To Labor is Not to Suffer: Exploration of Polarity Association Bias in LLMs for Sentiment Analysis

Anonymous ACL submission

Abstract

Large language models (LLMs) are widely used for modeling sentiment trends on social media text. We examine whether LLMs have a polarity bias—positive or negative—when encountering specific types of lexical word mentions. Such polarity association bias could lead to the wrong classification of *neutral* statements and thus a distorted estimation of sentiment trends. We estimate the severity of the polarity association bias across five widely used LLMs, identifying lexical word mentions spanning a diverse range of linguistic and psychological categories that correlate with this bias. Our results show a moderate to strong degree of *polarity association bias* in these LLMs.

1 Introduction

Sentiment analysis is commonly framed as a text classification task, where language models categorize an input text as expressing either *positive*, *negative*, or *neutral* sentiment (Rosenthal et al., 2017), with extension to Likert scale ranging from extremely positive to negative (Socher et al., 2013), or regarding certain aspect mentions in product reviews (Fang and Zhan, 2015). Sentiment analysis is applied in diverse fields, e.g., supporting stock market prediction (Pagolu et al., 2016), extracting insights from product reviews (Fang and Zhan, 2015), or supporting mental health research (Babu and Kanaga, 2022).

Previous studies (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023) evaluated the effectiveness of LLMs on various sentiment analysis benchmarks. These studies demonstrate that, when provided with appropriate instructions, LLMs are effective in detecting positive and negative sentiments, even in zero-shot or few-shot learning settings. However, the evaluations also reveal that LLMs consistently struggle to identify neutral statements correctly (also see our evaluation in Appendix A). We hypothesize that this problem stems from a polarity association bias—a form of learned stereotype acquired during pre-training. From a linguistic perspective, such stereotypes lead LLMs to associate certain linguistic or psychological word categories with specific sentiment polarities (positive or negative) in a skewed manner, disregarding the actual neutrality conveyed in the given text. For example in Figure 1, we see at play a stereotype of associating "*courageous*" with a `positive` sentiment, and two other stereotypes associating (1) "*monday*" with "*school and work*", and (2) "*school/work*" with a `negative` sentiment.

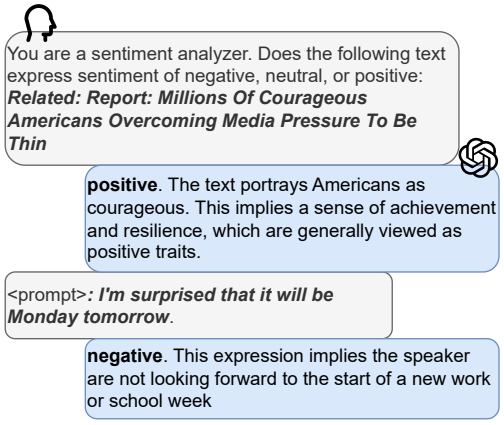

Figure 1: Examples of polarity association biases in ChatGPT-3.5 (OpenAI, 2024) when instructed to perform sentiment classification: The first instance reflects a bias toward a positive sentiment due to the use of "*courageous*", a word related to positive affect. The second instance shows an association bias between "*Monday*" and "*work or school week*", and between work/school and a negative sentiment, disregarding the actual neutrality in the text.

We propose a simple, yet effective, approach to estimate the polarity association biases using two benchmark datasets, SemEval-2017 (Rosenthal et al., 2017) and GoEmotions (Demszky et al., 2020). We employ that approach to examine these

biases in five representative LLMs. Our study reveals a moderate to strong occurrence of false negative errors in these LLMs when identifying neutral text instances, potentially induced by their underlying association biases. Additionally, we observe that the presence of certain linguistic or psychological word categories correlates with the tendency of these models to misclassify neutral text as either `positive` or `negative`, presenting consistent patterns as the aforementioned "stereotype". These key findings suggest that LLMs may have developed biased associations between particular word categories and a sentiment. We thus caution against the application of the current LLMs for large-scale sentiment classification, as, if the LLM employed has a polarity bias, it can exaggerate the estimation of sentiment trends (e.g., on social media) towards positive or negative.

2 Methodology

We estimate the severity of the **polarity association bias** through the measurement of false negative (FN) error rate in sentiment classifiers on *neutral* instances. We acknowledge that the observed FN could also be caused by other factors, such as underfitting. However, based on existing reports demonstrating the robust generalization capabilities of LLMs in predicting positive and negative sentiment (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023), we set aside underfitting as a primary cause in our study. We focus on measuring the FN rate on instruction-tuned foundational LLMs with zero-shot setting, rather than prompt-tuned or supervised fine-tuned LLMs using a sentiment classification dataset. We do so because FN in models continually tuned with sentiment classification data may be more attributable to extrinsic hallucination or noise rather than being a reflection of intrinsic bias in the pre-trained foundational LLMs (Ladhak et al., 2023).

Data: We collect $7,323$ neutral text instances from the SemEval-2017 Task 4 dataset (Rosenthal et al., 2017) and $12,748$ instances from the GoEmotions dataset (Demszky et al., 2020), all manually annotated. The datasets capture distinct posts from two different social media platforms. The SemEval2017 dataset was compiled from Twitter (which has become the platform X since the dataset construction). It primarily includes short and public-facing content often directed outward to a broad audience, focusing on trending topics like political events, product reviews, and news (e.g., "*gun control*", "*iPhone*"). The GoEmotions dataset consists of popular English subreddits comments collected for emotion annotation. Its posts were predominantly contributed by young male users (Duggan and Smith, 2013). They are typically shorter (i.e., fewer words) than the posts from SemEval-2017 (see average word count 23 > 12 in Appendix B), more self-reflective and emotional. Although the GoEmotions annotations focus on fine-grained emotional states like joy, sadness, and anger—differing from the polarity-based annotations of positive, negative, and neutral in SemEval-2017—the neutral emotional state aligns with the neutral sentiment, according to the annotation scheme (Demszky et al., 2020).

Models and Configurations: We set up a baseline by applying an "*off-the-shelf*" sentiment classifier (Camacho-Collados et al., 2022). This classifier is RoBERTa fine-tuned using the SemEval-2017 sentiment dataset. The GoEmotions dataset was not used in its finetuning process. We explore five instruction-tuned LLMs: `gemma2-2b-it` (Rivière et al., 2024), `Llama3.1-8B-Instruct` (Dubey et al., 2024), `deepseek-llm-7b-chat` (DeepSeek-AI, 2024), `gpt3.5-turbo` (OpenAI, 2024), and `gpt4-turbo` (OpenAI, 2023), by creating a uniform zero-shot sentiment classification prompt (see Appendix C). For each LLM, we experiment with three different temperature settings (Appendix D) to enable diverse reasoning paths on the same instance and consolidate the final prediction through a majority voting mechanism. We find that varying the temperature settings had minimal influence on the predictions across all the experimented LLMs. This indicates that the sentiment classification task exhibits low intrinsic randomness, and that the experimented LLM is highly confident in its predictions. This strongly suggests that the false negatives are more likely to be induced by the learned biases rather than the stochastic factors of language models.

Identification of Lexical Mentions Regarding Various Linguistic or Psychological Word Categories: We utilize Linguistic Inquiry and Word Count (LIWC2015) (Pennebaker et al., 2015) to identify lexical word mentions covering

a broad spectrum of linguistic and psychological categories (see Appendix E, the first column of Table 3), e.g., "*anger*", and "*family*". We calculate (with LIWC2015) the psychometric scores to characterize each text instance with respect to each of these LIWC categories. Let N denote the total word count in an input text instance and C_d the total number of words belonging to d, a specific LIWC2015 category, for the same instance. The psychometric score is calculated as $s_d = \frac{C_d}{N} \times 100, s_d \in [0, 100]$. In nature, the distribution of s_d is skewed, with most instances having $s_d \in [0, 30]$ (see Appendix E, Figure 5).

Correlation Between Categorical Word Mention and False Negative Likelihood: For a given word category d, we first sort all neutral instances in ascending order based on their s_d values. We then partition the sorted instances into smaller subgroups using logarithmic binning (see Appendix F for details). Within each subgroup, we compute the FN rate, denoted as ε_d, further categorized as:

- **directed FN rate**: The proportion of neutral instances misclassified as `positive` sentiment (ε_d^+) or as `negative` sentiment (ε_d^-) respectively within the subgroup.
- **undirected FN rate** ($\varepsilon_d^\pm$): The proportion of misclassified neutral instances in the subgroup.

For each subgroup, the computed ε_d is then assigned to all instances within that subgroup, representing the likelihood L_{ε_d} that an instance may be misclassified towards either `positive` sentiment $L_{\varepsilon_d}^+$ or `negative` sentiment $L_{\varepsilon_d}^-$.

Finally, we calculate the Pearson correlation coefficient[1] r_d and the underlying p-value (p) between s_d (categorical word mention score) and $L_{\varepsilon_d}^\pm$ (FN likelihood).

3 Results and Discussion

Severity of the Polarity Association Biases

All five LLMs exhibited an FN greater than RoBERTa, on both SemEval-2017 and GoEmotions dataset (except for `llama3` on GoEmotions), suggesting moderate to strong severity of bias (Figure 2). Notably, `gemma` showed the highest FN

[1] https://docs.scipy.org/doc/scipy-1.15.0/reference/generated/scipy.stats.pearsonr.html

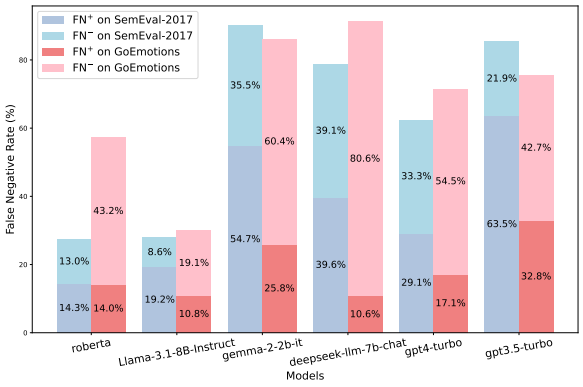

Figure 2: The LLMs' FN on the SemEval-2017 and GoEmotions datasets. The FN$^+$ denotes the proportion of mis-classifications as `positive` and FN$^-$ the proportion of mis-classifications as `negative`. The FN is the sum of FN$^\pm$ on each stacked bar.

(90.2%) on SemEval-2017, while `deepseek` had highest FN (91.2%) on GoEmotions. RoBERTa exhibits a lower FN on SemEval-2017 compared to GoEmotions, which is not surprising since it was fine-tuned on SemEval-2017. This difference in FN may suggest that, while fine-tuning can effectively reduce FN in identifying neutral instances when dealing with similar text (i.e., public-facing and intended for a broad audience) from the same social media platform, it may not effectively mitigate the intrinsic biases rooted in the pre-training stage of language models. Such biases could persist and re-emerge when applied to different texts (e.g., different social media platforms), exhibiting different language styles (i.e., inward-focused, self-reflective, and emotionally expressive, as in GoEmotions).

We also observed that all five LLMs tended to classify neutral instances in the SemEval-2017 dataset as `positive` but as `negative` in the GoEmotions dataset. Although both datasets contain more positive affect-related words (e.g., words expressing positive and negative emotions) than negative ones, the difference is smaller in GoEmotions (see `posemo` (positive emotions) and `negemo` (negative emotions) in Appendix E, Table 3). We assume that affect-related words play a key role in polarity association bias, and that, beyond this bias at the lexical level, additional factors (e.g., specific categories, like work or family) may also influence models to assign a `negative` sentiment.

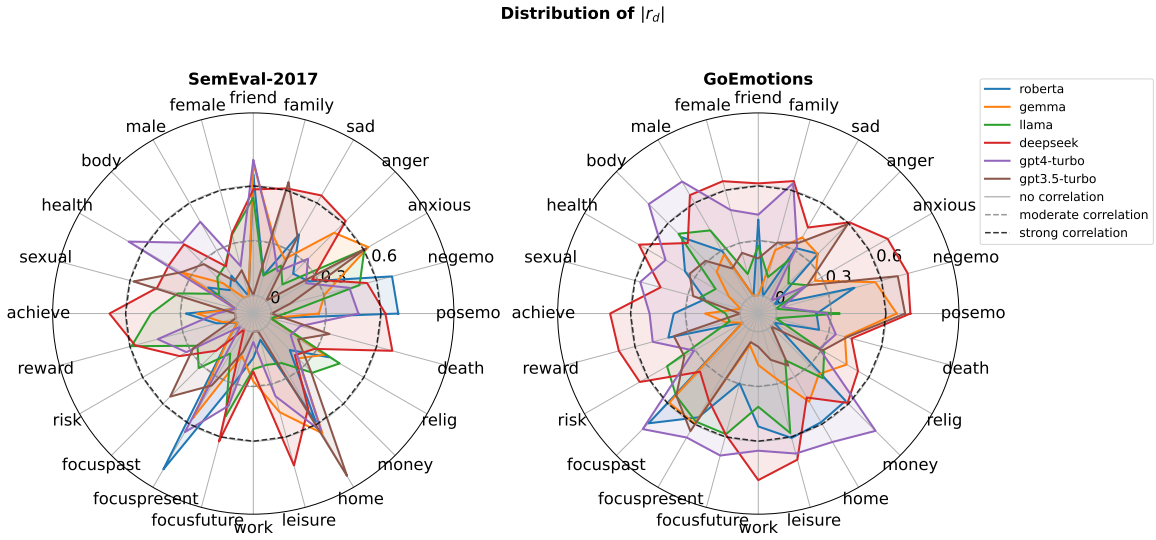

Figure 3: The distribution of $|r_d|$ on the SemEval-2017 and GoEmotions datasets. All negative r_d are converted into positive value for representing the intensity of the correlation. r_d are converted into 0 when there is no significant correlation ($p \geq 0.05$). $|r_d| \in [0.3, 0.6]$ denotes a moderate correlation, and $|r_d| > 0.6$ denotes a strong correlation.

Correlation Measure Between Categorical Word Mentions and False Negative Likelihood

We observed that both RoBERTa and the five LLMs exhibit a moderate to strong r_d between s_d (categorical word mention score) and the L_{ε_d} (FN likelihood) when classifying neutral text (Figure 3), suggesting that the models all exhibit polarity association biases. Notably, RoBERTa, which was fine-tuned on the SemEval-2017 dataset, showed a small r_d, but this value increased significantly when tested on the unseen GoEmotions dataset. This suggests that, while fine-tuning contextual language models can reduce r_d when classifying neutral instances, it may not effectively mitigate the bias, since the correlation persists when applied to unseen data.

The measured r_d on GoEmotions dataset among the five LLMs tends to be larger than on the SemEval-2017 dataset, reflected by the larger size of shadowed area on Figure 3. We hypothesize that, as text instances on GoEmotions are more subjective, self-reflective, and emotionally nuanced, they are more likely to trigger a polarity association bias.

The `deepseek` model appears to be the most affected by polarity association bias, with the highest r_d across almost all the experimented word categories. Additionally, we observed no significant reduction in r_d from `gpt3.5` to its more advanced version, `gpt4`, suggesting that the association bias can not be easily mitigated. `llama` resulted with moderate r_d on both datasets. We thus argue that `llama` may not be consistently reliable for sentiment analysis. In particular, if the data on which it is applied contains more words from the categories which showed a moderate r_d in our evaluation, this model may display a more pronounced bias and larger FN on processing neutral instances.

A detailed list of r_d and more comprehensive analysis is included in Appendix G. The results suggest that LLMs are not reliable for sentiment analysis, as polarity association biases can lead to mis-classifications of neutral instances and thus distort the estimation of sentiment trends.

4 Conclusion

Sentiment analysis has significant implications for many real-world applications. While LLMs can be instructed to conduct sentiment classification in a zero-shot setting, we argue that they are not yet fully reliable for large-scale sentiment analysis, especially when the data contains neutral texts.

Our study shows that LLMs have a high likelihood of misclassifying neutral text with either positive or negative sentiment. These mis-classification errors exhibit a moderate to strong correlation with the presence of specific categories of lexical words in the text. These findings suggest that LLMs may have developed a moderate to strong degree of polarity association bias, which can distort the estimation of sentiment trends.

Limitations

This study examines the polarity association bias in LLMs by only utilizing the datasets generated by English speakers, while cultural background can play a significant role in influencing the estimation of bias in LLMs (Imran et al., 2020). In real-world situations, multiple factors can contribute to polarity association bias, often making it challenging to isolate their individual effects. Our study focuses on lexical factors and assumes that each word category independently influences polarity association bias. Our experimental results on correlation do not establish a causal relationship between lexical word mentions and the likelihood of bias occurring.

We have not examined all the existing LLMs at the time of the submission. While it is not likely to change the findings, the experimental results represent a sub-set of all the LLMs that is available.

Ethical Concerns

We relied on the dataset providers to remove any material from the dataset that may reveal anyone's identity in their posts used in this study.

Our project has ethics approval from our affiliated organization. Details will be provided upon publication.

References

Nirmal Varghese Babu and E Grace Mary Kanaga. 2022. Sentiment analysis in social media data for depression detection using artificial intelligence: a review. *SN Computer Science*, 3(1):74.

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez-Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49.

DeepSeek-AI. 2024. DeepSeek LLM: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maeve Duggan and Aaron Smith. 2013. 6% of online adults are reddit users. *Pew Internet & American Life Project*, 3:1–10.

Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data*, 2:1–14.

Ali Shariq Imran, Sher Muhammad Daudpota, Zenun Kastrati, and Rakhi Batra. 2020. Cross-cultural polarity and emotion detection using sentiment analysis and deep learning on COVID-19 related Tweets. *IEEE Access*, 8:181074–181090.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Huang. 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 431–469.

OpenAI. 2023. GPT4. Version: firstcontact-gpt4-turbo 2023-03-15-preview.

OpenAI. 2024. ChatGPT 3.5. Version: chatgpt-35-turbo 2024-02-15-preview.

Venkata Sasank Pagolu, Kamal Nayan Reddy, Ganapati Panda, and Babita Majhi. 2016. Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350. IEEE.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *University of Texas at Austin*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, et al. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2023. Is ChatGPT a good sentiment analyzer? A preliminary study. *arXiv preprint arXiv:2304.04339*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

A Sentiment Classification on the SemEval-2017 Subset

We randomly selected $3,000$ manually annotated instances from the SemEval-2017 and applied the fine-tuned RoBERTa (Camacho-Collados et al., 2022), `gpt4` (OpenAI, 2023), and `llama3` (Dubey et al., 2024) to measure the F_1 score in the classification of `positive`, `negative`, and `neutral` instances (1000 instances within each class). The results suggest that LLMs are effective in identifying positive and negative instances but not for neutral instances (Table 1). The results on the SemEval-2017 dataset closely align with existing studies on other sentiment analysis datasets (Zhang et al., 2024; Qin et al., 2023; Wang et al., 2023; Laskar et al., 2023).

Model	Label Classes		
	`positive`	`negative`	`neutral`
`RoBERTa`	0.83	0.84	0.73
`gpt-4`	0.78	0.79	0.54
`Llama3`	0.65	0.64	0.59

Table 1: F_1 of sentiment classification by label class on SemEval-2017 Subsamples.

B Statistics of Word Count Per Post by Dataset

The distribution of post length (i.e., word count) by dataset.

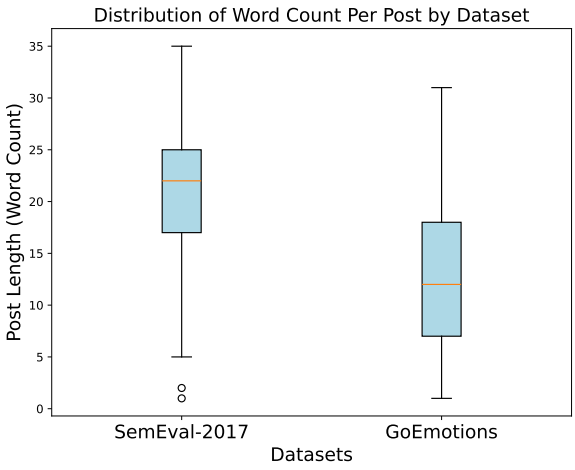

Figure 4: Distribution of word count in each post among SemEval-2017 and GoEmotions datasets.

C Prompt Instruction for Sentiment Classification

The following prompt template was used to instruct LLMs to perform sentiment classification with zero-shot settings: "`You are a sentiment analyzer. Does the following text express sentiment of negative, neutral, or positive: <text>. REQUIREMENT: only answer -1 as negative, 0 as neutral, 1 as positive.`" The "`<text>`" is filled with the full text content of each instance.

For the examples shown in Figure 1, we added "`and explain why`" to the "`REQUIREMENT`" to provide the explanation.

D Models and Configurations

The configuration for running RoBERTa and the open resource LLMs are shown in Table 2.

params	value
GPU	NVIDIA RTX 3500 Ada
context size	512
temperature	0.01; 0.5; 1
max_new_tokens	8
quantization*	4bits

Table 2: The environment setting and parameters in zero-shot learning setting. $*$ indicate that, except for the proprietary `gpt-3.5` and `gpt-4`, the open-source LLMs were loaded with 4-bits quantization on local server for computational feasibility. The `max_new_tokens` is set to 8 for all LLMs, as we only require the models to generate either -1, 0, and 1 for the representation of `negative`, `neutral`, and `positive` sentiment for a given input text.

E Statistics of Psychometric Scores by Word Category

The $mean$ of s_d by each LIWC word category is illustrated in Table 3. The distribution of s_d on the SemEval-2017 and GoEmotions dataset is illustrated in Figure 5.

Category	SemEval-2017	GoEmotions
posemo	2.1614	3.2057
negemo	1.4166	3.0153
anxious	0.1594	0.2492
anger	0.5243	1.1869
sad	0.3346	0.4648
family	0.3341	0.4960
friend	0.2282	0.6309
female	0.4765	1.0293
male	1.3611	2.1440
body	0.4841	0.8732
health	0.2824	0.6461
sexual	0.1611	0.2933
achieve	1.2716	1.2204
reward	1.1344	1.5200
risk	0.3300	0.6670
focuspast	2.2963	3.5040
focuspresent	8.2916	12.9773
focusfuture	3.2204	1.1631
work	1.5153	1.4472
leisure	1.9015	1.2825
home	0.2152	0.3159
money	0.5386	0.7860
religion	0.6162	0.3701
death	0.1959	0.2937

Table 3: Comparison of $mean$ of each LIWC2015 psychometric word category score (s_d) on the SemEval-2017 and GoEmotions Datasets. `posemo` and `negemo` denotes words related to positive and negative affect.

F Logarithm Binning for Psychometric Scores Partitioning

Since the distribution of the psychometric score (s_d) in both datasets are highly skewed (Figure 5), with most s_d falling below 30, we apply logarithmic binning with base equals to 10 to create 50 intervals with increasing width. Let x_i represent the edges of each interval,

$$x_i = 10^{\log_{10}(1)+i\cdot\Delta\log}, \quad \text{where } i = 0, 1, 2, \ldots, 50$$

Determining $\Delta\log$:

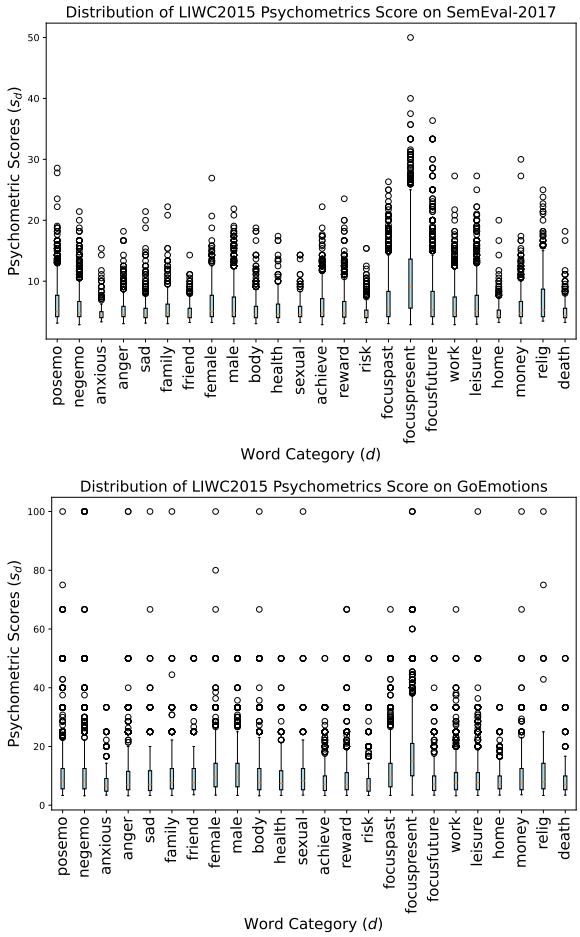

Figure 5: The distribution of psychometric scores (s_d) on the two benchmark datasets by LIWC2015 word category (d).

The range of the logarithmic scale is:

$$\log_{10}(100) - \log_{10}(1) = 2 - 0 = 2$$

Divide this range into 50 partitions:

$$\Delta\log = \frac{\log_{10}(100) - \log_{10}(1)}{50} = \frac{2}{50} = 0.04$$

This approach groups instances with lower s_d into narrower interval and instances with higher s_d into wider interval. Instances with s_d equals to 0 are excluded for the correlation study. Subgroups containing fewer than 10 instances are also excluded to ensure the stability and reliability of FN likelihood (L_{ε_d}) calculation.

G Results and Discussion of Correlation Coefficient Measure

We find that the presence of certain word categories shows a moderate to strong correlation with the models' tendency to misclassify neutral

text into either positive or negative (Table 6). For example, mentioning leisure-related words is strongly correlated (0.74) with a positive sentiment for gemma, whereas work-related words strongly correlate (0.77) with a negative sentiment for deepseek. This suggests that these models may have developed a significant ($p < 0.05$) bias in associating specific word mentions with a positive or negative sentiment, despite the text instances being neutral.

The fine-tuned RoBERTa and pre-trained foundational LLMs exhibit varying levels of r_d (correlation coefficient) between s_d (categorical word mention scores) and $L^{\pm}_{\varepsilon_d}$ (FN likelihood), ranging from moderate ($0.3 \leq |r_d| \leq 0.6$) to strong ($|r_d| > 0.6$), across both datasets (Table 6). In most cases, we can observe consistent signs of r_d between the two datasets for the same model. For example, both llama3.1-8B and gpt-4 demonstrated positive r_d between family-related lexical mentions and $L^{+}_{\varepsilon_d}$, while showing negative r_d with $L^{-}_{\varepsilon_d}$. The contrasting sign in the correlation suggests that a model is likely to classify an instance as positive if there are more family-related words in the text. However, a few models showed a contradictory sign direction in r_d between two datasets for a given word category. For example, in the "*religion*" category, llama3.1-8B had a negative r_d (-0.57) between s_d and $L^{+}_{\varepsilon_d}$ on the SemEval-2017 dataset, but a positive r_d (0.48) between the two on the GoEmotions dataset. The lack of consistency in the sign direction of the correlation suggests that this word category may not be the primary factor in the mis-classifications of neutral posts for the specific model. There might be other factors not covered by this study. However, we observed that such inconsistency in the correlation sign direction is rare in our study (see Table 6), suggesting that the existence of polarity association biases in these models mostly holds.

Comparing correlations across different models, we observe that deepseek exhibits close to a strong correlation with all evaluated word categories, whereas other models show only moderate to strong correlations with specific categories. This suggests that deepseek may have developed the most pronounced bias among the tested models. We hypothesize that deepseek may have been pre-trained using a cost-efficient approach with reduced exposure to a large volume of diverse data, which could have led to a more pronounced bias compared to other LLMs. Additionally, the five LLMs demonstrate a similar level of correlation severity to RoBERTa, implying that the bias may cause these models to overlook entire post content when assessing sentiment. Notably, there is no clear reduction in correlation when comparing larger models to smaller ones (e.g., llama3-8B cf. gemma2-2b) or in the transition from gpt-3.5 to gpt-4. This suggests that such correlations are likely learned biases from pre-training rather than underfitting.

In conclusion, while the presence of polarity association biases in these models is likely influenced by multiple factors, the moderate to strong correlations between lexical word mentions and FN likelihood on neutral text expressions suggest that these LLMs may have developed moderate to strong severity of polarity association bias for sentiment classification. We argue that addressing such bias is critical to ensuring fairer estimation of sentiment trend before applying on the analysis of large-scale social media text data which might contain a significant number of neutral instances.

H Estimation of Computational Cost

The total GPU hours for running open-resource LLM with 4-bits quantization on the local NVIDIA RTX 3500Ada GPU is shown in Table 4. The budget for running proprietary LLMs is shown in Table 5.

Model	SemEval-2017	GoEmotions
RoBERTa	< 0.2	< 0.2
Llama3	≈7	≈8
Deepseek	≈7	≈8
Gemma2	≈4	≈5

Table 4: Estimation of approximate GPU hour by LLM on each dataset.

Model	SemEval-2017	GoEmotions
gpt4-turbo	≈20.38	≈29.22
gpt3.5-turbo	≈1.02	≈1.48

Table 5: Estimation of total cost (USD) for running proprietary LLMs on each dataset.

	Category	RoBERTa		Gemma-2-2b		LLaMa3.1-8B		Deepseek-7b		GPT-4-turbo		GPT-3.5-turbo	
		r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-
Affect	posemo	0.77	-0.30	0.30	-0.24	0.42	-0.61			0.61	-0.50		
	negemo		0.66					-0.41	0.60		0.40	0.30	-0.57
	anxious	-0.34		0.91	0.88	-0.80	0.94	-0.77	0.75	-0.78	0.49	-0.89	0.64
	anger							-0.69	0.74		0.31		
	sad	-0.42							0.41			-0.35	
Social	family				-0.39		-0.66	0.39	0.66	0.56	-0.63		-0.66
	friend	-0.60	-0.45		-0.77	0.25	0.33	-0.80	0.95		-0.44	-0.78	0.51
	female					-0.52	0.25	-0.44	0.59				
	male					-0.38	0.28	-0.54	0.69		0.51	-0.35	0.30
Bio	body							-0.86	0.81				
	health		0.32	-0.41	0.40	-0.47	0.36	-0.78	0.34	-0.54		-0.45	0.34
	sexual						-0.26	-0.75	0.64				-0.48
Drives	achieve		-0.37				-0.52	-0.27	0.67				
	reward	0.30		0.43	-0.47		-0.53	-0.28	0.61		-0.59		
	risk							-0.54	0.55		-0.98	0.50	-0.91
Time	focuspast			-0.38	0.42			-0.64	0.55			-0.57	
	focuspresent		0.88	-0.62	0.79			-0.75	0.69		0.71	-0.69	0.55
	focusfuture		-0.40	0.46	-0.38		-0.68	-0.71	0.75		-0.57	0.33	-0.45
Personal Concerns	work				0.32	-0.33	0.24	-0.32	0.34				
	leisure		-0.35	0.74	-0.63	0.31	-0.26	0.54		0.26	-0.78	0.60	-0.74
	home		-0.64	-0.43			-0.76	-0.73	0.92	-0.48		-0.72	
	money						-0.57	-0.44	0.50		-0.35		
	religion	0.48			0.34	-0.57	0.24						
	death			-0.63	0.40	-0.27	0.45	-0.61	0.71			-0.66	0.35

(a) $r_d^{\pm}$ on SemEval-2017

	Category	RoBERTa		Gemma-2-2b		LLaMa3.1-8B		Deepseek-7b		GPT-4-turbo		GPT-3.5-turbo	
		r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-	r_d^+	r_d^-
Affect	posemo	0.91	-0.87	0.89	-0.87	0.84	-0.83			0.92	-0.90	0.93	-0.86
	negemo		0.47		0.39			-0.74	0.80			0.77	
	anxious							-0.32	0.86	-0.78	0.49		
	anger	0.44				0.62	-0.49	-0.54	0.68				0.45
	sad	-0.43			0.34			-0.64	0.64				0.33
Social	family			0.81	-0.73	0.72	-0.59	-0.29	0.79	0.58	-0.87	0.69	-0.68
	friend	0.66	-0.71	0.79	-0.77	0.62	-0.65	-0.55	0.62	0.70	-0.82	0.69	-0.59
	female					0.40	-0.58	-0.54	0.71	0.32	-0.57		
	male	0.62	-0.51	0.66	-0.68	0.45	-0.71	-0.50	0.70	0.45	-0.78	0.55	-0.47
Bio	body	0.43	-0.67					-0.48	0.56		-0.86	0.34	-0.55
	health	0.30	-0.59			0.38	-0.40	-0.69	0.74			-0.32	
	sexual	0.65	-0.61				-0.26	-0.52	0.55	0.67	-0.79	0.53	
Drives	achieve	0.67	-0.65	0.71	-0.67	0.86	-0.69	-0.62	0.77	0.79	-0.89		
	reward	0.73	-0.77	0.79	-0.76	0.73	-0.75	-0.58	0.79	0.83	-0.82	0.76	-0.61
	risk						-0.42	-0.59	0.76				0.46
Time	focuspast	0.44	-0.83			0.38	-0.65	-0.68	0.61		-0.79	-0.56	
	focuspresent	0.36	-0.58	0.46	-0.54	0.45	-0.77	-0.41	0.55	0.41	-0.70	0.34	-0.71
	focusfuture	0.76	-0.77	0.79	-0.74	0.68	-0.76	0.37	0.59	0.67	-0.90	0.88	-0.77
Personal Concerns	work		-0.63	0.40	-0.53	0.25	-0.50	-0.57	0.77	0.30	-0.76		
	leisure	0.33	-0.76	0.73	-0.70	0.43	-0.82	-0.47	0.73		-0.86	0.54	-0.58
	home		-0.62			0.53	-0.35	-0.40	0.69		-0.76		
	money		-0.64				-0.49	-0.59	0.76		-0.75		
	religion			0.68	-0.58	0.48	-0.76	-0.26	0.60	0.79	-0.76	0.78	-0.69
	death					0.46		-0.49	0.59		-0.38		

(b) $r_d^{\pm}$ on GoEmotions

Table 6: Correlation coefficients $r_d^{\pm}$ between word mention score s_d and FN likelihood $L_{\varepsilon_d}^{\pm}$ on the SemEval-2017 (SemEval) and GoEmotions (GoEmo) datasets. Positive r_d^+ indicates that the incremental s_d positively correlates with the model's preference to falsely identify a neutral text as an expression of `positive sentiment`. $r_d < 0.3$ with $p \geq 0.05$ is not displayed as there is no significant correlation.