Debiasing Multi-Entity Aspect-Based Sentiment Analysis with Norm-Based Data Augmentation

Scott Friedman, Joan Zheng, Hillel Steinmetz
friedman@sift.net, jzheng@sift.net, hsteinmetz@sift.net
SIFT, Minneapolis, MN USA

Abstract

Bias in NLP models may arise from using pre-trained transformer models trained on biased corpora, or by training or fine-tuning directly on corpora with systemic biases. Recent research has explored strategies for reduce measurable biases in NLP predictions while maintaining prediction accuracy on held-out test sets, e.g., by modifying word embedding geometry after training, using purpose-built neural modules for training, or automatically augmenting training data with examples designed to reduce bias. This paper focuses on a debiasing strategy for aspect-based sentiment analysis (ABSA) by augmenting the training data using norm-based language templates derived from previous language resources. We show that the baseline model predicts lower sentiment toward some topics and individuals than others and has relatively high prediction bias (measured by standard deviation), even when the context is held constant. Our results show that our norm-based data augmentation reduces topical bias to less than half while maintaining prediction quality (measured by RMSE), by augmenting the training data by only 1.8%.

Keywords: sentiment, bias, ABSA, NLP

1. Introduction

Pre-trained transformer models are prevalent components in modern NLP architectures (Devlin et al., 2019). These transformers help encode the semantics of words and phrases in context, based on word and phrase meanings in massive corpora of training data. Recent research has shown that—along with the contextual meanings of these words—the training data also encodes systemic bias (Bolukbasi et al., 2016; Lu et al., 2020), where biases in language models can actually predict stereotypes over time (Garg et al., 2018) and gender gaps at international scale (Friedman et al., 2019).

Systemic bias or anti-social sentiments in training data—such as gender bias, racism, homophobia, antisemitism, Islamophobia, and more—can lead to sub-optimal NLP predictions. Encoding and preserving biases may be useful if we aim to model the stereotypes and world-view of a single group, but if we seek to build generally-applicable models or build models that promote equity then these systemic biases will diminish performance.

In addition to biased *pre*-training data, these same types of biases and stereotypes may exist in the data used for domain-level or task-level training or fine-tuning. For example, suppose we want to train a model to detect emotions, pro-social or anti-social language, hate speech (Zheng et al., 2022), toxicity (Hosseini et al., 2017), or moral framing (Hoover et al., 2020). If we train on purpose-built corpora from online forums that have cohesive views, e.g., advocating for some ideologies i^+ and opposing other ideologies i^-, then our model will understandably learn to ascribe higher sentiment toward i^+ than i^- since doing so will consistently decrease its error during training and evaluation. If we later want to apply this learned model to another forum that advocates for i^- and opposes i^+, the learned biases will reduce performance. Ultimately, we desire models that capture the linguistic principles but can be practically applied to other domains.

This paper characterizes the measurable problem of topical bias in multi-entity, multi-dimensional aspect-based sentiment analysis (ABSA) and describes an approach for data augmentation that empirically reduces bias without diminishing model performance on held-out test sets. We continue with an overview of our problem setting of ABSA and debiasing, including related work, and then we describe our dataset, our approach, and our empirical results.

2. ABSA and Moral Disengagement

Aspect-based sentiment analysis (ABSA) (Yang et al., 2018) captures multiple aspects (i.e., dimensions) of sentiment on a message, e.g., for assessing both the value and the food quality of a restaurant review, since these can vary independently. Building on top of ABSA is multi-entity ABSA (ME-ABSA) (Tao and Fang, 2020; Zheng et al., 2022), which classifies multiple aspects on a per-entity basis in a message, e.g., to capture a **positive**, **neutral**, or **negative** sentiment label along each dimension for each entity of interest.

Unlike previous work on categorical (positive/neutral/negative) ME-ABSA, this paper's ABSA model uses continuous dimensions of sentiment, where each dimension is an *intensity score* that ranges from -1 to 1, with 0 being neutral. Sentiment aspects are derived in part from Bandura's (1999; 2016) theory of moral disengagement. We focus on the following dimensions,

LREC-COLING 2024, pages 4456–4461

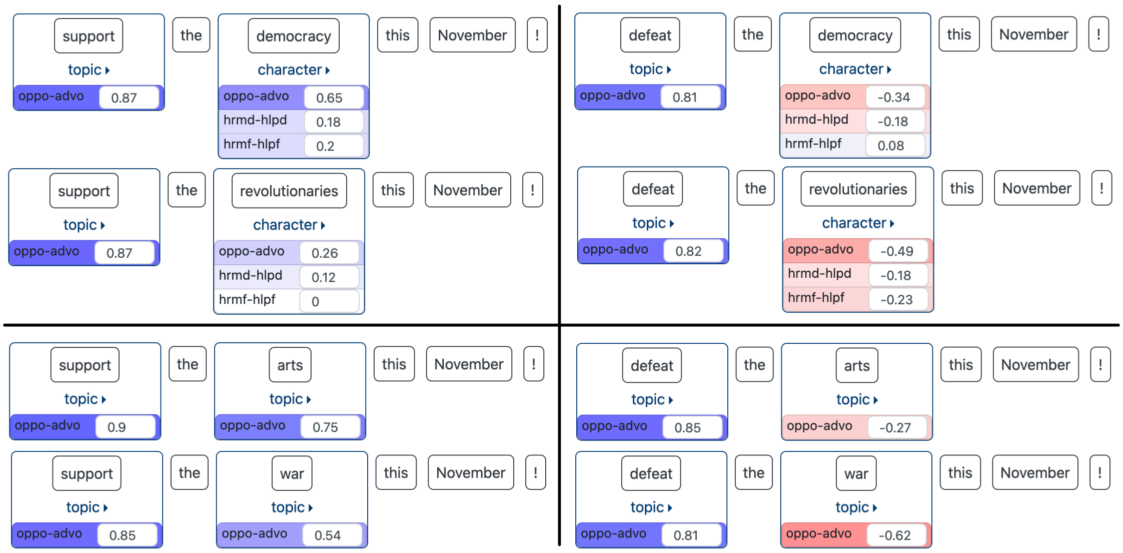

Figure 1: Multi-entity, multi-dimension ABSA results for eight different sentences using the same general text templates, advocating (left) or opposing (right) various characters/ideologies (top) or topics (bottom), where the sentiment values vary with the topic that is mentioned.

where each is an axis from an anti-social pole (-1) to a pro-social pole (1) that can be expressed over an entity or topic mentioned in a text:

- **Opposition—Advocacy**: The author disapproves or dislikes the topic (-1), or promotes and advances a topic (1). This aspect is the closest to a general sentiment measure in other work.
- **Harmed—Helped**: The author expresses that the entity is somehow harmed (-1) or benefited (1). Harmed is an indication of victimization in Bandura's theory, which may morally justify action on their behalf.
- **Harmful—Helpful**: The author expresses that the entity or topic brings harm to others (-1) or benefits others (1). Blame for harm may morally justify action against the party, per Bandura's theory.

Depending on the category of entity (from a NER or topic-detection model), the ME-ABSA model predicts all three aspects (for human-like *characters* that include pronouns, human roles, names, and ideologies), or only oppose-advocate (for non-human *topics* that cannot be victimized or morally blamed for harm in the Bandura sense).

To exemplify the ME-ABSA predictions, Figure 1 shows examples of ME-ABSA outputs for eight different parses using the same message using the template: "$\langle support, defeat\rangle$ the $\langle entity\rangle$ this November!" In the top two quadrants, $\langle entity\rangle$ is a character (i.e., human ideology, group, or individual). In the bottom two quadrants, $\langle entity\rangle$ is a non-human topic. For all outputs in all quadrants, the "support" and "defeat" are correctly predicted as being advocated buy the author (i.e., **oppo-advo** > 0). For all of the "support" instances (left quadrants), the model correctly predicts the advocacy of the subject, and for all "defeat" instances (right quadrants), the model correctly predicts opposition (i.e., **oppo-advo** < 0).

2.1. Sub-Optimal ABSA Bias

Despite being *qualitatively* accurate in the **oppo-advo** dimension, the results in Figure 1 include numerical discrepancies within each quadrant: "democracy" was predicted as more advocated and less opposed than "revolutionaries," in both upper quadrants, and "democracy" was also predicted as being more helpful (and "revolutionaries" as more harmful). In the bottom quadrants, "war" predicted as being more opposed than "arts" in both lower quadrants, despite the qualitative agreement.

We see that the *content* of the sentiment target (i.e., the word "war" versus "arts") impacts the ABSA judgment even when the surrounding *context* of the sentiment target (i.e., "support the _____ this November!") remains constant. This is likely due to a mixture of bias in the pre-trained transformer used in the architecture and additional bias learned from the labeled ABSA training data, where war is often opposed. This paper does not provide methods for pinpointing the source of the bias; rather, we focus on mitigating it.

3. Bias Mitigation

Previous work in NLP de-biasing techniques has focused on methods for instrumenting the training architecture (Attanasio et al., 2022), changing the geometry of word embeddings to neutralize biased components (Bolukbasi et al., 2016), or augmenting the training data in targeted fashions, e.g., using counter-factual data augmentation (CDA) to counter-factually reverse or perturb terms such as gendered pronouns (Lu et al., 2020). These each have unique benefits and constraints, e.g., for when you know the words or categories of bias you seek to mitigate (e.g., gender or racial bias).

Lu et al. (2020) showed that augmenting the training data using gender-targeted CDA outperformed word embedding modifications. This motivates this our data augmentation approach; however, unlike Lu et al., we are not targeting a single dimension of bias such as gender: as shown in Figure 1, the model may have a bias toward democracy and against warfare, in addition to gender, racial, and other biases. This makes a *single-dimension* CDA approach untenable for our open-ended ABSA.

Importantly, we do not want the model to disregard the entity's text *entirely*. For instance, when the ABSA target is "hero" or "terrorist" or "scumbag," the model should consider the helpful, harmful, and dehumanizing aspects of these terms since they carry sentiment in themselves. Masking the entity's text entirely would therefore reduce bias, but at the cost of model performance.

3.1. Norm-Based Data Augmentation

Our norm-based data augmentation approach generates templated examples that affirm sentiment norms, across a general set of topics. This norm-affirming approach differs from Lu et al.'s (2020) CDA approach that includes custom-made examples that violate [gender] stereotypes. We use Malle's (2020) listing of numerically-graded norms, where words vary in prescriptive strength. Table 1 shows the Oppose-Advocate intensity scores for various norm terms when used in the template "We should $\langle term \rangle$ $\langle entity \rangle$."

Int.	Norm Terms A	Norm Terms B
.9	demand; need	require; really want
.75	prefer; advocate for	recommend; want
.5	permit; allow	accept; welcome
-.5	discourage; frown upon	prevent; deter
-.75	disallow; not accept	reject; refuse
-.9	forbid; outlaw	prohibit; ban

Table 1: Intensity scores for different clusters of norms, primarily drawn from Malle's (2020) listing of terms with norm strength. Lexicon is divided into A/B groups, e.g., to train on A and test on B.

In addition, we created A/B groups of ABSA targets, where both groups have non-overlapping sets of human pronouns ("her," "him," "you," "them," etc.), human groups ("women," "men," "immigrants," "executives," "Muslims," "Christians," etc.), politicians ("Emmanuel Macron," "@macron," etc.), and topics (e.g., "abortion restrictions," "gun rights," "freedom of speech," "immigration rights," etc.).

We created norm corpora A and B by combining norm terms A and B with ABSA targets A and B, respectively, using the intensity scale in Table 1. This produces entries such as "We should reject economic reforms" in norm corpora B with Oppose-Advocate value of -0.75 for "economic reforms."

4. Experiments

Our experimental setup is adapted from previous debiasing approaches for data augmentation: we use the baseline model as a control condition, and then we augment the training data in three experimental conditions: norm Corpus A; norm Corpus B; or a corpus of randomly-perturbed data with the same number of examples in corpora A and B, combined. The perturbation condition is designed to measure the impact of using a competing data augmentation approach, since without any data augmentation, the baseline condition is operating with less training data.

The baseline dataset was labeled by three temporary workers for the ABSA attributes described above: Oppose-Advocate, Harmed-Helped, Harmful-Helpful, following the work of (Zheng et al., 2022). In total, 14,159 spans were annotated by all annotators. Krippendorff's alpha for three annotators is 0.808 for opposition-advocacy, 0.813 for harmful-helpful dimension, and 0.7 for harmed-helped. We train on 90% of this data and reserve 10% for testing, to measure the effect of our norm-based data augmentation.

We do not augment the 10% reserved test set with norm corpora; the test set is used to measure RMSE against annotated data, and the norm corpora (A and B) are solely used for training and for measuring bias (where B was used to measure bias of models trained on A, and the reverse).

Our ABSA NLP architecture uses a bert-base-uncased transformer model (Devlin et al., 2019), followed by max-pooling layer over all sub-words in the span to be ABSA-scored, followed by a linear layer for each of the three ABSA dimensions. We use mean-squared-error as the numerical ABSA loss function.

The A/B norm corpora—built from populating templates with the A/B norm terms and A/B ABSA targets described above—each have 252 spans to characterize.

5. Results

The metrics most relevant to this experiment are (1) RMSE on the held-out test set, to assess whether the debiasing intervention impacted accuracy and (2) standard deviation of ABSA judgments across spans in the same norm category, since this measures the distance of ABSA predictions in the same norm context.

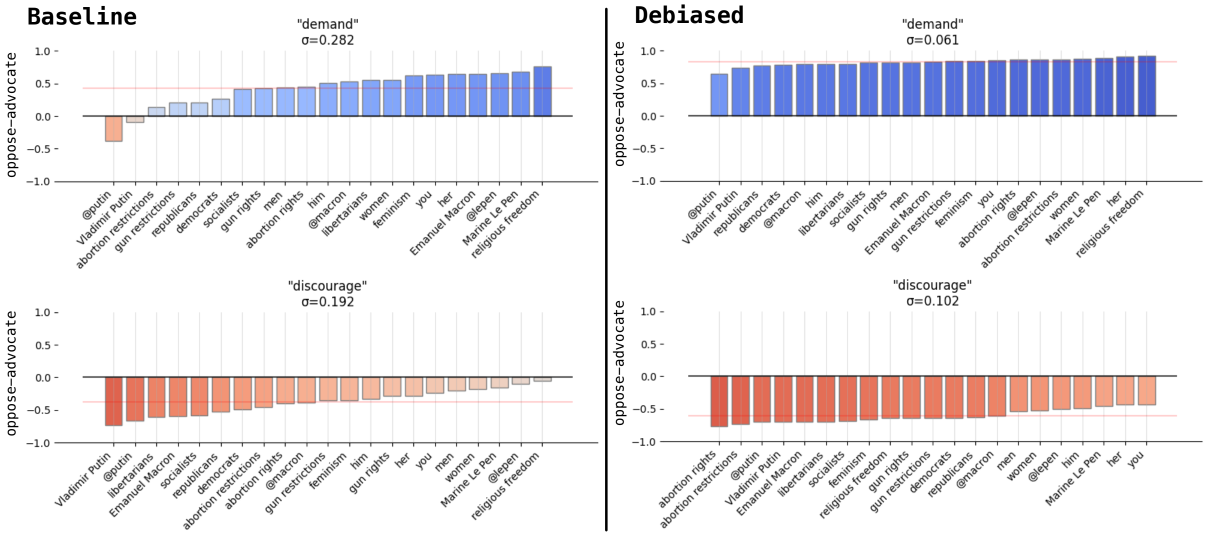

Figure 2: At left, baseline oppose-advocate predictions for two norm categories ("demand" and "discourage") across various ABSA target terms from the Norm Corpus A. At right, a model trained on Norm Corpus B makes the same Norm Corpus A predictions with lower variation across the same target terms.

Model	RMSE: oppo	RMSE: hrmd	RMSE: hrmf	stdev: oppo
Baseline	0.292	0.155	0.155	0.240
Pertub	0.289	0.154	0.155	0.233
Train-on-A	0.291	0.153	0.153	0.091
Train-on-B	0.293	0.154	0.152	0.115

Table 2: Results of augmenting two different norm corpora (A/B) into the training data, compared to the baseline.

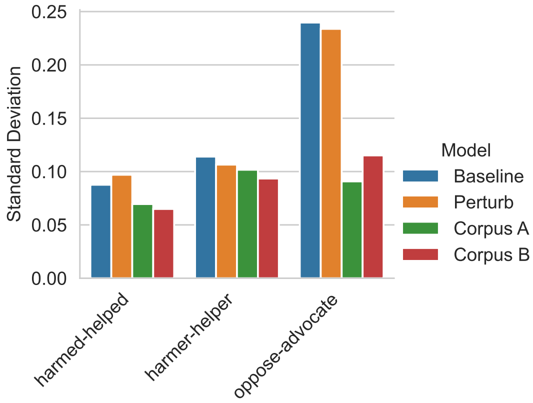

Figure 3: Standard deviation of per-aspect predictions across ABSA targets, comparing the baseline against training on two norm corpora (A and B).

As shown in Table 2, training by including the A and B norms corpora did not negatively impact the RMSE of any of the ABSA aspects. However, the standard deviation of ABSA judgments is substantially decreased for both debiasing conditions, reducing bias—as measured by standard deviation—to 38%-48% the bias of the baseline model. For Train-on-A and Train-on-B conditions, the standard deviation is assessed on opposite corpus (Corpus B and Corpus A, respectively), and for the baseline condition, standard deviation is assessed on both corpora.

Figure 2 shows examples of within-norm-category variation over different ABSA targets. In the left half of Figure 2, the baseline model is biased toward opposing a specific politician due to an abundance of training data opposing him. The opposition bias against this politician overcomes even the generally-positive context, as shown by the other ABSA targets. Likewise, the ABSA target "religious freedom" is difficult for the model to predict as opposed, even when explicitly "discouraged." On the right side of Figure 2, we see that debiasing with norm-based data augmentation reduces the variation across all terms and brings these ABSA targets into the fold. Figure 3 shows the standard deviations (y-axis) across the three ABSA aspects (x-axis), for four different conditions: the baseline (no debiasing applied); perturbation-based debiasing (with perturbation-based data augmentation); and the two norm-based data augmentation conditions (corpora A and B). The A and B debiasing corpora were labeled only in the oppose-advocate dimension, so intuitively, we see the largest decrease in standard deviation of predictions in that dimension. Random perturbation-based data augmentation did not have a substantial effect on bias, as measured by standard deviations. In Figure 3 we see that the other aspects (harmed-helped and harmful-helpful) were slightly debiased by training on either Corpus A or Corpus B. This suggests that the debiasing operations in the labeled oppose-advocate dimension impacted the upstream transformer component, diminishing bias less dramatically along other dimensions.

To test statistical significance of this debiasing approach against the baseline, we aggregate two conditions: (1) *perturb-condition inferences* on A+B and (2) *norm-debiased inferences* on A+B. We compare these conditions across the six grades of norms shown in Table 1. T-tests indeed show that all but one of these levels (for each of the norm levels in Table 1) are statistically significant: demand ($p = 5e\text{-}17$); prefer ($p = 2e\text{-}6$); permit (not significant); discourage: ($p = 9e\text{-}16$); disallow: ($p = 6e\text{-}14$); and forbid ($p = 6e\text{-}25$). These measures show that the debiasing operations significantly changed the distribution of scores.

6. Conclusion

This paper contributed an initial investigation of norm-based data augmentation to debias multi-entity ABSA models. In our two debiasing conditions, we extended the dataset by 1.8% to reduce measurable bias to less than 50% of the baseline in an ABSA dimension, without negatively impacting RMSE. We also present preliminary evidence that debiasing on one ABSA dimension can reduce bias in other dimensions, or at least, it does not *increase* bias on other dimensions.

The norm-based data augmentation approach presented here does not require targeting specific dimensions of bias; rather, it uses graded norms (Malle, 2020) as constant contexts while varying arbitrary topics to address many dimensions of bias.

Future work. Comparing norm-based debiasing with other approaches on the same dataset is a near-term goal, as well as exploring the effect of the size and diversity of the norm-based debiasing corpus and the number of ABSA target terms. Finally, the NLP neural layers may impact the model's ability to separate context from content, so representing context in the architecture, e.g., by pooling windows of preceding and subsequent subwords, could improve the model's receptiveness to debiasing with data augmentation.

Acknowledgments

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the view of the Defense Advanced Research Projects Agency (DARPA). We thank Dan Miller and COLING/LREC reviewers for helpful feedback.

7. Ethical Considerations

The ABSA model presented in this work is based on Bandura's (2016) psychosocial theory of moral disengagement. It predicts linguistic indicators that are relevant to this theory, but these indicators—even taken altogether—are not a complete solution to characterize that an author is morally disengaged, and our ABSA model should not be applied in this fashion.

This paper uses a standard deviation strategy to measure prediction bias in intensity-based ABSA scores, but even if the standard deviation is reduced (or zero) for some set of ABSA targets, systemic biases may exist for other ABSA targets—or for other ABSA contexts—that are not captured in the debiasing corpora. Consequently, an approach that measures standard deviation over a list of prediction targets should conclude that a model has been categorically de-biased.

8. Bibliographical References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. *arXiv preprint arXiv:2203.09192*.

Albert Bandura. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and social psychology review*, 3(3):193–209.

Albert Bandura. 2016. *Moral disengagement: How people do harm and live with themselves.* Worth publishers.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Scott Friedman, Sonja Schmer-Galunder, Anthony Chen, and Jeffrey Rye. 2019. Relating word embedding gender biases to gender gaps: A cross-cultural analysis. In *Proceedings of the first workshop on gender bias in natural language processing*, pages 18–24.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Joe Hoover, Gwenyth Portillo-Wightman, Leigh Yeh, Shreya Havaldar, Aida Mostafazadeh Davani, Ying Lin, Brendan Kennedy, Mohammad Atari, Zahra Kamel, Madelyn Mendlen, et al. 2020. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202.

Bertram Malle. 2020. Graded representations of norm strength. In *CogSci*.

Jie Tao and Xing Fang. 2020. Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data*, 7(1):1–26.

Jun Yang, Runqi Yang, Chongjun Wang, and Junyuan Xie. 2018. Multi-entity aspect-based sentiment analysis with context, entity and aspect memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Joan Zheng, Scott Friedman, Sonja Schmer-Galunder, Ian Magnusson, Ruta Wheelock, Jeremy Gottlieb, Diana Gomez, and Christopher Miller. 2022. Towards a multi-entity aspect-based sentiment analysis for characterizing directed social regard in online messaging. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 203–208.