Counterfactual Probing for the Influence of Affect and Specificity on Intergroup Bias

Venkata S Govindarajan **Kyle Mahowald** **David I. Beaver** **Junyi Jessy Li**
Department of Linguistics
The University of Texas at Austin
{venkatasg, mahowald, dib, jessy}@utexas.edu

Abstract

While existing work on studying bias in NLP focues on negative or pejorative language use, Govindarajan et al. (2023) offer a revised framing of bias in terms of intergroup social context, and its effects on language behavior. In this paper, we investigate if two pragmatic features (specificity and affect) systematically vary in different intergroup contexts — thus connecting this new framing of bias to language output. Preliminary analysis finds modest correlations between specificity and affect of tweets with supervised intergroup relationship (IGR) labels. Counterfactual probing further reveals that while neural models finetuned for predicting IGR reliably use affect in classification, the model's usage of specificity is inconclusive.

1 Introduction

Most work on bias in NLP only considers negative or pejorative language use (Kaneko and Bollegala, 2019; Sheng et al., 2019; Webson et al., 2020; Pryzant et al., 2020; Sheng et al., 2020). While recent work has delved into implicit bias (Rashkin et al., 2015; Sap et al., 2017, 2020), they are still limited as they rely on identifying specific demographic dimensions or an individual's intent. Crucially, language production is still taken to be 'unbiased' by default. Research in social psychology suggests a different framing of bias that encompasses all language use — we can analyze bias as changes in (language) behavior reflecting shifting social dynamics (Van Dijk, 2009). Under this view, all the language we produce is biased, with the nature of the bias determined by the social relationships between the speaker and target. Inspired by this idea, Govindarajan et al. (2023) proposed a new framing of bias by modeling intergroup relationships (IGR, in-group and out-group) in interpersonal English language tweets, potentially capturing more subtle forms of bias. This framing raises a question: *which linguistic features vary systematically in different intergroup contexts?*

The Linguistic Intergroup Bias (LIB; Maass et al., 1989; Maass, 1999) hypothesis offers some clues towards linguistic features that change with shifting intergroup contexts. LIB speculates that socially desirable in-group behaviors and socially undesirable out-group behaviors are encoded at a higher level of abstraction. The theory however relies on a restricted definition of abstractness that relies solely on predicates, and an ad-hoc analysis of 'social desirability' that doesn't permit large-scale analysis. We can do better by using two well-defined pragmatic features: **specificity** (Li, 2017) is a pragmatic feature of text that measures the level of detail (similar to abstract–concrete axis), while **affect** is a feature that measures the attitude of a speaker towards their target (Sheng et al., 2019) in an utterance (analogous to social desirability).

Specificity and affect are analogous to the LIB axes of language variation that are easy to annotate and compute. Furthermore, specificity is a more *general* property than abstractness in the LIB — specificity is a property of the whole sentence rather than just the predicate. Thus, our study focuses on **intergroup bias** more generally, rather than the narrow parameterization of the LIB. Similar to the LIB, our formulation of intergroup bias predicts that positive affect in-group utterances and negative affect out-group utterances are encoded with *lower specificity* (i.e. more generally). Tables 1 and 2 compare the predicted language variation between the LIB and our formulation.

In this work, we perform the first large-scale study of linguistic differences in intergroup bias by analyzing its nature in the corpus of English tweets from Govindarajan et al. (2023), which makes use of naturally occurring labels for in-group vs. out-group. This distinguishes us from existing work in LIB which mostly relies on artificial responses from participants in studies, rather than natural language use in the wild. To bolster our probing investigation, we also explore it causally: exploiting

Findings of the Association for Computational Linguistics: ACL 2023, pages 12853–12862
July 9-14, 2023 ©2023 Association for Computational Linguistics

	In-group	Out-group
socially desirable	abstract	concrete
socially undesirable	concrete	abstract

Table 1: Predicted language variation in the LIB.

	In-group	Out-group
positive affect	low specificity	high specificity
negative affect	high specificity	low specificity

Table 2: Predicted language variation in our more general formulation, using specificity and affect

the quantitative nature of our formulation to study if a neural model finetuned for IGR prediction uses pragmatic features such as specificity and affect in its decision-making process through counterfactual probing techniques (Ravfogel et al., 2021).

To summarize our findings, we find a modest positive correlation between affect and IGR in our data, with a positive causation effect as well — making a tweet's affect more positive makes it more likely to be in-group regardless of its specificity. We find no correlation between specificity and IGR in our data. Surprisingly, we discover a causal effect of low specificity on IGR prediction that is uniform across affect, but none for high specificity. We hypothesize that this could be because of damage to the underlying language model, but we leave further investigation to future work. We release our code and data at `github.com/venkatasg/intergroup-probing`.

2 Background

Intergroup bias The Linguistic Intergroup Bias (LIB) theory (Maass et al., 1989; Maass, 1999) tries to explain how stereotypes are transmitted and persist in communication by hypothesizing that socially desirable in-group behaviors and socially undesirable out-group behaviors are encoded at a higher level of abstraction . The LIB has been reproduced in various psychological experiments and analyses (Anolli et al., 2006; Gorham, 2006); it has also been used as an indicator for a speaker's prejudicial attitudes (Hippel et al., 1997), and racism (Schnake and Ruscher, 1998).

Table 1 describes the LIB asymmetry and the parameters used. As stated earlier, the LIB relies on ad-hoc and hand-coded concepts such as 'social desirability' and abstractness of predicates (Semin and Fiedler, 1988) . Our proposed experiments *generalize* beyond the LIB by utilizing parameters that are easily computable, and are a function of the whole utterance. We also build upon the dataset and work in Govindarajan et al. (2023), which is the first large-scale analysis of intergroup bias on naturally occurring speech.

Specificity Specificity is a pragmatic concept of text that measures the level of detail and involvement of concepts, objects and events. Louis and Nenkova (2011) introduced the first dataset and model for sentence specificity prediction, and in later work Li (2017) illustrated the role of specificity in discourse coherence. Furthermore, Gao et al. (2019) expanded the scope of specificity analysis from the news domain to social media.

Affect There is a wealth of work studying emotions and sentiment in social media text (Mohammad, 2012; Wang et al., 2012; Mohammad and Kiritchenko, 2015; Abdul-Mageed and Ungar, 2017; Desai et al., 2020; Demszky et al., 2020). Govindarajan et al. (2023) introduced the first dataset annotated for *interpersonal* emotion (defined as only emotions expressed towards or in connection with a target), using the Plutchik wheel (Plutchik, 1980, 2001) as a framework. While fine-grained, this approach isn't amenable to the experimentation we propose easily. Inspired by the concept of *regard* by a speaker towards a demographic in an utterance (Sheng et al., 2019), we introduce annotations for a coarse-grained feature we term *affect* that estimates how a speaker feels towards the target they mentioned in an interpersonal utterance.

Table 2 describes the intergroup language variation as hypothesized in our experimentation, using specificity and affect. Analogous to LIB, our hypothesis is that positive affect utterances directed at in-group individuals, and negative affect utterances directed at out-group individuals are encoded with *lower specificity*.

AlterRep AlterRep (Ravfogel et al., 2021) is a probing technique that tests if a neural network *uses* a property, rather than just testing if the model's learned representations correlate with the property. The method uses Iterative Nullspace Projection (INLP; Ravfogel et al., 2020) to iteratively train a linear classifier on the model's internal representations to pick out a particular feature, using the parameters learned by the classifier to intervene on the embedding and alter it in a systematic way. The AlterRep method based on INLP has been used

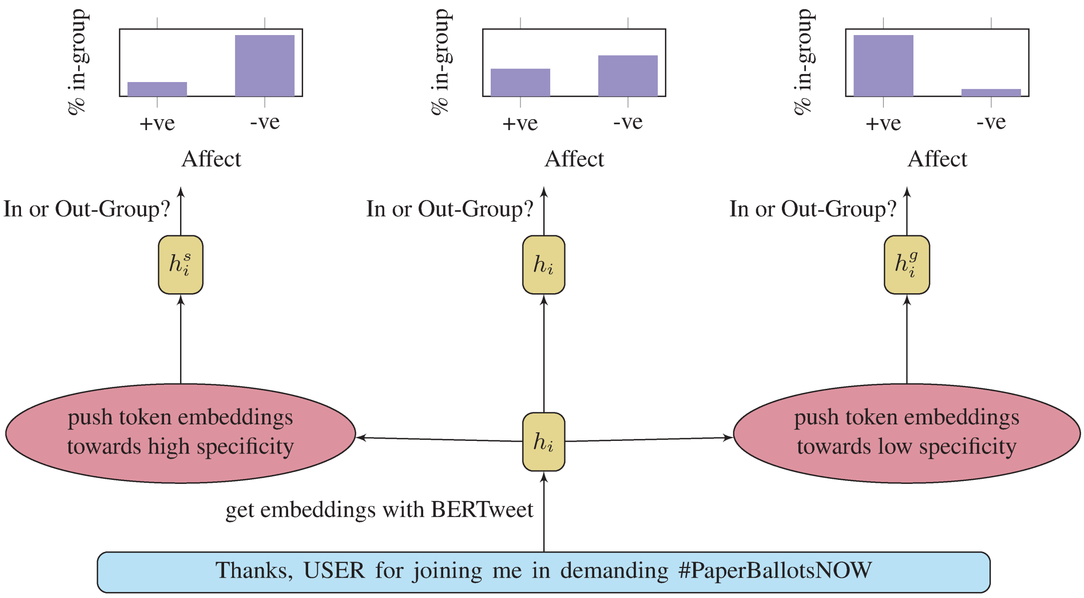

Figure 1: Flowchart describing the specificity intervention experiment and expected results.

to probe for syntactic phenomena such as subject-verb number agreement (Ravfogel et al., 2021). To our knowledge, ours is the first work probing if a model learns and uses higher-level pragmatic features like affect and specificity using AlterRep.

3 Data & Experiments

3.1 Data & Annotations

We use the same dataset of tweets from Govindarajan et al. (2023), which consists of tweets by members of US Congress that @-mention other members in the same tweet, with 'found-supervision' for the IGR labels of every tweet. A tweet is in-group if it is targeted at another member of the same party as the writer of the tweet, else it is out-group.

Affect We build upon the dataset's fine-grained annotations for interpersonal emotion by adding annotations for affect. We presented annotators on Mechanical Turk with tweets from our dataset with the target mention masked (with the placeholder Doe, to minimize potential biases of the annotator), and asked the following questions:

a. How does the writer feel in general about Doe? *warmly, coldly, neutral, mixed*
b. How does the writer feel in general about Doe's actions/behavior? *approval, disapproval, neutral, mixed*

Annotators are given the option to select one of the 4 options listed above for each question. For each tweet, we collect annotations from 3 annotators, obtaining an aggregate label for each question by majority vote. We report an inter-annotator agreement score (Fleiss's kappa; Fleiss, 1971) of 0.53 for the first question, and 0.56 for the second.

We derive a binary affect label (± 1) from our annotations using a simple rule: If the writer of a tweet is deemed to either feel warmly towards the target, or if they approve of the target's actions, the affect is set to be positive; else it is set to be negative. An analysis of our collected annotations on the data shows that there is a small positive (Pearson's) correlation (r=0.2, $p < 0.001$) between binary affect and IGR.

Specificity Specificity of the tweets in the dataset are calculated using the specificity prediction tool from Gao et al. (2019). Their specificity predictor is trained on tweets, and uses surface lexical features, as well as syntactic, semantic and distributional features to calculate a specificity score between 1 and 5. We note that on our dataset, there was *no correlation between specificity and IGR* (r=-0.07, $p < 0.001$), unlike affect. On further inspection of our dataset, we find that tweets with very high/low specificity scores (gathered by excluding specificity scores between 3 and 4, similar to excluding the middle in Gelman and Park, 2009) have a small but statistically significant negative correlation with IGR labels (r=-0.13, $p < 0.001$).

3.2 Interventions

Model We use BERTweet (Nguyen et al., 2020), a language model pre-trained on 850M English tweets, the same model used in Govindarajan et al. (2023). All intervention experiments are carried out with the best performing *finetuned* version of this model — where the model is finetuned on the task of predicting IGR labels. The input to the model is only the tweet with no other context, and the target masked with a placeholder `@USER`.

We use the model's representations from layer 11 for the INLP procedure since it shows the most reliable effects. INLP (Ravfogel et al., 2020) works by learning a series of linear classifiers on the representations from an encoder. In each iteration, the embeddings are projected onto the intersection of nullspaces of the classifiers learned so far, meaning the information used by the existing classifiers is removed from the model. Every subsequent classifier we learn removes more information of the property of interest from the model's representations. We find that higher layers offer a good balance between feature extractability and language model stability (see Appendix D) for our features.

After training INLP, AlterRep uses the classifier's decision space to project model embeddings into a null component that contains no information from the feature of interest, and an orthogonal component, that contains all the information from the feature of interest. These two components thus enable us to perform the counterfactual intervention — pushing model embeddings towards having more, or less, of a particular property. When AlterRep uses INLP classifiers with more iterations, the strength of the intervention is greater. Figure 1 offers an illustration of our intervention experiment on specificity, and the expected results.

Affect Using the binary affect labels we derived from annotations that we described in § 3.1, we perform interventions to test if the model uses affect causally in its decision. We sample 3 tokens at random from each sentence in the training and validation split of our dataset, train an iterative linear classifier on the model's representations of these tokens using INLP (against the affect label of the tweet), and use the decision boundary learned by the classifier to intervene by pushing model representations to have more positive affect or have more negative affect. We set the hyperparameter α in AlterRep to 4.

Specificity The INLP classifier for specificity is learned using the same procedure as for affect. We train the classifier on only the tweets with high and low specificity scores in our dataset (scores below 3 and above 4; scores taken from the specificity prediction tool in Gao et al. (2019)), excluding the middle to ensure effective learning of the decision boundary (Gelman and Park, 2009). Thus, we are effectively pushing the model representations to have high or low specificity. For both affect and specificity, once the INLP classifier is learned, we perform the intervention on a random subset of 30% of the tokens of a tweet (to control for tweet length). We also report the results of random interventions as a control, where random interventions are generated by sampling from a standard gaussian instead of using the decision matrix generated by INLP.

Hypotheses We report the percentage of tweets in the test split of our dataset that are predicted to be in-group by our classifier model with increasing strength of the intervention (number of INLP iterations, 0 being pre-intervention). Thus, we have the following hypotheses on the effects of our intervention on the data based on our intergroup bias framework described in Table 2:

1. Interventions towards positive affect should induce the model to predict low specificity tweets to be in-group and high specificity tweets to be out-group, while interventions towards negative affect should affect the model conversely.
2. Interventions towards higher specificity should induce the model to predict positive affect tweets as out-group and negative affect tweets as in-group, while interventions towards lower specificity should affect the model conversely.

4 Results & Analysis

The results for the interventions on affect are presented in Figure 2, while those for specificity are presented in Figure 3. Overall, we observe that in both cases, interventions had the same effect on tweets that were annotated with positive affect as they did on tweets with negative affect (and similarly for tweets with high and low specificity) — so we only show the percentage of *all* tweets in the test split classified as in-group.

Affect As Figure 2 shows, pushing model representations towards having more positive affect causes almost all tweets in the test split of our data to be classified as in-group after 32 iterations of INLP. The randomness after 40 iterations of INLP could be attributed to the underlying RoBERTa

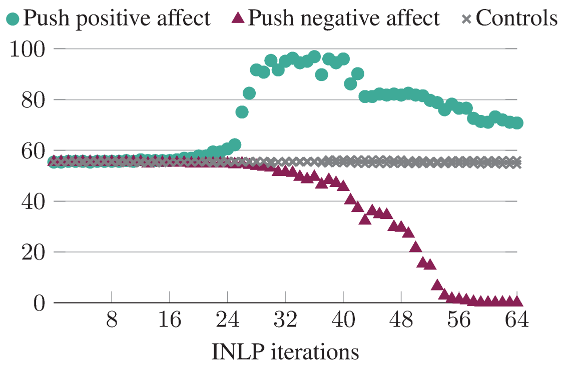

Figure 2: % of test set classified as in-group plotted against number of INLP interventions for affect.

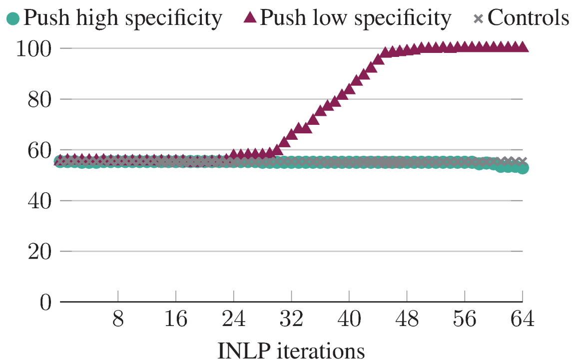

Figure 3: % of test set classified as in-group plotted against number of INLP interventions for specificity.

language model being destroyed, as the LM Top-100 accuracy plot in Appendix D shows. Pushing the model's representations towards negative affect shows the inverse effect as expected, although the nature of the drop appears different. We hypothesize that this is because most of the tweets in our dataset (75.2%) have positive affect. An intervention pushing the representations towards negative affect would be slower and require stronger intervention forces, which is borne out in Figure 2.

Specificity Figure 3 shows that pushing model representations towards being more specific has no effect on model behavior and is indistinguishable from the control; but pushing towards lower specificity has a noticeable effect — interventions after 48 iterations of iNLP lead to all the data being predicted as in-group. Our hypothesis states that general language is more likely in positive affect in-group contexts; however we find no difference in the model's behavior on positive versus negative affect tweets as reported earlier.

Overall our findings indicate that while the model does use affect towards making its decision on the interpersonal group relationship prediction task (albeit uniformly across specificity), it doesn't use specificity as we had predicted. The discrepancy between high and low specificity interventions could be because the average specificity of tweets in our training data is 3.49 ($\sigma = 0.54$) — meaning that interventions towards lower specificity act in opposition to most of our data in representation space. But these results requires further investigation to understand them better.

Qualitative error analysis Digging into the results further, we wanted to investigate if the interventions function the way we wanted them to. We analyzed the tokens that the model predicts before and after intervention for example (1). Firstly, finetuning the model for IGR prediction leads to degradation in LM abilities — a vanilla model predicts *birthday, anniversary* for the masked token in (1), but the finetuned model predicts nonsensical tokens like *sworn, opport__* even before any interventions. Pushing towards negative affect causes it to predict tokens with negative connotations (*killing, ass, opposition*), but degrades the underlying LM even further. The specificity interventions are especially hard to interpret due to the semantically and syntactically implausible tokens being selected (*opport__, mug__, ask__*)

(1) Happy <mask> @USER! I got you a new bill: #IIOA

While some of the interventions push the model's predictions to be in the general lexical space desired (which probably explain the affect intervention results), the lack of contextual fit due to LM degradation may explain the inconclusive results, and lack of interaction between affect and specificity.

5 Conclusion

Studying bias in language use through an interpersonal lens opens up new questions, such as which linguistic features vary systematically in changing interpersonal contexts. We perform a correlational and causal analysis of two pragmatic features, specificity and affect, on a dataset of interpersonal tweets in English, to establish how they influence intergroup relationship prediction. We find modest correlations between our features and IGR labels, while counterfactual probing reveals mixed results. Affect influences IGR prediction causally but without interacting with specificity, while specificity only influences IGR prediction in one direction.

Limitations

Future work must look into the generalizability of the results presented here in other domains of language use, and other languages. While we present the utterances as constituting natural speech by one speaker (the congressperson who sent the tweet), it is likely most congresspeople employ social media teams that help in crafting the language of some of their tweets. However, we believe for the sake of interpersonal group membership, the relationship between the speaker(or speakers) and their target(s) would not be affected.

Techniques like INLP extract information that is linearly extractable. While we've shown that it is possible to extract and manipulate language information using such simple linear techniques, more complex methods like those proposed by Ravfogel et al. (2022) might be able to manipulate more non-linearly encoded properties.

The AlterRep procedure, as can be seen in our results and in Ravfogel et al. (2021), is sensitive to parameters like α and the number of INLP iterations. Picking these parameters is tricky and we have done it in a manner that preserves information in the language model. It is possible that a different set of settings not explored here could lead to different results.

Ethics Statement

For the corpus of tweets on which we performed annotations, we downloaded the tweets using the official Twitter API. In accordance with the Twitter Terms of Service, we release tweet IDs and usernames, but not the tweet text itself. Our dataset was built through crowdsourced annotations on Amazon Mechanical Turk. To ensure annotators were paid a fair wage of at least $10 an hour, we paid annotators $0.50 per HIT. Each HIT involved annotating 3 tweets, which we estimate to take on average 3 minutes to complete.

Acknowledgements

This research is partially supported by Good Systems[1], a UT Austin Grand Challenge to develop responsible AI technologies, and NSF grants IIS-2145479, IIS-2107524. We acknowledge the Texas Advanced Computing Center (TACC)[2] at UT Austin and AWS for many of the results within this paper. Kyle Mahowald was funded in part by NSF Grant 2104995.

[1] https://goodsystems.utexas.edu

[2] https://www.tacc.utexas.edu

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

Luigi Anolli, Valentino Zurloni, and Giuseppe Riva. 2006. Linguistic intergroup bias in political communication. *The Journal of General Psychology*, 133:237 – 255.

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, Online. Association for Computational Linguistics.

Shrey Desai, Cornelia Caragea, and Junyi Jessy Li. 2020. Detecting perceived emotions in hurricane disasters. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5290–5305, Online. Association for Computational Linguistics.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Yifan Gao, Yang Zhong, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2019. Predicting and analyzing language specificity in social media posts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1):6415–6422. Number: 01.

Andrew Gelman and David K. Park. 2009. Splitting a predictor at the upper quarter or third and the lower quarter or third. *The American Statistician*, 63(1):1–8.

Bradley W. Gorham. 2006. News media's relationship with stereotyping: The linguistic intergroup bias in response to crime news. *Journal of Communication*, 56(2):289–308. Place: United Kingdom Publisher: Blackwell Publishing.

Venkata Subrahmanyan Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David Beaver, and Junyi Jessy Li. 2023. How people talk about each other: Modeling generalized intergroup bias and emotion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2488–2498, Dubrovnik, Croatia. Association for Computational Linguistics.

W. Hippel, Denise Sekaquaptewa, and P. Vargas. 1997. The linguistic intergroup bias as an implicit indicator of prejudice. *Journal of Experimental Social Psychology*, 33:490–509.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Junyi Jessy Li. 2017. *From Discourse Structure To Text Specificity: Studies Of Coherence Preferences*. phdthesis, University of Pennsylvania.

Annie Louis and Ani Nenkova. 2011. Text specificity and impact on quality of news summaries. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 34–42, Portland, Oregon. Association for Computational Linguistics.

Anne Maass. 1999. Linguistic intergroup bias: Stereotype perpetuation through language. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 31, pages 79–121. Academic Press.

Anne Maass, Daniel Anthony Salvi, Luciano Arcuri, and Gün R. Semin. 1989. Language use in intergroup contexts: the linguistic intergroup bias. *Journal of personality and social psychology*, 57 6:981–93.

Saif Mohammad. 2012. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255, Montréal, Canada. Association for Computational Linguistics.

Saif M. Mohammad and Svetlana Kiritchenko. 2015. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31:301 – 326.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for english tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. Association for Computational Linguistics.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Robert Plutchik. 2001. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist*, 89(4):344–350.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2015. Connotation frames: A data-driven investigation. *arXiv: Computation and Language*.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Grusha Prasad, Tal Linzen, and Yoav Goldberg. 2021. Counterfactual interventions reveal the causal effect of relative clause representations on agreement prediction. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 194–209, Online. Association for Computational Linguistics.

Shauli Ravfogel, Francisco Vargas, Yoav Goldberg, and Ryan Cotterell. 2022. Adversarial concept erasure in kernel space.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Conference on Empirical Methods in Natural Language Processing*.

Sherry B Schnake and Janet B Ruscher. 1998. Modern racism as a predictor of the linguistic intergroup bias. *Journal of Language and Social Psychology*, 17(4):484–491.

G. R. Semin and K. Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. 54:558–568. Publisher: American Psychological Association.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Affect	Train	Dev	Test
Positive	1813	226	242
Negative	589	80	83

Table 3: Distribution of affect in train-dev-test split

Teun A Van Dijk. 2009. *Society and discourse: How social contexts influence text and talk*. Cambridge University Press.

Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P. Sheth. 2012. Harnessing twitter "big data" for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 587–592.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "undocumented workers" the same as "illegal aliens"? Disentangling denotation and connotation in vector spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix

A Implementation

We use `bertweet-base` models from VinAI's Huggingface models repository, and the `transformers` package for all of our probing experiments (Wolf et al., 2020). Language ID classifiers were trained using `LinearSVC` classifier from `sklearn`. For training these classifiers, equal number of tokens from both labels were sampled. We used a batch size of 32, and a maximum sequence length of 128 when performing the intervention experiments. The interpersonal group relationship prediction model was reproduced from Govindarajan et al. (2023) using the same experimental settings and hyperparameters for the probing experiments.

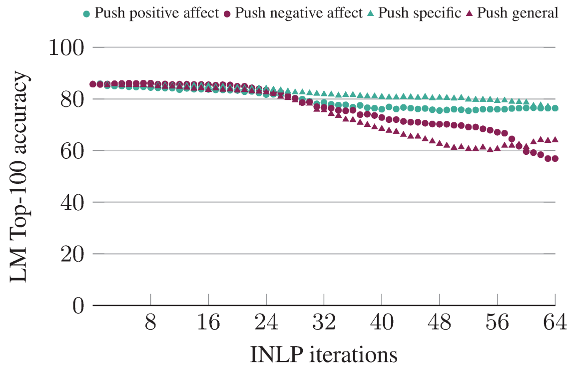

Figure 4: LM Accuracy on train plus validation split for different interventions

B Data & Annotation

To obtain reliable annotations, we prequalify annotators using a qualifying task. Annotators were recruited on Mechanical Turk using a qualifying task where they were asked to annotate 6 tweets using the schema detailed in§ 3.1. We restricted the qualification task to annotators living in the USA who had attempted at least 500 HITS and had a HIT approval rate $\geq 98\%$. After manual inspection, 6 anonymous annotators were qualified for bulk annotation. Each tweet was annotated by three different annotators. To ensure annotators were paid a fair wage of at least 10$ an hour, we paid annotators $0.50 per HIT. Each HIT involved annotating 3 tweets, which we estimate to take on average 3 minutes to complete. In total, 3,033 tweets between 2010 and 2021 were annotated with perceived affect.

C Dataset Statistics

We present preliminary statistics for the annotations on the dataset of tweets in Table 3.

D LM Accuracy over INLP iterations

AlterRep directly alters the LM's representations, which inevitably harms the model's internal structure. Figure 4 shows the LM's top-100 accuracy at predicting randomly masked tokens on our dataset, proving that the interventions are meaningful while still maintaining the LM's integrity.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the Limitations and Ethics Section

☑ A2. Did you discuss any potential risks of your work?
In the Limitations and Ethics section

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 and Appendix A,B

☑ B1. Did you cite the creators of artifacts you used?
Section 3 and Appendix A, B

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 3 and Appendix B

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 3 and Appendix B

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 3 and Appendix B

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Appendix C

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No. We used standard off the shelf models, and didn't pre-train any new models. The total computational budget for fine-tuning for our analysis is very small.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3.2 and Appendix A

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3, 4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section 3 and Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix B

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Appendix B

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B