Tell-tale Signs of Implicit Bias: Language Abstraction for Automated Bias Analysis

Xulang Zhang[1], **Rui Mao**[1], **Mengshi Ge**[2], **Erik Cambria**[1]
[1] Nanyang Technological University, Singapore
[2] National University of Singapore, Singapore
{xulang.zhang, rui.mao, cambria}@ntu.edu.sg, mengshi.ge@nus.edu.sg

Abstract

Our implicit biases are often reflected in our utterances. Existing AI research works are limited by their narrow focus on predefined, explicit biases. Thus, we propose a more generalized approach to bias analysis by implementing Linguistic Intergroup Bias (LIB) theory, which suggests people tend to use abstract terms to describe in-group positive and out-group negative behaviors, and concrete terms for the opposite. To leverage LIB, we propose a novel task named language abstraction span extraction, and finetuned an LLM for this task. With the model, we analyze the intergroup bias between the political left and right on Twitter and news datasets. Results show that our method can be used for bias analysis both quantitatively and qualitatively. More importantly, our paper provides the first large-sample-based empirical evidence to support LIB, affirming the correlation between abstraction and intergroup bias in social media and news domains.

Keywords: Bias analysis; natural language processing; linguistic intergroup bias; large language model

Introduction

Bias has long been a focus of study in the NLP research community. Analyzing bias in societies is crucial for identifying and addressing systemic inequalities, advancing fairness, and strengthening social cohesion (Muchnik et al., 2013). The rise of Large Language Models (LLM) and their bias analysis have highlighted the presence of diverse biases inherent in human-produced texts, which these models learn from and propagate through their outputs (Mao et al., 2023, 2025). As such, understanding how bias is expressed in text is critical to mitigating the propagation of discriminatory patterns in LLMs and preventing the reinforcement of structural inequities within societal systems (Cambria, 2024).

Recent research on bias in NLP primarily focuses on identifying and mitigating harmful biases, such as those related to hate speech and stereotypes, particularly in the contexts of gender and race (Cambria, Mao, Zhang, et al., 2026; Hartvigsen et al., 2022; Nadeem et al., 2025; Raza et al., 2024; Sap et al., 2020). These models, while effective, require diverse data encompassing various social and cultural groups to provide sufficient background. However, in the age of social networks, social biases, such as stereotypes, are rapidly evolving. These models must consistently mine for knowledge to catch up. For instance, "the blue-haired girl" is a relatively new and niche prejudice that is not captured by many existing stereotype datasets and trained NLP models.

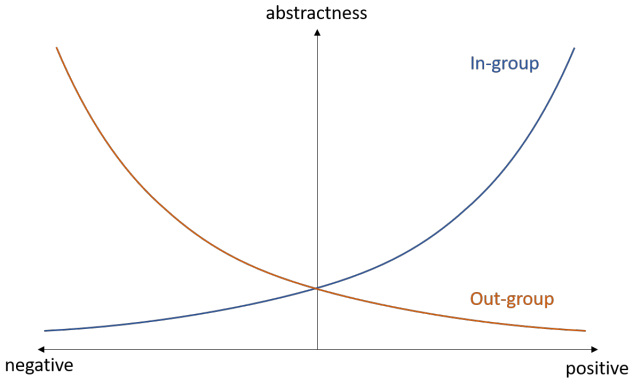

Figure 1: Illustration of the concept of LIB theory.

In addition, they tend to focus only on widely recognized disadvantaged groups, since they are primarily concerned with reducing harms such as sexism and racism. Thus, a critical limitation of existing NLP works lies in the narrow focus on pre-defined, widely recognized forms of bias, which are undoubtedly important but insufficient to capture the complexity of evolving intergroup and social prejudices.

Taking inspiration from psychology and sociolinguistics, we would like to approach bias from a more general sense, where it could be directed towards any groups or individuals (Greenwald & Banaji, 1995; Van Dijk, 2009). Specifically, we are informed by Linguistic Intergroup Bias (LIB) theory (Maass, 1999; Maass et al., 1996). The main observation of the theory is that positive behaviors of in-group members and negative behaviors of out-group members will more likely be described using abstract language, while the same behaviors by the opposite group will be depicted using concrete language (see Figure 1).

That is, bias manifests from the differential usage of abstraction in describing behaviors based on group membership, which can reinforce and perpetuate stereotypes. This differentiation of abstraction in depicting behaviors based on group membership contributes to bias formation, as it shapes the perception of in-group and out-group members in fundamentally different ways. For instance, if someone committed a violent act, an in-group member may describe this person as "hurting somebody" (concrete), whereas an out-group member may describe this person as "aggressive" (abstract).

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

This form of biased language use has been proven to occur in a variety of intergroup dynamics, including sports teams, nations, gender, political and interest groups, etc (Fiedler et al., 1993; Maass, 1999; Schmid, 1999). LIB is rooted in the fact that the more abstract a statement is, the more it suggests temporal and cross-situational consistency, and the more difficult it is to be verified and to present refuting evidence (Semin & Fiedler, 1988). Unlike biases that are narrowly focused on predefined groups, or on distinct forms of discriminatory language like hate speech, the LIB approach provides a more comprehensive framework for examining how bias is embedded within the broader structure of language. Further, unlike other biased expressions, language abstraction is unlikely to be consciously controlled. While people may be intentionally depicting a group or individual in a positive or negative light, they are rarely aware of the level of abstraction they are using (Douglas et al., 2008; Franco & Maass, 1996; Maass et al., 1996; Schmid & Fiedler, 1998). As such, it can be used to detect subtle and unintentional bias. Especially in domains that purport neutrality, such as news, court cases, and education, abstraction could serve as a useful measure for detecting and mitigating implicit bias, particularly those rooted in subconscious attitudes and perceptions.

Therefore, we proposed a novel span-based sequence labeling task that extracts the descriptive words and phrases in an input sentence, and labels their abstraction level, based on the Linguistic Category Model (LCM) (Schmid & Fiedler, 1998; Semin & Fiedler, 1991) adopted by LIB researchers. We created a ranking algorithm to select the top 300 most diverse and representative sentences from the Wall Street Journal (WSJ) dataset and manually annotated them. Using this dataset, we adopted task-specific instruction-based finetuning to train an LLM for the abstraction span extraction task. We applied the LLM to conduct bias analysis on an intergroup Twitter dataset and a new article dataset. Results show that our method can be employed for automated LIB analysis, and further confirmed LIB theory. Finally, we perform a case study on selected news articles from the latter dataset to showcase the capability of our proposed method in explainable qualitative bias analysis.

The contributions are: (1) We proposed a novel abstraction span extraction task useful for automated bias analysis, along with a small high-quality labeled dataset. (2) We presented an abstraction span extraction model for an explainable, human-in-the-loop bias analysis.[1] (3) Using the proposed model, our paper is the first to validate LIB theory on a larger scale in two different domains, i.e., Twitter and news.

Related Work

NLP-based Bias Analysis

NLP research on bias in text is predominantly concerned with detecting inappropriate language use towards certain demographics, relating to aspects such as gender, race, nationality, sexuality, etc (Cambria, Mao, Hussain, et al., 2026; Pair et al., 2021; Pan et al., 2023; Raza et al., 2024; Sap et al., 2020; Spinde et al., 2021). LLM-driven bias detection methods largely rely on data to learn the diverse scenarios of bias (Hartvigsen et al., 2022; Kumar et al., 2024; Raza et al., 2024; Yu et al., 2024). As such, they can only be as competent as the datasets taught them to be, which limits the scope of bias they can look for.

Closest to our paper, some existing works have explored intergroup bias, which presented annotated datasets for training models to identify interpersonal group relationships and affects, revealing that LLMs can learn and reflect some language patterns that signal group membership depending on the domains of text (Dong et al., 2024; V. Govindarajan et al., 2024; V. S. Govindarajan et al., 2023). However, the language abstraction aspect of LIB has not been studied by the NLP research community to the best of our knowledge.

LCM-based Bias Analysis

In the field of psychology, language abstraction is considered to be integrally linked with bias, because of its perceived high level of stability, dispositional inference, and likelihood of repetition (Maass, 1999). Specifically, the abstraction levels are measured by the LCM, which defines four categories of terms. The most concrete terms are Descriptive Action Verbs (DAV), which describe specific, observable events such as *"A hit B"*. The second category is the Interpretive Action Verbs (IAV), which describe a specific event in a slightly more abstract way, e.g., *"A hurt B"*. The most abstract verb category is the State Verbs (SV), which describe prolonged psychological states that go beyond a specific event, such as *"A hates B"*. Lastly, the most abstract terms are the adjectives (ADJ), which describe general, cross-situational dispositions such as *"A is aggressive"*. More recent developments of LIB theory (Carnaghi et al., 2008; Geschke et al., 2010; Graf et al., 2013) extend the LCM to include a fifth category more abstract than ADJ, nouns (NOUN), such as *"A is an aggressor"*, because such nouns tend to conjure up richer inferences and assumptions about the described target than adjectives, e.g., *"an artist"* versus *"artistic"*. Generally, in bias analyses, DAVs and IAVs are considered as concrete terms, whereas SVs, ADJs, and NOUNs are considered as abstract terms. It is worth noting that under this definition of abstraction, more "detailed" does not equal less abstract. For instance, *"A young female scholar attended the conference"* is more abstract than *"A scholar attended the conference"*. The definition of LCM categories is given in the preliminary section.

Using this framework, studies have been conducted to analyze bias in different contexts. It is found that the LIB phenomenon is present in a variety of scenarios. To name a few, Maass (1999) conducted experiments that indicate LIB occurs in both prompted human survey settings, as well as spontaneous settings such as sports and political reports. Geschke et al. (2010) explored and verified the bias in mass media, especially 8 German news reports on migrants. Similarly, Dragojevic et al. (2017) examined the evidence of LIB on 4,663 sentences from news reports about U.S. immi-

[1] Code and data available at: github.com/SenticNet/LIB-analysis

gration, confirming the occurrence of LIB in news reporting.

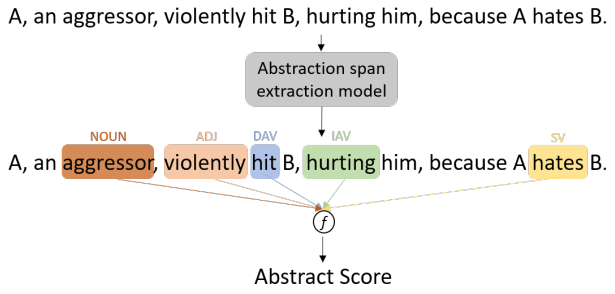

Figure 2: Illustration of abstract span extraction and score computation.

While effectively proving the universality of LIB, it can be gleaned from these prior works that LCM-based bias analysis is labor-intensive, making it especially difficult to conduct on large-scale corpora. As such, the psychology research community has come up with tools to automate LCM coding. However, these tools are mostly lexicon-based (Johnson-Grey et al., 2020; Seih et al., 2017; Snefjella & Kuperman, 2015), which heavily rely on abstractness lexicons to identify abstract and concrete terms in the input sentence. The limitations are two-fold: (1) Existing lexicons are not extensive enough to handle texts from various domains, particularly on informal texts such as tweets; (2) Lexicon-based methods are not well-equipped to handle the nuances in context. For instance, the LCM manual (Coenen et al., 2006; Johnson-Grey et al., 2020) stated that colors are not coded as ADJs. However, some color words have word senses that should be coded as ADJs, e.g., *"blue"* can be used to signify melancholy or the U.S. left-wing. Therefore, an LCM coding tool that takes into account both syntax (Zhang et al., 2023) and semantics (Mao, He, et al., 2024) is essential for abstraction-based bias analysis.

Methodology

Preliminary

In this work, we adopted the specific definition of abstractness in LCM. Following the LCM manual and its recent updates (Graf et al., 2013; Johnson-Grey et al., 2020), we use five categories for descriptive terms. The most concrete terms are DAVs, which provide a concrete and objective description of a specific behavioral event. Specifically, all actions that a DAV can describe invariantly share a physical feature. For instance, all the actions that *"talk"* can be applied to involve the physical feature of mouth. Examples of DAV include: *walk, read, dance*. IAVs are slightly more abstract as they describe a larger class of behaviors, although they maintain a clear reference to a specific behavior in a specific situation. They differ from DAV in that they do not share an invariant physical aspect. Examples of IAV include: *do, have, make*. The most abstract verb category is SV, which describes lasting psychological states that generalize beyond specific situations and behaviors. These states can be either cognitive or affective, and cannot be objectively verified. Examples of SV include: *think, understand, love, hate*. The fourth category is ADJ, which includes adjectives and adverbs describing a general disposition that applies across situations, behaviors, and objects. Notably, verbs that qualify the object of the sentence, e.g., the underlined part in *"I pass people waiting for the train"*, are considered ADJ. The most abstract category is NOUN, which refers specifically to nouns that qualify as the subject of a sentence. Although such nouns are coded as ADJ in the original LCM manual, recent updates (Carnaghi et al., 2008; Graf et al., 2013) suggest them as a standalone category. Aligning with the theory, only predicative nouns are viewed as NOUN, e.g., *"His father is a thief"*.

Abstraction Span Extraction Model

In this section, we introduce how we obtained the abstraction span extraction model. First, we proposed a ranking algorithm that samples the most diverse and representative sentences in the WSJ dataset, and labeled the descriptive text spans in each sentence based on LCM. Then, we created LCM-based prompts and employed instruction-based finetuning to train an LLM on our annotated dataset. Lastly, we present how the abstract score is calculated for quantitative analysis. Figure 2 shows the computational process of span extraction and abstract score computation.

Data Selection and Annotation. Our goal of data selection is to find a small set of sentences that encompasses a variety of language abstraction degrees and scenarios, to reduce manual labor while keeping the dataset effective for finetuning. As shown in the LCM manual and earlier automation efforts (Johnson-Grey et al., 2020), the abstract level of a term depends greatly on the sentence's syntax. Therefore, our intuition is to select sentences that are diverse in both topic and syntactic structure, while also filtering out the uncommon outliers that might lead to overfitting.

To this end, we used the WSJ dataset as our selection pool, as each data sample is labeled with part-of-speech tags. We pre-processed the WSJ dataset with sentence splitting so that each data sample contains one sentence. We modified the Maximum Marginal Relevance (MMR) sampling (Carbonell & Goldstein, 1998) to select the sentences as it balances diversity and relevance. Specifically, when comparing two sentences, s_1 and s_2, we took into consideration both the semantic similarity and the syntactic similarity. The former is computed as the cosine similarity of the embedding space, using RoBERTa-base (Liu et al., 2019) as the encoder: $\text{sem\_sim}(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|}$, where $v_i = Encoder(s_i)$. The latter is computed using the corresponding part-of-speech sequences p_1 and p_2, provided by the WSJ dataset: $\text{syn\_sim}(p_1, p_2)' = 1 - \frac{\sum_{l=1}^{\min(L_1,L_2)} \delta(p_1^l, p_2^l) + |L_1 - L_2|}{\max(L_1,L_2)}$, where L_1 and L_2 denote the length of p_1 and p_2, p_1^l and p_2^l denote the part-of-speech tag of l-th word in the respective sentence, and the function δ is defined as:

$$\delta(a,b) = \begin{cases} 1 & \text{if } a \neq b, \\ 0 & \text{if } a = b. \end{cases}$$

We normalize the syntactic similarity such that it also falls in the range of [-1, 1]: $\text{syn\_sim}(p_1, p_2) = 2 \cdot \text{syn\_sim}(p_1, p_2)' - 1$. The similarity of p_1 and p_2 is calculated as the weighted sum of their semantic similarity and syntactic similarity: $\text{sim}(p_1, p_2) = \gamma \cdot \text{sem\_sim}(p_1, p_2) + (1-\gamma) \cdot \text{syn\_sim}(p_1, p_2)$, where $\gamma \in [0, 1]$ is a hyperparameter. Then,for the MMR algorithm, S denotes the set of selected sentences, and C candidate sentences. We randomly initialize S. The MMR computes the diversity d and relevance r of each candidate $c \in C$: $d(c, S) = \max_{s \in S}(\text{sim}(c, s))$, $r(c, S) = \frac{1}{|S|} \sum_{s \in S} \text{sim}(c, s)$, and iteratively adds qualified candidates into S until it contains 300 sentences: $\text{MMR}(S, C) = \arg\max_{c \in C} (\lambda \cdot d(c, S) - (1 - \lambda) \cdot r(c, S))$, where $\lambda \in [0, 1]$ is a hyperparameter.

Subsequently, we annotated the data as a span-based sequence labeling task. The motivation for this task setting, in contrast to score regression, lies in its potential to facilitate the extraction of abstract terms, which could be valuable for qualitative analyses, enabling the construction of specific bias profiles. Following task convention, we adopt BIO annotation (Ramshaw & Marcus, 1999), where "B" stands for the beginning of a span, "I" for inside a span, and "O" for outside of any span. Each text span is labeled as DAV, IAV, SV, ADJ, or NOUN. For instance, the sentence *"She gently whispered."* would be labeled as "`O B-ADJ B-DAV O`". Each sentence is labeled by two qualified annotators with 0.86 Cohen's kappa, and disagreements are resolved via discussions.

Instruction-based Finetuning. With the labeled dataset, we can create the instruction to finetune an LLM, as shown below. We select an LLM as the backbone of our model due to its proficiency in handling a wide range of NLP tasks and its extensive linguistic knowledge (Mao, Chen, et al., 2024).

Abstract Score Computation. We adopt the computation method in the LCM manual with the addition of the NOUN category (Carnaghi et al., 2008; Graf et al., 2013). The abstract score (abst) of a given text is computed as below:

$$\text{abst} = \frac{\text{DAV}+2\cdot\text{IAV}+3\cdot\text{SV}+4\cdot\text{ADJ}+5\cdot\text{NOUN}}{\text{DAV}+\text{IAV}+\text{SV}+\text{ADJ}+\text{NOUN}}, \quad (1)$$

where DAV, IAV, SV, ADJ, and NOUN denotes the number of terms in each categories.

Experiments

Datasets. The **Interpersonal Group Relationship (IGR)** dataset (V. S. Govindarajan et al., 2023) contains U.S. Congress tweets mentioning other members, annotated as in-group or out-group and labeled with 8 emotions: *(admiration, anger, distrust, fear, interest, joy, sadness, surprise)*. We derive sentiment labels as follows: *admiration* and *joy* contribute 1 positive point; *anger*, *distrust*, and *fear* contribute 1 negative point; *interest* contributes 1 neutral point; and *surprise* contributes 0.5 positive and 0.5 negative points. Tweets with tied or zero positive/negative scores are labeled neutral; otherwise, they take the higher-scoring polarity. Since *sadness* does not necessarily indicate negative sentiment towards the target, tweets labeled only as *sadness* are excluded, yielding 2,796 tweets. See Table 1 for statistics.

Table 1: The IGR dataset used for experiments.

Group	Sentiment		
	Positive	Neutral	Negative
In-group	735	642	29
out-group	446	655	289

Table 2: The articles selected from POLUSA.

Leaning	Target	Sentiment	
		Negative	Non-Negative
Right	Republican	72	100
	Democrat	100	100
Left	Republican	100	100
	Democrat	100	100

The **POLUSA** dataset (Gebhard & Hamborg, 2020) contains 0.9M politically labeled news articles from 18 U.S. outlets. We selected left- and right-leaning articles mentioning Republicans or Democrats in the headline and prompted an LLM to determine whether each article negatively portrays the referenced party. We avoided standard sentiment polarity labels because overtly positive political reporting is uncommon. The final dataset contains 772 articles (6,747 sentences). See Table 2 for statistics.

Setup. We set $\gamma = 0.5$ and $\lambda = 0.7$ for data selection empirically for more grammatically diverse samples. We employed LLaMA-3.2-3b-instruct[2] as our LLM. For instruction-based finetuning, we used the AdamW optimizer (Loshchilov, 2017). The model is trained for 15 epochs with a learning rate set to 2e-4, and warmup steps set to 10. We used spaCy[3] to split the data in the WSJ dataset and the articles in the POLUSA dataset into sentences.

Evaluation. To evaluate the trained model's reliability, we randomly selected 100 of its outputs on the IGR dataset. Two annotators were tasked to evaluate whether each token received the correct label. They were instructed according to the LCM manual. The finetuned model achieved 80.3% in accuracy and 76.9% in F1 score with 0.89 Cohen's Kappa.

Results

Linguistic Intergroup Bias in Tweets

We applied our model to the IGR dataset to obtain the abstract scores of each tweet using Equation 1. After filtering the invalid samples with zero abstract score, we analyze the correlation between language abstraction and intergroup bias on the remaining 2,205 data samples. Table 3 shows the means and standard deviations of the abstract scores.

First, to verify LIB theory that in-group positive and out-group negative behaviors tend to be described more abstractly, while out-group positive and in-group negative behaviors tend to be described more concretely, we conduct the two-sample t-tests on the abstract scores of the in-group and out-group positive tweets, as well as the out-group and in-group negative tweets.

[2] kaggle.com/models/metaresearch/llama-3.2

[3] spacy.io/usage/linguistic-features

Table 3: The mean and standard deviation of abstract scores of different group-sentiment categories in IGR.

Category	Mean	Standard Deviation
Ingroup Positive	3.15	0.61
Ingroup Neutral	2.79	0.65
Ingroup Negative	2.62	0.54
Outgroup Positive	2.90	0.98
Outgroup Neutral	2.87	0.57
Outgroup Negative	3.05	0.47

Table 4: The mean and standard deviation of abstract scores of different group-sentiments in POLUSA.

Category	Mean	Standard Deviation
Ingroup Non-negative	2.76	0.14
Ingroup Negative	2.81	0.26
Outgroup Non-negative	2.78	0.18
Outgroup Negative	2.89	0.21

The results indicated that in-group positive tweets ($M = 3.15, SD = 0.61$) are significantly more abstract than out-group positive tweets ($M = 2.90, SD = 0.98$), $t = 4.45$, $p < .001$, with Cohen's $d = 0.40$ indicating medium effect size. Similarly, the t-test also indicates that the out-group negative tweets ($M = 3.05, SD = 0.47$) are significantly more abstract than the in-group negative tweets ($M = 2.62, SD = 0.54$), $t = 3.68$, $p = .001$, with Cohen's $d = 0.90$ indicating a large effect size. The t values suggest a significant difference between the abstract scores of the in-group positive and out-group positive samples, and the same for the out-group negative and in-group negative samples, which is in line with the LIB phenomenon.

To further investigate how abstraction manifests in statements of different sentiments and group memberships, two-way ANOVA was conducted to test the effects of the abstract score on intergroup sentiment. The non-neutral group-sentiment interaction $F(2, 1233) = 20.17$, $p < .001$ indicates that there is a significant interaction effect between group membership and sentiment on the dependent variable abstract score. The effect of sentiment shows no statistical significance, while that of memberships cannot be interpreted independently. If neutrals are included, the group-sentiment interaction $F(3, 2199) = 20.16$, $p < .001$ is significant, though less than the main effect of sentiment $F(3, 2199) = 27.93$, $p < .001$, showing that there is a clear difference in abstraction usage for neutral and emotive utterances.

Linguistic Intergroup Bias in News

In this section, we examine the LIB phenomenon in the selected POLUSA data. We applied our model to label the sentences in each news article, and computed the abstract score on the article level. Sentences that are invalid to be LCM-coded, i.e., consist solely of O labels, are excluded when calculating the abstract score. Table 4 shows the means and standard deviations of the abstract scores.

Similar to the previous experiment, we conduct the two-sample t-test on the abstract scores of different group-sentiment samples. The result of in-group and out-group negative, $t = 3.23$, $p = 0.001$, with Cohen's $d = 0.35$ indicating a small to medium effect size, revealed that out-group negative reports ($M = 2.89, SD = 0.21$) were significantly more abstract than in-group negative ($M = 2.81, SD = 0.26$) reports. Additionally, the t-test between the non-negative news reports of in-group and out-group yields no statistical significance, which is likely because the majority of non-negative reports are neutral, and thus they all tend to be more concrete. This can be observed by comparing the average abstract scores of negative samples and non-negative samples, i.e., $M = 2.85, SD = 0.24$ versus $M = 2.77, SD = 0.18$.

To further illustrate the difference in abstraction level in negative and non-negative reportings from in-group and out-group, we also conducted ANOVA on the abstract scores. The results indicate that the interaction effect between group-membership and sentiment is significant on abstract scores, i.e., $F(1, 768) = 4.39$, $p = 0.04$, suggesting that the LIB phenomenon is observable in news reporting. Further, ANOVA on the right-leaning news articles shows statistical significance of the main effect of group membership on abstraction, i.e., $F(1, 370) = 4.10$, $p < 0.001$, with out-group reporting significantly more abstract than in-group reporting, while that for the left-leaning news show no particular significance, suggesting that the right reports Democrats more abstractly regardless of sentiment, but not the other way around.

Case Study

In this section, we conduct a case study to investigate language abstraction beyond the scores. Table 5 presents sentences labeled by our abstraction span extraction model.

The 1st sentence highlights the model's robustness in handling informal language, as it successfully identifies the hashtag *#TureAmericanHero* used as a predicative noun. The 2nd and 3rd sentences illustrate that the abstract score, while effective in large-scale quantitative analysis, might miss some nuances. Although the two sentences receive the same abstract score, the latter clearly paints a more abstract picture. The span extraction task setting could serve as a jumping board for more refined quantitative measures in the future. It is also worth noting that *A true public servant* may be labeled as a NOUN, since it is not grammatically complete and qualifies the main subject. It shows that training data ought to be expanded to include several sentences for better generalizability. The 4th to 7th sentences are from the reporting of the right- and left-leaning outlets on both political parties. Although all sentences are correctly labeled according to the LCM manual, some extracted spans do not account for the subjects of interest, e.g., *"rough"* in the 4th sentence, *"American"* in the 6th sentence, and the main bias that contributed to the score of the 5th sentence is aimed towards immigrants, i.e., *mass legal and illegal immigration*. If we exclude such spans when calculating the abstract scores, we would get an even starker contrast between in-group and out-group negative statements. As such, it is necessary for the model to output text spans instead of a single abstract score for human-in-the-loop bias analysis, depending on the analytic context.

Table 5: Case study. Score is the abstract score. Type indicates the group-sentiment category, where 'in' is in-group, 'out' is out-group, 'pos' is positive, 'neg' is negative.

Content	Score	Type
[Enjoying]$_{SV}$ [listening]$_{DAV}$ to my friend and colleague @repjohnlewis [speak]$_{DAV}$ at the #MOW50 – He is a [#TrueAmericanHero]$_{NOUN}$ #MarchOnWashington	2.50	In-Pos
Congratulations Rep. @john_dingell on becoming [longest-serving]$_{ADJ}$ member of U.S. Congress ever	4.00	Out-Pos
Congratulations to @john_dingell, my [beloved]$_{ADJ}$ colleague who today becomes the [longest-serving]$_{ADJ}$ Member in House history. A [true]$_{ADJ}$ public servant.	4.00	In-Pos
Republicans are [floundering]$_{IAV}$ on health care in the Obamacare age, but [looking to]$_{DAV}$ government to [solve]$_{IAV}$ the [rough]$_{ADJ}$ edges that Obamacare [created]$_{IAV}$ is [playing]$_{IAV}$ right into liberals' hands.	2.16	In-Neg
Democrats have [increasingly]$_{ADJ}$ [swept]$_{IAV}$ congressional districts whose demographics have been [quickly]$_{ADJ}$ [changed]$_{IAV}$ due to the country's policy of [mass]$_{ADJ}$ [legal]$_{ADJ}$ and [illegal]$_{ADJ}$ immigration.	3.42	Out-Neg
Democrats '[lied]$_{IAV}$ to the [American]$_{ADJ}$ people' over Mueller probe, now have to [answer]$_{IAV}$ to [American]$_{ADJ}$ people: Chaffetz.	2.50	In-Neg
[Vulnerable]$_{ADJ}$ Republicans [seek]$_{IAV}$ distance from Trump in [new]$_{ADJ}$ Congress.	3.33	Out-Neg

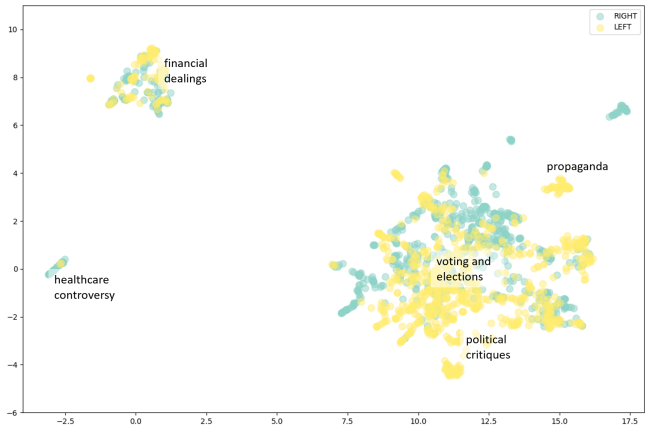

Figure 3: Clustering of the top 30% most abstract sentences from out-group reportings in POLUSA.

Visualization Analysis

In this section, we employed UMAP (McInnes et al., 2018) and HDBSCAN (McInnes et al., 2017) on RoBERTa-base sentence embeddings to investigate what kind of bias commonly appeared in the U.S. left and right discourse using the POLUSA dataset. Following the intuition that the more abstract a statement, the more it evokes bias (Maass, 1999), we clustered the sentences with top 30% abstract scores from out- and in-group reportings.

In Figure 3, the largest cluster is about *voting and elections*, particularly, both the left- and right-leaning media were using biased language to discuss the opposing party's actions and strategies for the 2020 election. Another noteworthy cluster is *financial dealings*, which contains mostly financial scandals and thus involves plenty of moral judgment. The *political critiques* and *propaganda* clusters mainly consist of the left's reporting, which criticizes the policies and media sway of the

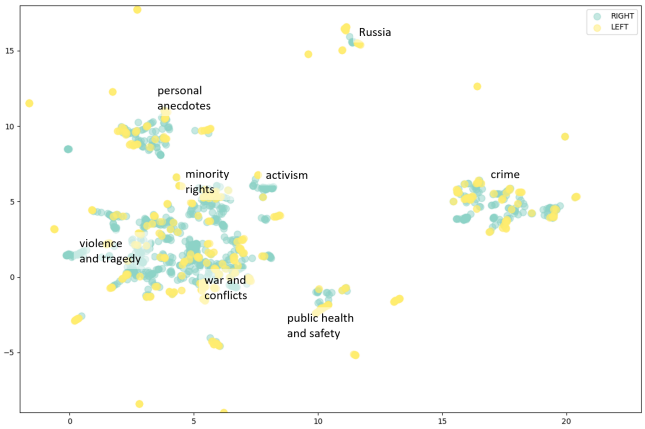

Figure 4: Clustering of the top 30% most abstract sentences from in-group reportings in POLUSA.

right. The *healthcare controversy* cluster, mostly from the right discusses polarizing topics such as abortion and racial health disparities. From Figure 4, we can observe that the high abstraction topics from in-group reporting are not actually directed towards the in-group party, but foreign powers (*Russia* and *war and conflicts*), convicted or perceived criminals (*crime* and *violence and tragedy*), and minority groups (*activism* and *minority rights*), which are indeed bias-riddled aspects. On the other hand, sentences in the *personal anecdotes* cluster receive high abstract scores mainly because they are statements about internal emotions and struggles, and thus are inherently biased by the definition of LIB.

Discussion

From the t-test results on the IGR dataset, we confirm that in-group positive and out-group negative tweets tend to be more abstract than the out-group positive and in-group negative ones. Similarly, results on the POLUSA dataset indicate that out-group negative news articles use more abstract language than in-group negative ones. ANOVA results on both datasets show that sentiments do not affect the abstraction level of utterances by themselves, but their interaction with group memberships has a significant effect.

Furthermore, when neutral samples from the IGR dataset were included, the ANOVA results suggest that they primarily drive the main effect of sentiment, i.e., they significantly differ from the positive and negative samples, which aligns with the LIB intuition that neutral utterances tend to be more concrete. Additionally, ANOVA on the different leanings from POLUSA suggests that the right-leaning outlets tend to depict Democrats more stereotypically, since a higher degree of abstraction across sentiment is a sign of LIB driven by differential expectancy (Maass, 1999).

For future work, we plan to improve our model, and explore how the proposed task can be used for quantitative bias analysis in varied domains, e.g., aggregating and analyzing the various stereotypes associated with a particular social group.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL). The work is also supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (MOE-T2EP20123-0005).

References

Cambria, E. (2024). Understanding natural language understanding. *Springer, ISBN 978-3-031-73973-6*.

Cambria, E., Mao, R., Hussain, A., Oatley, K., & Hinton, G. (2026). Artificial intelligence as the fourth decentering revolution: From cosmic, biological, and psychological displacement to cognitive decentering. *Cognitive Computation*, *18*(20), 1–13. https://doi.org/10.1007/s12559-026-10569-8

Cambria, E., Mao, R., Zhang, X., Xiao, L., Shen, T., & Anand, A. (2026). SenticNet 9: Generative commonsense for emotion AI via conceptual primitive discovery and time shift mechanism. *IEEE Transactions on Computational Social Systems*, *13*, 10.1109/TCSS.2026.3677113.

Carbonell, J., & Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 335–336.

Carnaghi, A., Maass, A., Gresta, S., Bianchi, M., Cadinu, M., & Arcuri, L. (2008). Nomina sunt omina: On the inductive potential of nouns and adjectives in person perception. *Journal of personality and social psychology*, *94*(5), 839.

Coenen, L. H., Hedebouw, L., & Semin, G. R. (2006). Measuring language abstraction: The linguistic category model (lcm).

Dong, W., Zhunis, A., Chin, H., Han, J., & Cha, M. (2024). I am not them: Fluid identities and persistent out-group bias in large language models. *arXiv preprint arXiv:2402.10436*.

Douglas, K. M., Sutton, R. M., & Wilkin, K. (2008). Could you mind your language? an investigation of communicators' ability to inhibit linguistic bias. *Journal of Language and Social Psychology*, *27*(2), 123–139.

Dragojevic, M., Sink, A., & Mastro, D. (2017). Evidence of linguistic intergroup bias in us print news coverage of immigration. *Journal of Language and Social Psychology*, *36*(4), 462–472.

Fiedler, K., Semin, G. R., & Finkenauer, C. (1993). The battle of words between gender groups a language-based approach to intergroup processes. *Human Communication Research*, *19*, 409–441.

Franco, F. M., & Maass, A. (1996). Implicit versus explicit strategies of out-group discrimination: The role of intentional control in biased language use and reward allocation. *Journal of Language and Social Psychology*, *15*(3), 335–359.

Gebhard, L., & Hamborg, F. (2020). The polusa dataset: 0.9 m political news articles balanced by time and outlet popularity. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 467–468.

Geschke, D., Sassenberg, K., Ruhrmann, G., & Sommer, D. (2010). Effects of linguistic abstractness in the mass media. *Journal of Media Psychology*.

Govindarajan, V., Zang, M., Mahowald, K., Beaver, D., & Li, J. J. (2024). Do they mean 'us'? Interpreting referring expression variation under intergroup bias. *Findings of EMNLP*, 9772–9785.

Govindarajan, V. S., Atwell, K., Sinno, B., Alikhani, M., Beaver, D. I., & Li, J. J. (2023). How people talk about each other: Modeling generalized intergroup bias and emotion. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2496–2506.

Graf, S., Bilewicz, M., Finell, E., & Geschke, D. (2013). Nouns cut slices: Effects of linguistic forms on intergroup bias. *Journal of Language and Social Psychology*, *32*(1), 62–83.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, *102*, 4.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Johnson-Grey, K. M., Boghrati, R., Wakslak, C. J., & Dehghani, M. (2020). Measuring abstract mind-sets through syntax: Automating the linguistic category model. *Social Psychological and Personality Science*, *11*(2), 217–225.

Kumar, S., Kholkar, G., Mendke, S., Sadana, A., Agrawal, P., & Dandapat, S. (2024). Socio-culturally aware evaluation framework for llm-based content moderation. *arXiv preprint arXiv:2412.13578*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Loshchilov, I. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Maass, A. (1999). Linguistic intergroup bias: Stereotype perpetuation through language. In *Advances in experimental social psychology* (pp. 79–121, Vol. 31). Elsevier.

Maass, A., Ceccarelli, R., & Rudin, S. (1996). Linguistic intergroup bias: Evidence for in-group-protective motivation. *Journal of Personality and Social Psychology*, *71*(3), 512.

Mao, R., Chen, G., Li, X., Ge, M., & Cambria, E. (2025). A comparative analysis of metaphorical cognition in ChatGPT and human minds. *Cognitive Computation*, *17*(35), 1–12. https://doi.org/10.1007/s12559-024-10393-y

Mao, R., Chen, G., Zhang, X., Guerin, F., & Cambria, E. (2024). GPTEval: A survey on assessments of ChatGPT and GPT-4. *Proceedings of LREC-COLING 2024*, 7844–7866.

Mao, R., He, K., Zhang, X., Chen, G., Ni, J., Yang, Z., & Cambria, E. (2024). A survey on semantic processing techniques. *Information Fusion*, *101*, 101988.

Mao, R., Liu, Q., He, K., Li, W., & Cambria, E. (2023). The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection. *IEEE Transactions on Affective Computing*, *14*(3), 1743–1753.

McInnes, L., Healy, J., Astels, S., et al. (2017). Hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Muchnik, L., Aral, S., & Taylor, S. J. (2013). Social influence bias: A randomized experiment. *Science*, *341*, 647–651.

Nadeem, M., Sohail, S. S., Cambria, E., Schuller, B. W., & Hussain, A. (2025). Gender bias in text-to-video generation models: A case study of Sora. *IEEE Intelligent Systems*, *40*(3), 10–15.

Pair, E., Vicas, N., Weber, A. M., Meausoone, V., Zou, J., Njuguna, A., & Darmstadt, G. L. (2021). Quantification of gender bias and sentiment toward political leaders over 20 years of kenyan news using natural language processing. *Frontiers in Psychology*, *12*, 712646.

Pan, Z., Peng, K., Ling, S., & Zhang, H. (2023). For the underrepresented in gender bias research: Chinese name gender prediction with heterogeneous graph attention network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(12), 14436–14443.

Ramshaw, L. A., & Marcus, M. P. (1999). Text chunking using transformation-based learning. In *Natural language processing using very large corpora* (pp. 157–176). Springer.

Raza, S., Garg, M., Reji, D. J., Bashir, S. R., & Ding, C. (2024). Nbias: A natural language processing framework for bias identification in text. *Expert Systems with Applications*, *237*, 121542.

Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. *Association for Computational Linguistics*.

Schmid, J. (1999). Pinning down attributions: The linguistic category model applied to wrestling reports. *European Journal of Social Psychology*, *29*(7), 895–907.

Schmid, J., & Fiedler, K. (1998). The backbone of closing speeches: The impact of prosecution versus defense language on judicial attributions 1. *Journal of Applied Social Psychology*, *28*, 1140–1172.

Seih, Y.-T., Beier, S., & Pennebaker, J. W. (2017). Development and examination of the linguistic category model in a computerized text analysis method. *Journal of Language and Social Psychology*, *36*(3), 343–355.

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, *54*(4), 558.

Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range. *European review of social psychology*, *2*(1), 1–30.

Snefjella, B., & Kuperman, V. (2015). Concreteness and psychological distance in natural language use. *Psychological science*, *26*, 1449.

Spinde, T., Plank, M., Krieger, J.-D., Ruas, T., Gipp, B., & Aizawa, A. (2021). Neural media bias detection using distant supervision with babe-bias annotations by experts. *Findings of EMNLP*, 1166–1177.

Van Dijk, T. (2009). *Society and discourse: How social contexts influence text and talk*. Cambridge University Press.

Yu, Z., Sen, I., Assenmacher, D., Samory, M., Fröhling, L., Dahn, C., Nozza, D., & Wagner, C. (2024). The unseen targets of hate: A systematic review of hateful communication datasets. *Social Science Computer Review*, 08944393241258771.

Zhang, X., Mao, R., & Cambria, E. (2023). A survey on syntactic processing techniques. *Artificial Intelligence Review*, *56*, 5645–5728.