news consumption by analyzing over 12,448 randomized controlled trials on upworthy.com. Applying “lexical” sentiment analyses, the authors make two central claims: first, they find that headlines with negative words significantly increase click-through rates (CTR). Second, they find that positive words in a headline reduce a news headline’s CTR. Our reproducibility efforts include two different techniques: using the same data and procedures described in the study, we successfully reproduce the two claims through a blind computational approach, with only minor and inconsequential discrepancies. When using the authors’ codes, we reproduce the two claims with identical numerical results. Examining the robustness of the authors’ claims in a pre-registered third step, we validate and apply a “semantic” sentiment analysis using two large language models to re-compute their independent variables describing negativity and positivity. While we find support for the negativity bias, we do not find semantic (in contrast to lexical) positivity to reduce online news consumption.

## 1 Introduction

This report is prepared as part of a collaboration between the Institute for Replication and Nature Human Behaviour (see Brodeur, Dreber, et al. 2024) and responds to recent calls for replication studies in communication science (Bowman 2024; Breuer and Haim 2024; Dienlin et al. 2021; Freiling et al. 2021). This collaboration aims to systematically reproduce and replicate studies published in Nature Human Behaviour from 2023 onward, spanning fields

---

\* [M.Reiss@leibniz-hbi.de](mailto:M.Reiss@leibniz-hbi.de) , Leibniz-Institute for Media Research, Hamburg, Germany.

\*\* [Hauke.Roggenkamp@unisg.ch](mailto:Hauke.Roggenkamp@unisg.ch) , Institute of Behavioral Science and Technology, University of St. Gallen, Switzerland and Faculty of Economics & Social Sciences, Helmut-Schmidt-University, Hamburg, Germany. Both authors contributed equally.

including anthropology, epidemiology, economics, management, politics, and psychology, with findings compiled into a meta-paper that will be considered for publication as a research article (subject to peer review).

As part of this broader initiative, we examine the reproducibility and robustness of the two central claims from Robertson et al. (2023), who investigate the impact of negative language on online news consumption. The authors use a large dataset that reports results of over 22,000 A/B-tests with about 105,000 different variations of news headlines on [Upworthy.com](https://www.upworthy.com) (see Matias et al. 2021). Applying a *lexical* sentiment analysis paradigm by using a dictionary to count the frequency of positive and negative words within a headline, they analyze 12,448 of these A/B-tests (which account for 53,699 headline variations) conducted between January 24, 2013 and April 30, 2015. The authors find that headlines with negative words significantly increase click-through rates (CTR). Conversely, positive words in headlines decrease the CTR.

Taken together, the authors make two central claims:

1. “Consistent with the ‘negativity bias hypothesis’, the effect for negative words is positive [ $\beta = 0.018$ ,  $SE = 0.003$   $z = 6.942$ ,  $P < 0.001$ , 99% CI=(0.011, 0.025)] suggesting that a larger proportion of negative words in the headline increases the propensity of users to access a news story.” (p. 814)<sup>1</sup>
2. “In contrast, the coefficient for positive words is negative [ $\beta = -0.017$ ,  $SE = 0.003$   $z = -6.589$ ,  $P < 0.001$ , 99% CI=(-0.023, -0.010)], implying that a larger proportion of positive words results in fewer clicks.” (p. 814)

Considering the authors’ title, theory development and reporting of results, we consider the former to be their main claim, whereas the latter claim is secondary.

To examine these claims, we consider two distinct yet complementary research questions. While Robertson et al. (2023) asked whether the frequency of negative and positive words (*lexical* sentiment) affects click-through rates, we investigate whether the sentiment expressed in headlines, when analyzed in context (*semantic* sentiment), drives user engagement. This shift from a lexical to a semantic approach reflects both established and recent advances in computational methods. Specifically, we employ two language models: DistilBERT, based on the BERT architecture that was available at the time of the original study, and GPT-4, representing the current state of the art. This dual approach allows us to examine the robustness of the negativity (and positivity) bias with a different sentiment conceptualization.

We build on the authors’ publicly available replication package to conduct both reproducibility and robustness checks. Doing so, we proceed in four steps: first, we check the data availability after searching the cleaning and analysis codes for data inputs. Second, we reproduce the

---

<sup>1</sup> On page 814 the authors report results of a multilevel binomial random effects model (random intercepts, fixed slopes). Because their preferred model is a specification with random intercepts and *random* slopes, we report these results instead. We discuss this further in Section 5.3.

relevant parts of their analysis by only following the procedure described in their methods section (i.e., blind, without looking at their code). Next, we check the unblind computational reproducibility, that is, the extent to which results in the original study can be reproduced using both the data and the code from the replication package (Brodeur, Mikola, et al. 2024, 5). We are able to blindly reproduce the authors' main claims with minor numerical differences. The unblind procedure leads to identical results.

When accounting for context using large language models (i.e., DistilBERT and GPT-4o) in a final step, we find mixed results. With our preferred specification we find the main claim to be robust to our sentiment conceptualization: we find a statistically significant effect of semantical negativity in the same direction as in the authors' analysis. Whereas the effect size is larger, it is not significantly different from the authors' reported lexical effect (see Table 7). However, for the secondary claim about positivity, we find a clear difference: unlike lexical positivity which reduced CTRs, semantic positivity shows inconsistent results that differ with variations in magnitude, significance, and even direction.

This document is organized as follows: we briefly report the results of the data availability and blind recoding as well as the unblind reproduction check in Section 2, Section 3, and Section 4, respectively. We then describe the robustness checks, as well as the corresponding methods, validation, and results in Section 5. Section 6 concludes. Throughout the document, we focus on the two claims identified above and refer to our osf repository, which can be found here: <https://osf.io/bfhdw/>

## 2 Original Code and Data Availability

The original replication package contains complete cleaning and analysis code as well as complete analysis data. However, the raw data is not complete. Specifically, three data files are missing for varying reasons:

- `Data/raw/upworthy-archive-confirmatory-packages-03.12.2020.csv` can be retrieved from Matias et al. (2021)'s [osf repository](#) which is indicated in the article.
- `Data/LIWC2015/liwc2015_dict.RData` is a proprietary dictionary that *was* distributed by [Pennebaker Conglomerates](#). Due to fairly widespread misuse (e.g., people redistributing and reselling copies of the LIWC dictionary, people building commercial products on the LIWC dictionary, etc.) Pennebaker Conglomerates is no longer distributing the LIWC dictionary files. However, in special cases, they can send a copy of the 2015 English dictionary file after agreeing to usage conditions (that prevent us from sharing the file).
- `Data/TopicModeling/upworthy_confirmatory_topics_7.csv` is missing by mistake. The authors provided that file upon request.

Robertson et al. (2023) conducted a user study to validate their sentiment analyses. Even though the user study is not directly related to the numerical reproduction of the two central

claims, it can be considered as the central claims’ backbone as it is intended to establish the study’s validity. We therefore sought to reproduce the user study and found an incomplete replication package where a clear documentation as well as some information to reproduce the validation was missing.<sup>2</sup> However, the authors provided the required information upon request.

Table 1: Availability of data (required to reproduce the two central claims)

	Fully	Partial	No
Raw data provided (or indicated)		x	
Cleaning code provided	x		
Analysis data provided	x		
Analysis code provided	x		

### 3 Blind Computational Reproduction

We successfully computationally reproduced the two central claims of Robertson et al. (2023) (in R) using the same data they used (see Matias et al. 2021)—but blindly. Accordingly, we ignored the authors’ code and followed the procedure they describe in their methods section (p. 818-819).

Table 2: Summary of blind computational reproducibility results

	Identical Numerical Results	Minor Differences	Major Differences
Reproducible from raw data		x	
Reproducible from analysis data	x		

Even though we find that positive and negative language in news headlines are both important determinants of CTRs, confirming the authors’ central claims, we also identified small differences that are most pronounced for the coefficient of negative words. Considering both the random effects model as well as the random effects model with random slopes reported in Supplementary Table 3 as well as Table 3 (p. 815), we find the corresponding estimate,  $\beta_2$ , to be considerably smaller than reported in the article: 0.009 instead of 0.015 (random effects model reported in Supplementary Table 3) and 0.006 instead of 0.018 (random effects model with random slopes reported in Table 3). However, the magnitude of these discrepancies decreases for negativity if we do look at the authors’ code and implement a pre-processing step during

<sup>2</sup> See [02\\_LIWC\\_validation.qmd](#) or [02\\_LIWC\\_validation.html](#) for details.

the text-mining called *stemming*. Stemming is the process of reducing words to their base or root form by removing prefixes and suffixes, in order to group related words together for more effective text analysis. For example, the words “connect,” “connecting,” and “connection” can all be reduced to the same stem, “connect” through the stemming process. This stemming process is reasonable but was not explicitly reported. For this reason, we report the results with and without stemming in Table 3 where all coefficients are statistically significant (with p-values  $\leq 0.005$ ). For more information, see also the [blind reproduction report](#) in our osf repository.

Table 3: Comparison of the blind reproduction’s key estimators by model with and without stemming procedure

	log(Odds Ratio)					
	Original		Click-through rate (CTR)		Blind with stemming	
	(1)	(2)	(3)	(4)	(5)	(6)
Positivity	−0.008*** (0.001)	−0.017*** (0.003)	−0.009*** (0.001)	−0.017*** (0.002)	−0.007*** (0.001)	−0.015*** (0.003)
Negativity	0.015*** (0.001)	0.018*** (0.003)	0.009*** (0.001)	0.006** (0.002)	0.014*** (0.001)	0.015*** (0.003)
Include Random Slopes	No	Yes	No	Yes	No	Yes
Observations	53,699	53,699	53,755	53,755	53,755	53,755
Akaike Inf. Crit.	532,321.800	498,467.600	533,115.800	505,071.100	532,983.200	498,978.500

Note:

\*p<0.05; \*\*p<0.01; \*\*\*p<0.001

Standard errors in parentheses.

As in Robertson et al. (2023), all models control for a headline’s word count, complexity and relative age.

In addition, we would describe the dimensions of the data differently compared to the authors’ formulations on page 814: while the authors report to analyze 53,699 *different headlines* that correspond to 12,448 A/B-tests, we could (naively) reproduce similar but not identical dimensions. Our data comprises 12,473 A/B-tests and 53,755 *rows*.<sup>3</sup> However, to our understanding, the number of rows does not represent the number of *different* headlines. Instead, we count 47,399 unique headlines (i.e., 13.29% fewer headlines than 53,699) and 5,126 headlines that are associated to multiple (up to 8) A/B-tests (i.e., `clickability_test_ids`).<sup>4</sup> Matias et al. (2021), who collaborated with *Upworthy*’s developers and editors to archive the data, informed us that this is likely driven by editors who tested headlines multiple times and that human error might have been another source of these “duplicates”.

<sup>3</sup> After inspecting the authors’ code, we conclude that these differences stem from the fact that they first processed the raw data before they filtered out headlines (see p. 818). We filtered the data first before we processed it.

<sup>4</sup> In addition, we find the numbers reported in the abstract not ideal because the authors describe the data they had access to, but not the (filtered) data they analyzed eventually. Specifically, they write, that they “conducted [their] analyses using a series of randomized controlled trials (N = 22,743)” even though they analyzed (only) about 12,000.



We also tried to blindly reproduce the proposed validation of the sentiment ratings obtained by the LIWC-dictionary (p. 819). Although this does not directly concern the calculations of the main analysis, it concerns the validity of the approach chosen for the main analysis, namely using the LIWC-dictionary for sentiment analysis. Robertson et al. (2023) report on their validation on p. 35f in Supplementary Material H. We could not blindly reproduce the identical correlations between the human benchmark and the LIWC sentiments that were reported by the authors. However, we did also find moderate and significant correlations that differed only little from the reported correlations, casting no doubt on the statements made by the authors on the validity of their approach.<sup>5 6</sup>

## 4 Unblind Computational Reproduction

We successfully computationally reproduced the two central claims with identical numerical results using their codes as well as their analysis data and raw data with minimal effort.

Table 4: Summary of unblind computational reproducibility results

	Identical Numerical Results	Minor Differences	Major Differences
Reproducible from raw data	x		
Reproducible from analysis data	x		

If we devote our attention to results other than the central claims, however, we find discrepancies that are two-fold. First, there are inconsistencies in the authors’ reporting. Robertson et al. (2023, 814) write, for instance: *“There are considerable differences between positive and negative language in news headlines (Fig. 1b). We find that positive words are more prevalent than negative words (Kolmogorov-Smirnov (KS) test:  $D = 0.574$ ,  $P < 0.001$ , two-tailed).”* In the header of that exact same Figure 1b, they report different statistics (i.e., *“KS test:  $D = 0.084$ ,  $P \approx 0$ ”*), though.

The second kind of discrepancies are differences between results we obtain by running their code using their (raw and analysis) data and the results the authors report in the article. For instance the authors write (on page 814) *“Overall, 2.83% of all words in news headlines are categorized as positive words, whereas 2.62% of all words are categorized as negative words.”* which is a statement we cannot reproduce with their code.<sup>7</sup>

<sup>5</sup> When consulting the code, we were able to replicate the results, but we could not reconstruct how the respective sentiment variable that was used by the authors was created.

<sup>6</sup> Following a literate programming approach (Knuth 1984) we provide the documentation of our analyses in our osf repository (see [01\\_main\\_claims.qmd](#) or [01\\_main\\_claims.html](#)).

<sup>7</sup> Interestingly, the authors also report values of 3.7% and 2.8% in Supplementary Table 1, which can be reproduced with their code.

Trying to identify the root of these differences, we experienced problems recreating the authors' software environments. We acknowledge that the R-software as well as some package versions are specified on page 819 but two issues remain. First, we miss an exhaustive list, specifying the version of tidyverse, texreg, or xtable, for instance. Second, some of the specified package versions are incompatible with other required packages. For example, the package `quanteda.dictionaries` requires `quanteda` 3.0.0 and is not compatible with version 2.0.1, as specified. Similarly, we could not use the version 1.1.23 of the `lme4` package, likely due dependency issues with other packages when using R 4.0.2. A package version control system such as `packrat`, `renv` or, nowadays, `groundhog` would improve the replication package. Second, a documentation describing the files or an order in which the codes shall be processed would be of great benefit, too. For instance, there are two files with the name `regression_df.csv` containing very different data and variables. Furthermore, across different data sets they use identical names for variables that are not identical (i.e., contain differing content).<sup>8</sup> We could imagine that, in part, an inadequate separation of calculations, code and files between stage 1 and stage 2 of the registered report, as well as some inconsistent variable naming between the two stages, might have contributed to the reported confusions.

To increase comprehension and reproducibility, the code would have benefited from a reorganization, for example by a more intuitive and unique naming of files and variables, clearly separating main and secondary analyses in the code, and ensuring that tables and figures are referenced appropriately in the comments or text annotations of the scripts.

Importantly, despite some organizational challenges in the code and documentation that may stem from the two-stage registered report process, we were able to successfully reproduce the central claims, validating the main claims of the paper.

## 5 Robustness Checks

For their lexical sentiment analyses, Robertson et al. (2023) apply dictionaries to detect negative and positive words within the news headlines. They primarily rely on LIWC2015 (p. 818) for their main analysis, but also use the NRC and SentimentStrength dictionaries for robustness checks (p. 815). While dictionaries are a common and transparent method for text analysis, they also come with several shortcomings such as their difficulty in capturing the context of words or limitations in handling synonyms and polysemous words. Consequently, several studies have demonstrated that more recent and complex methods like machine learning and deep learning outperform dictionary based text analysis in various contexts, including sentiment analysis (Hartmann et al. 2023, 2019; Barberá et al. 2021; Nelson et al. 2021; Wouter van Atteveldt and Boukes 2021).

---

<sup>8</sup> Here, we are referring to `UserStudy/regression_df.csv` and `validation_sample.csv` that both contain the variables `liwc_negemo`, `liwc_posemo`, and `liwc_sentiment` which hold differing values for identical headlines.

Due to their architecture, large language models (such as BERT, see Devlin et al. 2019 and ChatGPT or GPT-4o), can capture the context in which words are used, allowing them to interpret the sentiment of phrases and sentences that might be ambiguous or have different meanings based on context (see, e.g., Hussain et al. 2023). Rathje et al. (2024) as well as Hartmann et al. (2023, Table 1), report that state of the art language models outperform dictionary approaches by 20 percentage points on average and argue that they are well suited for this task that is characterized by two classes (positive vs. negative) and sentence-level analyses. Consequently, in this robustness reproduction, we apply two recent large language models and generate two new sets of independent variables, effectively taking a complementary perspective on the central claims by Robertson et al. (2023). Specifically, we applied DistilBERT (Yuan 2023) and OpenAI’s GPT-4o to predict the positivity and negativity of each headline. These robustness checks were pre-registered in advance. The corresponding pre-re-analysis plan can be found in [our osf repository](#).

First, we chose DistilBERT because it is open-source and both smaller and faster than the standard BERT model while preserving over 95% of BERT’s performances. As such, we use algorithms, tools, and workflows that were available during the original study period and adhere to FAIR principles (Wilkinson et al. 2016) to increase transparency, reproducibility, and reusability. In addition, Yuan (2023)’s model perfectly fits our purpose as it conveniently predicts scores for both positivity and negativity.

Second, we use GPT-4o, the most recent and advanced model family by OpenAI at the time of writing to determine the sentiment of each headline. Specifically, we employ `gpt-4o-mini-2024-07-18`, a cost-efficient version of GPT-4o via OpenAI’s python API. GPT-4o (including GPT-4o-mini) is an advancement of the previous GPT-4 model family, which already has been shown to be capable of outperforming human annotators (Törnberg 2023).

When Robertson et al. registered their analyses in 2020, many of these modern tools were either not available or were in earlier stages of development. GPT-4 was released more than two years later, and much of the evidence demonstrating that large language models can effectively analyze sentiment has started to emerge at that time. The authors’ choice of lexical methods therefore reflected a *ne* established and *accessible* practice, particularly for ensuring reproducibility and transparency in scientific research. Our use of both contemporary (DistilBERT) and new (GPT-4) semantical methods thus complements rather than criticizes their methodological choices, demonstrating how rapidly evolving computational tools can provide additional perspectives.

## 5.1 Methods

Our goal is to compare the results we obtain from using large language models to the original lexical results. Hence, we use DistilBERT’s and GPT-4o’s sentiment scores and follow the

authors’ approach in analyzing the data closely: We estimate the effect of semantical positivity and negativity on the CTR ( $\theta_{ij}$ ) and capture between-experiment heterogeneity through a multi-level structure. Like Robertson et al. (2023), we start by specifying the random intercept, random slopes model in Equation 1. We also control for other characteristics across headline variations, namely length, text complexity and the relative age of a headline. The first regression model is then given by

$$(1) \quad \text{logit}(\theta_{ij}) = \alpha + \alpha_i + \beta_1 \text{Positive}_{ij} + \beta_2 \text{Negative}_{ij} + \mathbf{X}'_{ij} \gamma$$

where  $j$  denotes a headline variation in an RCT  $i$  ( $i = 1, \dots, N$ ) and where “Positive” and “Negative” are generic terms for the three pairs of sentiment scores described above (i.e., LIWC, DistilBERT and GPT-4o). The vector  $\gamma$  represents the effects of length, text complexity and the relative age of a headline, which are captured in the transposed vector  $\mathbf{X}'_{ij}$ . This mirrors the random intercept specification with fixed slopes described on page 819 in Robertson et al. (2023).

Second, we estimate the authors’ preferred specification in Equation 2, that is, a random intercept model with random slopes for  $\beta_1$  and  $\beta_2$  to allow the receptivity to language to vary across news articles (for example, if the receptivity of negative language differs between political and entertainment news).

$$(2) \quad \text{logit}(\theta_{ij}) = \alpha + \alpha_i + (\beta_1 + \beta_{1i}) \text{Positive}_{ij} + (\beta_2 + \beta_{2i}) \text{Negative}_{ij} + \mathbf{X}'_{ij} \gamma$$

Importantly, the dictionary approach applied by Robertson et al. (2023) calculates sentiment scores by determining the proportion of negative or positive words in a headline. Accordingly, a higher proportion of negative (positive) words is interpreted as a higher level of negativity (positivity). As mentioned above, the applications of DistilBERT and GPT-4o go beyond word counts and predict sentiments based on the entire headline without any unit of measurement (such as the number of words in a sentence).

## 5.2 Validation

To assess whether the predictions by the two large language models are valid (Reiss 2023), we build on the authors’ user study in which human judges rated the sentiment of a random subset of headlines. Accordingly, we use each of the two models to predict the positivity and negativity of news headlines for the authors’ subset of 213 headlines. We then correlate the average human sentiment rating for the 213 headlines with the sentiment ratings obtained from DistilBERT and GPT-4o. The resulting Spearman’s rank correlation coefficients are reported in Table 5, where we compare the correlations between the sentiment ratings by the human judges and the various sentiment ratings (i.e., LIWC, DistilBERT, and GPT-4o). All of the coefficients are significant (at  $p < 0.001$ ) and both large language model approaches outperform

the lexical LIWC approach—potentially because large language models can go beyond simple word counts and analyze context, tone, intent, and the relationships between words to capture nuances, sarcasm, and idiomatic expressions, similar to how humans interpret meaning.<sup>9</sup>

Table 5: Correlations between the human sentiment ratings and model predictions

	LIWC	DistilBERT	GPT-4o
Positivity	0.1968	0.3098	0.7906
Negativity	-0.2023	-0.2923	-0.7771

### 5.3 Results

We find support for the authors’ main claim that negativity drives online news consumption. For the authors’ secondary claim, positivity, we find inconsistent evidence, complementing the author’s findings.

Table 6: Summary of robustness checks

	Support for Claim	Inconsistent Results
Main claim: Negativity	x	
Secondary claim: Positivity		x

The results of our analyses, both for the fixed and random slopes model, are presented in Figure 1. The estimates resulting from the LIWC-dictionary approach by Robertson et al. (2023) serves as the benchmark.<sup>10</sup> Focusing on their main claim, *negativity drives online news consumption*, the left panel of Figure 1 illustrates the effect of negative sentiment on the CTR across the different sentiment analysis paradigms.

<sup>9</sup> We provide the corresponding scripts in our osf repository (see [validation\\_DistilBERT.qmd](#) and [validation\\_gpt4o.qmd](#)).

<sup>10</sup> When we refer to LIWC-based estimates in this section, we always refer to the results reported in the original study. To compute these results, we used their data, which contains fewer observations than the data we used to predict the large language models’ sentiment score (see Section 3 for more details). In Appendix A (see Table 8 & Table 9), we show that the difference in the number of observations does not affect the results.

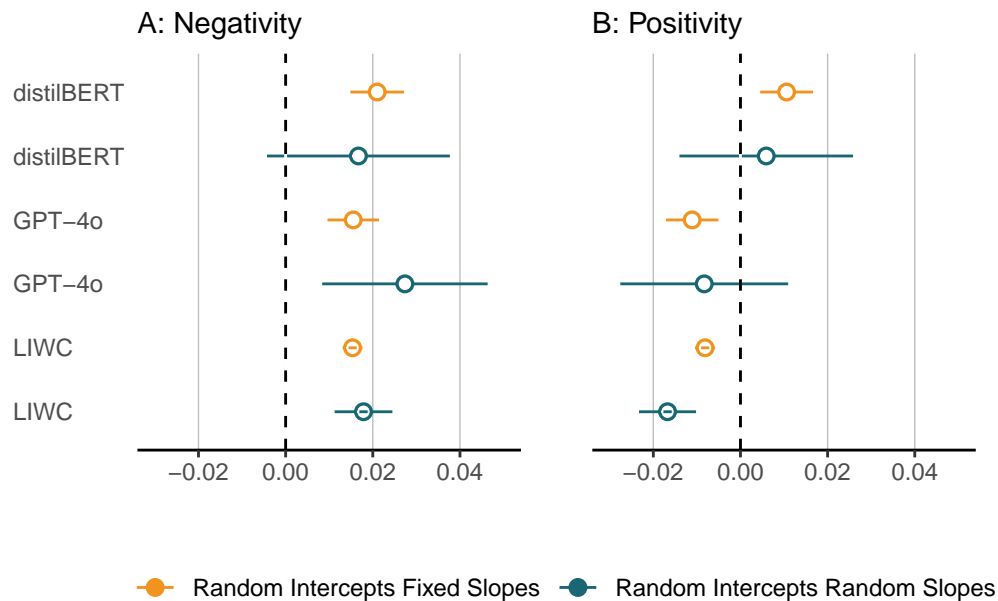


Figure 1: Forest plot of effect estimates negativity (left) & positivity (right) with 99% confidence intervals across different sentiment analysis paradigms where LIWC represents the authors’ original estimates and serves as a benchmark.

In line with the LIWC benchmark, the fixed slopes model (orange) yields statistically significant positive effects of negativity for both the DistilBERT and GPT-4o approaches.<sup>11</sup> When we consider the random slopes model (blue), the results are more nuanced. The GPT-4o approach continues to show statistically significant positive effects, consistent with its fixed slopes results. However, the DistilBERT random slopes estimator does not yield statistically significant results, although we interpret it as suggestive evidence ( $p \approx 0.040$ )<sup>12</sup> which is consistent with the other estimations. See also Table 7 for the full regression results.

The consistency of these results across the two different sentiment analysis paradigms and model specifications strengthens the case for a positive relationship between negative sentiment and CTR in online news articles. We therefore conclude that the negativity effect (i.e., the authors’ main claim) is robust across different sentiment analysis paradigms (i.e., a lexical and a semantic approach).

We now shift our attention to the authors’ secondary claim, *positivity reduces online news consumption*, for which the right panel of Figure 1 and the first row of Table 7 show inconsistent results: considering the fixed slopes model, both the DistilBERT and GPT-4o approach yield statistically significant—yet contradicting—estimates. Whereas we find a positive effect of positivity with DistilBERT scores (contradicting the authors’ results on lexical positivity)

<sup>11</sup> We excluded seven observations from the GPT-4o-based models because GPT-4o did not provide meaningful sentiment scores for these seven headlines. For more details, please refer to Appendix B3 with Table 10.

<sup>12</sup> See Appendix B2.

Table 7: Results of the regression model with random intercepts with and without random slopes for the varying sentiment scores

	log(Odds Ratio)					
	Click-through rate (CTR)					
	Random Intercept	Fixed Slopes		Random Intercept	Random Slopes	
	(1)	(2)	(3)	(4)	(5)	(6)
Positivity	−0.008*** (0.001)	0.011*** (0.002)	−0.011*** (0.002)	−0.017*** (0.003)	0.006 (0.008)	−0.008 (0.007)
Negativity	0.015*** (0.001)	0.021*** (0.002)	0.016*** (0.002)	0.018*** (0.003)	0.017* (0.008)	0.027*** (0.007)
Length	0.013*** (0.0003)	0.040*** (0.001)	0.039*** (0.001)	0.014*** (0.0004)	0.038*** (0.001)	0.036*** (0.001)
Complexity	−0.001*** (0.0002)	−0.003*** (0.001)	−0.003** (0.001)	−0.001 (0.0003)	−0.003** (0.001)	−0.004*** (0.001)
Platform Age	−0.002*** (0.00003)	−0.311*** (0.005)	−0.309*** (0.005)	−0.002*** (0.00003)	−0.312*** (0.006)	−0.307*** (0.006)
Constant	−3.928*** (0.014)	−4.474*** (0.006)	−4.475*** (0.006)	−3.961*** (0.015)	−4.498*** (0.006)	−4.493*** (0.006)
Sentiment	LIWC	DistilBERT	GPT-4o	LIWC	DistilBERT	GPT-4o
Observations	53,699	53,755	53,748	53,699	53,755	53,748
Akaike Inf. Crit.	532,321.800	533,175.000	532,707.300	498,467.600	493,216.900	493,786.400

Note: \*p<0.05; \*\*p<0.01; \*\*\*p<0.001  
Standard errors in parentheses.

the effect we estimate with GPT-4o scores is negative and in line with the authors’ findings. Focusing on the random slopes model, none of our approaches results is statistically significant or suggestive evidence in favor of a positivity bias. In their [registered report stage 1 protocol](#) (p. 19), the authors give precedence to the random slopes model in cases of conflicting results between different models.<sup>13</sup> If we follow this guideline and focus on random slopes models, we still observe a considerable variation in magnitude, significance levels and signs. Because none of our random slope estimators is statistically distinguishable from zero, we conclude that we do *not* find a statistically significant effect of semantic positivity.

Taken together, we find the effect of positivity is neither robust within semantic nor across lexical and semantic sentiment analysis paradigms. The validation study we report in Section 5.2 shows that GPT4-o scores exhibit by far the highest correlation with the average human rater’s judgement (i.e.,  $r_s^{GPT-4o} > 0.75$ ). LIWC’s and DistilBERT’s moderate correlations  $r_s^{dBERT} \approx |0.3|$  compared to  $r_s^{LIWC} \approx |0.2|$ ) can be interpreted as a limitation in validity that could affect the reliability of the analysis, particularly when dealing with small or borderline effects. However, we also want to stress that the validation study was rather small ( $N = 213$ ) and the agreement between the raters was moderate.

## 6 Conclusion

The present paper contributes to recent calls for more replication studies in communication science (Bowman 2024; Breuer and Haim 2024; Dienlin et al. 2021; Freiling et al. 2021) and tests the reproducibility and robustness of the central claims in Robertson et al. (2023), who investigate the impact of positive and negative language on online news consumption. The original paper leverages a large data set comprising more than 12,000 A/B-tests and finds that negativity drives online news consumption whereas positivity reduces it.

Using the same data, we were able to computationally reproduce their claims using two different approaches. First, we only followed the procedure described in their methods section (without looking at their code) to obtain very similar results. Second, we used the authors’ code to obtain results that are numerically identical to those reported in their article. There were some issues regarding data availability and documentation, but thanks to the swift responses from the original authors, these were easily resolved.

In the second step, we evaluated the robustness of the central claims made by Robertson et al. (2023) using two alternative approaches for assessing the sentiment of news headlines. Instead of employing the authors’ lexical method to count the frequency of positive and negative words in a news headline, we applied a semantic approach using two large language models, DistilBERT and GPT-4, to re-compute the positive and negative sentiment.

---

<sup>13</sup> In Table 1 Robertson et al. (2023, 813) the authors deviate from their pre-registered statement: “We consider evidence to be conclusive only in cases where both model fits to the data agree in their qualitative conclusions about the effect of negative words.” (The authors do not mention “positive words” in that context).



The different results between lexical and semantic analyses provide insights into how we measure and understand emotional content in news headlines. The finding that lexical and semantic negativity both showed similar effects on click-through rates supports the original study's main finding about negativity driving news consumption. This convergence across methodologies suggests that a general notion of a negativity bias in news consumption is robust.

In contrast, the positivity bias does not generalize as well. The lexical analysis suggests that positive words reduce engagement, whereas our semantic analyses show no consistent effects, with variations in magnitude, significance, and even direction. This divergence may stem from theoretical differences: positivity might not provoke as strong a reaction as negativity. The authors themselves argue that *“negative information may be more ‘sticky’ in our brains; people weigh negative information more heavily than positive information”* (p. 812). From a methodological perspective, the relatively strong validation of the semantic approach suggests it may better capture contextual nuances in sentiment. However the validation study does not answer the question of how people evaluate emotions in headlines “in the wild”. While we found preliminary evidence suggesting that context is important, it is also plausible that people rely on heuristics, reacting strongly to specific words without fully evaluating their contextual meaning.

Using the same data, Banerjee and Urminsky (2024) also note that while some factors can systematically improve engagement, their effects often defy predictions from both practitioners and academics. This unpredictability underscores the need for further research employing diverse methods, designs, and contexts (see Berger, Moe, and Schweidel 2023). Building on this, we hope that our perspective on what drives online news consumption inspires further exploration of how language and linguistic cues influence engagement and attention.

## Acknowledgements

We would like to thank Abel Brodeur for his guidance on the replication process and the authors of the original study for their swift responses and helpful comments. Their willingness to provide additional information greatly facilitated our replication efforts. Special thanks goes to J. Nathan Matias for his valuable comments.

## Code & Data Availability

Code and data (with exception of the LIWC dictionary) that support the findings of our reproductions and robustness checks can be found in our osf repository <https://osf.io/bfhdw/> alongside the pre-re-analysis plan that builds the foundation of this report.

## Competing Interests

The authors declare no competing interests.

## References

- Banerjee, Akshina, and Oleg Urminsky. 2024. “The Language That Drives Engagement: A Systematic Large-Scale Analysis of Headline Experiments.” *Marketing Science* Articles in Advance. <https://doi.org/10.1287/mksc.2021.0018>.
- Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. “Automated Text Classification of News Articles: A Practical Guide.” *Political Analysis* 29 (1): 19–42. <https://doi.org/10.1017/pan.2020.8>.
- Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, et al. 2018. “Redefine Statistical Significance.” *Nature Human Behaviour* 2 (1): 6–10. <https://doi.org/10.1038/s41562-017-0189-z>.
- Berger, Jonah, Wendy W. Moe, and David A. Schweidel. 2023. “What Holds Attention? Linguistic Drivers of Engagement.” *Journal of Marketing* 87 (5): 793–809. <https://doi.org/10.1177/00222429231152880>.
- Bowman, Nicholas. 2024. “On the Continued Need for Replication in Media and Communication Research.” *Media and Communication* 12: 1–5. <https://doi.org/10.17645/mac.7935>.
- Breuer, Johannes, and Mario Haim. 2024. “Are We Replicating yet? Reproduction and Replication in Communication Research.” *Media and Communication* 12: 1–7. <https://doi.org/10.17645/mac.8382>.
- Brodeur, Abel, Anna Dreber, Fernando Hoces de la Guardia, and Edward Miguel. 2024. “Reproduction and Replication at Scale.” *Nature Human Behaviour* 8: 2–3. <https://doi.org/10.1038/s41562-023-01807-2>.
- Brodeur, Abel, Derek Mikola, Nikolai Cook, Thomas Brailey, Ryan Briggs, Alexandra de Gendre, Yannick Dupraz, et al. 2024. “Mass Reproducibility and Replicability: A New Hope.” IZA Discussion Paper 16912. IZA - Institute of Labor Economics. <https://doi.org/10.1257/pol.20200092>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” <https://arxiv.org/abs/1810.04805>.
- Dienlin, Tobias, Niklas Johannes, Nicholas David Bowman, Philipp K Masur, Sven Engesser, Anna Sophie Kümpel, Josephine Lukito, et al. 2021. “An Agenda for Open Science in Communication.” *Journal of Communication* 71 (1): 1–26. <https://doi.org/10.1093/joc/jqz052>.
- Freiling, Isabelle, Nicole M Krause, Dietram A Scheufele, and Kaiping Chen. 2021. “The Science of Open (Communication) Science: Toward an Evidence-Driven Understanding of Quality Criteria in Communication Research.” *Journal of Communication* 71 (5): 686–714. <https://doi.org/10.1093/joc/jqab032>.

- Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp. 2023. “More Than a Feeling: Accuracy and Application of Sentiment Analysis.” *International Journal of Research in Marketing* 40 (1): 75–87. <https://doi.org/10.1016/j.ijresmar.2022.05.005>.
- Hartmann, Jochen, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. “Comparing Automated Text Classification Methods.” *International Journal of Research in Marketing* 36 (1): 20–38. <https://doi.org/10.1016/j.ijresmar.2018.09.009>.
- Hussain, Zak, Marcel Binz, Rui Mata, and Dirk U Wulff. 2023. “A Tutorial on Open-Source Large Language Models for Behavioral Science.” PsyArXiv. <https://doi.org/10.31234/osf.io/f7stn>.
- Knuth, Donald E. 1984. “Literate Programming.” *The Computer Journal* 27 (2): 97–111. <https://doi.org/10.1093/comjnl/27.2.97>.
- Lakens, Daniel, Federico G. Adolphi, Casper J. Albers, Farid Anvari, Matthew A. J. Apps, Shlomo E. Argamon, Thom Baguley, et al. 2018. “Justify Your Alpha.” *Nature Human Behaviour* 2 (3): 168–71. <https://doi.org/10.1038/s41562-018-0311-x>.
- Matias, J. Nathan, Kevin Munger, Marianne Aubin Le Quere, and Charles Ebersole. 2021. “The Upworthy Research Archive, a Time Series of 32,487 Experiments in u.s. Media.” *Scientific Data* 8: 195. <https://doi.org/10.1038/s41597-021-00934-7>.
- Nelson, Laura K., Derek Burk, Marcel Knudsen, and Leslie McCall. 2021. “The Future of Coding: A Comparison of Hand-Coding and Three Types of Computer-Assisted Text Analysis Methods.” *Sociological Methods & Research* 50 (1): 202–37. <https://doi.org/10.1177/0049124118769114>.
- Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjeh, Claire E. Robertson, and Jay J. Van Bavel. 2024. “GPT Is an Effective Tool for Multilingual Psychological Text Analysis.” *Proceedings of the National Academy of Sciences* 121 (34): e2308950121. <https://doi.org/10.1073/pnas.2308950121>.
- Reiss, Michael V. 2023. “Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark.” <https://arxiv.org/abs/2304.11085>.
- Robertson, Claire E., Nicolas Pröllochs, Kaoru Schwarzenegger, Philip Pärnamets, Jay J. Van Bavel, and Stefan Feuerriegel. 2023. “Negativity Drives Online News Consumption.” *Nature Human Behaviour* 7: 812–22. <https://doi.org/10.1038/s41562-023-01538-4>.
- Törnberg, Petter. 2023. “ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning.” <https://arxiv.org/abs/2304.06588>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.
- Wouter van Atteveldt, Mariken A. C. G. van der Velden, and Mark Boukes. 2021. “The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms.” *Communication Methods and Measures* 15 (2): 121–40. <https://doi.org/10.1080/19312458.2020.1869198>.
- Yuan, Lik Xun. 2023. “Distilbert-Base-Multilingual-Cased-Sentiments-Student (Revision 2e33845).” Hugging Face. <https://doi.org/10.57967/hf/1422>.

## Appendix

### A: Additional Analyses

The results associated to the LIWC benchmark reported in Table 7 (see Section 5.3) are based on the authors’ original data. The remaining columns of that table (columns number 2, 3 and 5, 6) report sentiment scores based on data recomputed in the blind reproduction. Hence, as described above, the number of observations are not equal across regression models. To account for effects that might simply occur due to the small difference in  $n$ , in this section, we mirror Table 7 but hold the number of observations constant across all models.

#### A1 Our set of 53,755 rows with 47,399 unique headlines

Table 8 displays the results when using the number of observations we retrieved in the blind reproduction (without stemming, see Section 3) for all regression models. Importantly, as the data set is larger, there were some observations for which we could not use a provided LIWC-score as these observations were not included in the original data. Hence, we computed the LIWC-scores for all observations, which are different compared to the original data for the exact same observations.

Table 8: Results of the regression model without and with random slopes for the varying sentiment scores based on the headlines we retrieved from our blind reproduction

	log(Odds Ratio)					
	Click-through rate (CTR)					
	Random Intercept	Fixed Slopes		Random Intercept	Random Slopes	
	(1)	(2)	(3)	(4)	(5)	(6)
Positivity	−0.009*** (0.001)	0.011*** (0.002)	−0.011*** (0.002)	−0.017*** (0.002)	0.006 (0.008)	−0.008 (0.007)
Negativity	0.009*** (0.001)	0.021*** (0.002)	0.016*** (0.002)	0.006** (0.002)	0.017* (0.008)	0.027*** (0.007)
Sentiment	LIWC	DistilBERT	GPT-4o	LIWC	DistilBERT	GPT-4o
Observations	53,755	53,755	53,748	53,755	53,755	53,748
Akaike Inf. Crit.	533,115.800	533,175.000	532,707.300	505,071.100	493,216.900	493,786.400

*Note:* \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$   
Standard errors in parentheses.  
As in Robertson et al. (2023), all models control for a headline’s word count, complexity and relative age.

## A2 The original set of 53,699 rows with 47,301 unique headlines

Additionally, Table 9 shows the results when using the number of observations as in the original study. Again, these are the LIWC-scores computed by us in the blind computational reproduction and they are different compared to the original LIWC-scores for the identical observations, leading to slightly different results for the LIWC-based regressions (compare Table 7). However, more importantly, both the LIWC-based results and the results based on DistilBERT and GPT-4o are robust against the small changes in  $n$ . Moreover, the differences to the LIWC-based results observed in Table 7 remain for all observed scenarios.

Table 9: Results of the regression model without and with random intercepts for the varying sentiment scores based on the headlines we retrieved from our blind reproduction

	log(Odds Ratio)					
	Click-through rate (CTR)					
	Random Intercept	Fixed Slopes		Random Intercept	Random Slopes	
	(1)	(2)	(3)	(4)	(5)	(6)
Positivity	−0.009*** (0.001)	0.011*** (0.002)	−0.011*** (0.002)	−0.017*** (0.002)	0.006 (0.008)	−0.007 (0.007)
Negativity	0.009*** (0.001)	0.021*** (0.002)	0.016*** (0.002)	0.007** (0.002)	0.017* (0.008)	0.028*** (0.007)
Sentiment	LIWC	DistilBERT	GPT-4o	LIWC	DistilBERT	GPT-4o
Observations	53,699	53,699	53,692	53,699	53,699	53,692
Akaike Inf. Crit.	532,502.800	532,562.100	532,093.500	504,457.200	492,603.400	493,178.700

Note:

\* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$

Standard errors in parentheses.

As in Robertson et al. (2023), all models control for a headline's word count, complexity and relative age.

## B: Deviations from the pre-re-analysis plan

### B1 Input Data

In our pre-re-analysis plan, we specified to run our sentiment analyses on raw *unique* headlines in the data we processed during the blind reproduction. We slightly deviate from this plan and compute the sentiment scores for *all* of the raw headlines in the data we processed during the blind reproduction.

```
origi <- fread("../00_original_files/scripts/Data/regression_df.csv")

blind <- fread("../01_blind_computational_reproduction/data/blind_regression_df.csv")
```

```

setnames(x = blind,
        old = c("V1", "liwc_posemo", "liwc_negemo"),
        new = c("unique_identifier", "liwc_posemo_count", "liwc_negemo_count"))
setnames(x = blind,
        old = c("positive", "negative"),
        new = c("liwc_posemo", "liwc_negemo"))

fwrite(x = blind,
       file = "../03_robustness_checks/data/input/robustness_sample_input.csv",
       sep = ";")

```

This procedure is slightly more costly, as we compute scores in 53,755 rows (with duplicated headlines) instead of just 47,399 but removes any room for error as we can match the headlines unambiguously afterwards.

## B2 Threshold for “Statistical Significance”

Robertson et al. (2023) did not explicitly define a threshold for statistical significance (Lakens et al. 2018; Benjamin et al. 2018). We did not do so either up until we computed the DistilBERT scores (and before we computed the GPT-4o scores). For consistency with the reporting in Table 3 of the original study, we will consider results with  $p < 0.01$  as “statistically significant” and those with  $p < 0.05$  as “suggestive evidence”. All tests conducted will be two-sided.

## B3 Exclusion of seven cases for the GPT-4o-based analyses

In our pre-re-analysis plan, we did not specify any exclusion criteria. In our analysis, however, we excluded seven headlines from the GPT-4o-based regression models because GPT-4o did not provide meaningful and sensible sentiment scores for these headlines:

In five instances, the headlines were variations of: *“POP QUIZ: If BP Made \$36 Billion In 2010, And Their Tax Rate Is 35%, How Big Was Their Tax Bill?”*<sup>14</sup> Instead of returning sentiment scores, GPT-4o interpreted the quiz as a prompt, solved it, and responded with “12.6”. In two other instances, the headlines read: *“Can You Read The Words In This Image? If Your Answer Is No, Find Out Why.”*<sup>15</sup> Again, instead of returning a sentiment score, GPT-4o replied: *“I’m sorry, but I cannot read images. However, if you provide me with the text of the headline, I can help,”*.

<sup>14</sup> See the following `unique_identifier` in `robustness_sample_predicted_distilBERT.csv`.

<sup>15</sup> See the following `unique_identifier` 22360, 22363, 41673, 41678, 41679, 41680, 41681 in `robustness_sample_predicted_distilBERT.csv`.

No other responses from GPT-4o fell outside the required range of 0 to 1. Unfortunately, we did not anticipate such (rare but obvious) misinterpretations. Therefore, we also present the regression results for GPT-4o, adhering strictly to the pre-registration, which includes the five instances where GPT-4o responded with “12.6” but excludes the two non-numeric responses. Consequently, the sample size is  $N = 53,753$ .

The full regression results can be found in Table 10. Although the coefficients differ somewhat compared to the models with those seven cases excluded, the overall conclusions remain unchanged.

Table 10: Results of the regression model with random intercepts with and without random slopes for the varying sentiment scores

	log(Odds Ratio)					
	Click-through rate (CTR)					
	Random Slopes	Fixed Intercept		Random Slopes	Random Intercept	
	(1)	(2)	(3)	(4)	(5)	(6)
Positivity	−0.008*** (0.001)	0.011*** (0.002)	−0.011*** (0.002)	−0.017*** (0.003)	0.006 (0.008)	−0.008 (0.007)
Negativity	0.015*** (0.001)	0.021*** (0.002)	0.016*** (0.002)	0.018*** (0.003)	0.017* (0.008)	0.027*** (0.007)
Length	0.013*** (0.0003)	0.040*** (0.001)	0.039*** (0.001)	0.014*** (0.0004)	0.038*** (0.001)	0.036*** (0.001)
Complexity	−0.001*** (0.0002)	−0.003*** (0.001)	−0.003** (0.001)	−0.001 (0.0003)	−0.003** (0.001)	−0.004*** (0.001)
Platform Age	−0.002*** (0.00003)	−0.311*** (0.005)	−0.309*** (0.005)	−0.002*** (0.00003)	−0.312*** (0.006)	−0.307*** (0.006)
Constant	−3.928*** (0.014)	−4.474*** (0.006)	−4.475*** (0.006)	−3.961*** (0.015)	−4.498*** (0.006)	−4.493*** (0.006)
Sentiment	LIWC	DistilBERT	GPT-4o	LIWC	DistilBERT	GPT-4o
Observations	53,699	53,755	53,748	53,699	53,755	53,748
Akaike Inf. Crit.	532,321.800	533,175.000	532,707.300	498,467.600	493,216.900	493,786.400

*Note:* \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$   
Standard errors in parentheses.

## B4 Version Control

We modified the original pre-re-analysis plan after submitting it and after retrieving parts of the data. To avoid confusion and ambiguities, we removed these edits and recovered the original files. As this is not salient, we mention it explicitly and direct the interested reader to [OSF's version control feature](#).

**C: Session info**

```

R version 4.4.1 (2024-06-14)
Platform: x86_64-apple-darwin20
Running under: macOS Sonoma 14.4.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Europe/Berlin
tzcode source: internal

attached base packages:
[1] stats      graphics  grDevices  utils      datasets  methods   base

other attached packages:
[1] kableExtra_1.4.0    patchwork_1.2.0      ggplot2_3.5.1
[4] broom.mixed_0.2.9.5 stargazer_5.2.3      gt_0.11.0
[7] gtsummary_2.0.0     lme4_1.1-35.5        Matrix_1.7-0
[10] stringr_1.5.1       knitr_1.48           data.table_1.15.4
[13] magrittr_2.0.3

loaded via a namespace (and not attached):
[1] gtable_0.3.5      xfun_0.46           lattice_0.22-6      vctrs_0.6.5
[5] tools_4.4.1       generics_0.1.3      parallel_4.4.1     tibble_3.2.1
[9] fansi_1.0.6       pkgconfig_2.0.3     lifecycle_1.0.4    farver_2.1.2
[13] compiler_4.4.1    tinytex_0.52        munsell_0.5.1      codetools_0.2-20
[17] htmltools_0.5.8.1 yaml_2.3.10         pillar_1.9.0       furrr_0.3.1
[21] nloptr_2.1.1      tidyr_1.3.1         MASS_7.3-61        boot_1.3-30
[25] nlme_3.1-165      parallelly_1.38.0   tidyselect_1.2.1   digest_0.6.36
[29] stringi_1.8.4     future_1.34.0       dplyr_1.1.4        purrr_1.0.2
[33] listenv_0.9.1     labeling_0.4.3      forcats_1.0.0      splines_4.4.1
[37] fastmap_1.2.0     grid_4.4.1          colorspace_2.1-1   cli_3.6.3
[41] utf8_1.2.4        broom_1.0.6         withr_3.0.1        scales_1.3.0
[45] groundhog_3.2.1   backports_1.5.0     rmarkdown_2.27     globals_0.16.3
[49] evaluate_0.24.0   viridisLite_0.4.2   rlang_1.1.4        Rcpp_1.0.13
[53] glue_1.7.0        xml2_1.3.6          svglite_2.1.3      rstudioapi_0.16.0
[57] minqa_1.2.7       jsonlite_1.8.8      R6_2.5.1           systemfonts_1.1.0

```