Disentangling Dialect from Social Bias via Multitask Learning to Improve Fairness

Maximilian Spliethöver[†], **Sai Nikhil Menon**[*], **and Henning Wachsmuth**[†]
[†]Leibniz University Hannover, Institute of Artificial Intelligence
[*]Paderborn University, Department of Computer Science
{m.spliethoever,h.wachsmuth}@ai.uni-hannover.de

Abstract

Dialects introduce syntactic and lexical variations in language that occur in regional or social groups. Most NLP methods are not sensitive to such variations. This may lead to unfair behavior of the methods, conveying negative bias towards dialect speakers. While previous work has studied dialect-related fairness for aspects like hate speech, other aspects of biased language, such as lewdness, remain fully unexplored. To fill this gap, we investigate performance disparities between dialects in the detection of five aspects of biased language and how to mitigate them. To alleviate bias, we present a multitask learning approach that models dialect language as an auxiliary task to incorporate syntactic and lexical variations. In our experiments with African-American English dialect, we provide empirical evidence that complementing common learning approaches with dialect modeling improves their fairness. Furthermore, the results suggest that multitask learning achieves state-of-the-art performance and helps to detect properties of biased language more reliably.

1 Introduction

The term *social bias* is used broadly in the field of NLP. Existing works approach various facets, such as the affected social group (Sap et al., 2020), the tasks for which bias is evaluated (Blodgett et al., 2020), and the limited fairness of NLP systems in real-world settings, which may put specific social groups at a disadvantage while favoring others (Hovy and Spruit, 2016; Wu et al., 2022).

A specific fairness issue arises when a bias detection model is predominantly trained and evaluated on standard language but applied to texts with *dialect* (Jurgens et al., 2017). Dialects appear in regional and social communities and introduce syntactic and lexical variations (Blodgett et al., 2016). As such, dialects may notably diverge from the source language, posing a challenge to models trained primarily on standard language (Belinkov and Bisk, 2018; Ebrahimi et al., 2018; Kantharuban et al., 2023). If not explicitly accounted for, the lack of dialect understanding may subsequently lead to unfair decisions towards dialect speakers (Ziems et al., 2022), originating in data and label imbalances, but also in selected dialect terms that may be considered offensive in non-dialect contexts. For example, while the use of the N-word can be acceptable when used among African-American English (AAE) speakers, its use is considered inappropriate in Standard American English (SAE) (Rahman, 2012; Widawski, 2015; Talat et al., 2018).

Figure 1: Two texts from the corpus of Sap et al. (2020), showcasing some of the five social bias aspects tackled in this paper: Neither text is *lewd*, talks about some *target group*, or is from an *ingroup* member. Unlike (a), however, (b) is *offensive* and *intentional*. While (a) contains elements common in AAE, i.e., the habitual *be* and dropped copula (Ziems et al., 2022), (b) does not.

Most NLP models are, however, developed without consideration for dialect patterns and may, if any, only learn them implicitly through language modeling on large corpora. For example, at the time of writing, common language models, such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023a), Llama-2 (Touvron et al., 2023b), or DeBERTaV3 (He et al., 2023), all have not explicitly been trained or evaluated for dialects. The resulting disparities between dialect and standard

Findings of the Association for Computational Linguistics: ACL 2024, pages 9294–9313

languages have been widely recognized (Blodgett and O'Connor, 2017; Duarte et al., 2017; Tatman, 2017; Davidson et al., 2019; Resende et al., 2024). So far, however, few works focus on dialect-related fairness issues for social bias detection.

In this work, we investigate how to improve the fairness of social bias detection by explicitly modeling dialect language. We follow the *bias* definition of Sap et al. (2020), and consider *fairness* as equal classification performance between texts written in some dialect and other texts (Halevy et al., 2021). Concretely, we ask:

> *How to improve fairness in automated social bias detection between dialect and non-dialect language, while maintaining detection performance?*

To study this question, we propose a multitask learning approach that treats the dialect detection jointly with social bias classification tasks. Our hypothesis is that, by modeling dialect as an auxiliary task, the model gains a better internal representation of dialect language patterns and bias aspects. We expect that this does not only allow it to differentiate between dialect and biased language more easily, but also benefit non-dialect texts.

For evaluation, we focus on AAE as dialect, adopting the demographic-aligned definition of Blodgett et al. (2016), and five aspects of bias, namely offensiveness, lewdness, intention, targeting a group, and being part of the target group (cf. Section 4). Figure 1 illustrates how two examples relate to the dialect labels and the bias aspects.

In experiments, we compare the classification performance and fairness of our approach to baselines from related work as well as ablations that do not explicitly model dialect. We evaluate common performance and fairness metrics overall and per dialect. A perfectly fair model would show high performance without differences for dialect splits. Since, at the time of writing, no corpus for bias detection includes annotations for dialect use, we employ an automated data augmentation method.

The results of our experiments reveal performance disparities between AAE and non-AAE texts, and suggest that modeling multiple bias aspects helps to detect biased language more reliably. The proposed multitask learning approach improves over the best baseline and over single-task learning for four out of five bias aspects. Moreover, the dialect auxiliary task improves fairness for texts with dialect language and also benefits non-dialect texts. Learning five bias aspects and dialect detection simultaneously shows the most stable fairness and performance improvements across tasks.

To summarize, our main contributions are:

1. A multitask learning approach to jointly learn dialect and social bias detection.

2. Evidence that simultaneously modeling multiple bias aspects and dialect language improves the classification performance and fairness for (AAE) dialect speakers.[1]

2 Related Work

Social bias can be defined as stereotypical thinking or prejudices against social groups (Fiske, 1998). In NLP, it can manifest in hidden representations (Spliethöver and Wachsmuth, 2020) or unfair predictions (Angwin et al., 2016). Identifying social bias in data is an important step towards debiasing NLP models, since models adopt and amplify pre-existing biases which can have harmful effects (Zhao et al., 2017; Shwartz and Choi, 2020).

In related work, Wald and Pfahler (2023) analyze bias in fine-tuned large language models (LLM) as a proxy for bias in data. Focusing on single texts, Sap et al. (2020) introduce the Social Bias Inference Corpus (SBIC) and train a model to predict multiple bias aspects. Prabhumoye et al. (2022) apply few-shot learning to instruction-tuned LLMs on the same data. We build on the work of Sap et al. (2020), but extend it by considering dialects.

Relevant to dialects, another perspective to social bias is fairness regarding performance across social groups (Tolan, 2018). For example, Tatman (2017) find that video captioning systems perform worse for dialect speakers and women. In NLP, Blodgett et al. (2016) highlight disparities between dialect and standard language, as well as social groups in language identification. Resende et al. (2024) find negative biases in hate speech detection towards AAE texts due to underrepresentation in datasets. However, no work so far has considered fairness for the interplay of dialects and social bias. Ziems et al. (2022) find, similar to Joshi et al. (2024), that models perform worse on AAE compared to SAE in natural language understanding tasks.

Many existing fairness evaluations target toxicity detection. Mozafari et al. (2020) observe that a fine-tuned model labels AAE texts more often as

[1] Code at: https://github.com/webis-de/ACL-24

hate speech than SAE texts and propose a mitigation strategy. Similarly, Halevy et al. (2021) aim to mitigate this bias by introducing a dialect detection and dialect-specific toxicity classifiers. Their ensemble reconsiders positive toxicity predictions for AAE texts. Badjatiya et al. (2019) show that hate speech detection can be improved by removing bias-sensitive words commonly used in dialects, and Xia et al. (2020) propose an adversarial model to prevent false classifications for AAE text.

We aim to identify multiple aspects of biased language rather than just toxicity, and intend to incorporate dialect language into the model using multitask learning instead of separate models. Closest to our approach, Talat et al. (2018) try to overcome socio-demographic differences in hate speech annotations that arise from diversity in contexts and definitions. They train a multitask learning model on texts from various domains, annotated with separate definitions for hate speech. In experiments, their model outperforms existing approaches and generalizes better to unseen data. Unlike us, however, they do not account for dialect language, nor evaluate for respective social groups. Moreover, we use multitask learning to explicitly incorporate socio-demographic knowledge, namely dialect language patterns, into a unified model.

For data availability reasons, we automatically augment the SBIC. Modern language models have been tested with respect to their capabilities to augment training data. For example, Faggioli et al. (2023) evaluate the utility of LLMs as annotators for supervised learning and find that, for relevance judgment tasks, automatically-generated annotations show promising results. Zhang et al. (2023) extend the idea by introducing active learning. In contrast, we fine-tune an encoder model on a separate AAE dialect dataset and use it to annotate the bias aspect dataset for dialect language usage.

In terms of dialect data, Ziems et al. (2022), propose a rule-based approach to create a new benchmark derived from GLUE (Wang et al., 2018). Blodgett et al. (2016) introduce the TwitterAAE dataset containing around 60 million tweets with semi-supervised annotations. While many corpora focus on AAE vs. SAE (Groenwold et al., 2020), the TwitterAAE annotations make no further assumption about the non-AAE text. While a notable portion may be SAE, some might use other dialects. Since our work focuses on AAE vs. non-AAE and on texts from the internet, we employ the TwitterAAE corpus to train a dialect classifier.

3 Methodology

This work focuses on making model classifications fairer for dialect texts. In the following, we present our approach to improve fairness by integrating dialect with bias aspect detection in a multitask learning architecture. As dialect annotations for bias detection data are unavailable so far, we describe how we augment existing data automatically.

3.1 Joint Modeling of Social Bias and Dialect

Research has shown that a primary task's performance can improve through multitask learning, in which a trained model can transfer knowledge between primary and auxiliary tasks (Caruana, 1998).

In this work, we hypothesize that learning to detect dialect language as an auxiliary task improves the performance of a primary task for texts written in the given dialect. The addition of dialect aids the model in distinguishing simple dialect use from actual bias markers. For example, the text "*We was at some random-ass bar*" (Ziems et al., 2022) might not be lewd in a context where the dialect is commonly used, such as conversations between AAE speakers. However, similar use of the word *ass* in non-AAE contexts might be perceived as lewd or obscene (Ziems et al., 2022).

Furthermore, we expect that multitask learning does not only improve fairness, but also the reliability of identifying bias aspects. Given that bias aspects such as offensiveness, intentionality, and targeting a group are not fully independent from each other, multitask learning may leverage interdependencies to make more accurate predictions.

To operationalize our hypotheses, we propose a weight-sharing joint learning architecture (Collobert and Weston, 2008) that uses a shared encoder and separate classification heads for each task using a standard cross-entropy loss (Jurafsky and Martin, 2021), computed for each sample separately:

$$L_(\hat{\mathbf{y}}, \mathbf{y}) \quad := \quad -\sum_{i=1}^{n} y_i \log \hat{y}_i, \qquad (1)$$

where n is the number of samples, y_i the true label, and $\hat{y}_i$ the softmax output at position i. Here, the used labels alternate between the different dialect and bias aspect tasks in a round-robin manner.

Figure 2 illustrates our joint multitask learning architecture with k task-specific classification heads. We further add a classification head for the auxiliary task of learning to detect a specific dialect. At training time, each head is conditioned

on one task (i.e., bias aspect and dialect), whereas the shared encoder is fine-tuned for all $k+1$ tasks. In our specific case, the loss for the encoder model is calculated by alternating round-robin between tasks and individually being backpropagated to the encoder. For inference, only the classification head for the primary task is used.[2]

3.2 Data Augmentation with Dialect Labels

Previous work has shown that parallel data benefits multitask learning, as correlations between multiple labels are easier to identify, positively affecting all learned tasks (Pfeiffer et al., 2020). As noted in Section 2, however, no bias corpus with dialect labels exist. Furthermore, relying on a separate corpus for the auxiliary task (Collobert and Weston, 2008; Talat et al., 2018) may easily cause domain transfer problems and introduce noise.[3]

Therefore, we employ a data augmentation method. To this end, we train a dialect classifier on a separate corpus and then add dialect labels to the main corpus, as detailed in Section 4. Augmenting an existing corpus has two main advantages:[4]

1. It enables multitask learning approaches to transfer knowledge between the primary and auxiliary tasks more efficiently.
2. It is more generally applicable to other dialects, since dialect-specific classifiers can be developed independent from the approach.

4 Experiments

This section evaluates the effectiveness of multitask learning in improving fairness for dialects. We focus on African-American English (AAE) as a dialect and five social bias classification tasks. Below, we describe the experimental setups of our dialect data augmentation and our social bias detection approach. Using the augmented data, we test whether multitask learning improves fairness for dialect speakers if the dialect is modeled explicitly.

4.1 Data

We use the following two corpora for the auxiliary task of dialect classification and for the primary task of social bias classification, respectively.

[2]While joint learning requires re-training for new tasks or dialects, continual learning (Phang et al., 2019; Scialom et al., 2022) faces similar issues, and adapter fusion (Pfeiffer et al., 2021) performed notably worse in preliminary tests.

[3]Preliminary tests confirmed this assumption.

[4]Data augmentation also makes the experiments more controlled and less dependent on the content of the dialect corpus.

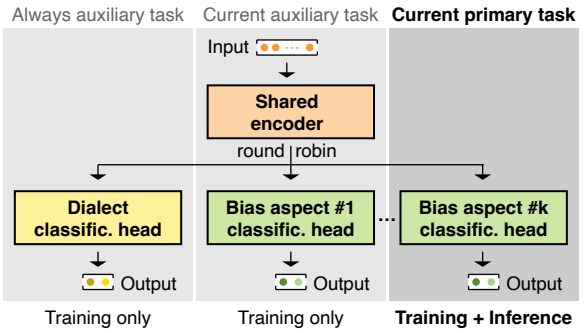

Figure 2: Our joint learning architecture: *Dialect classification* is added as an additional head to the classification of the bias aspects. During training, all classification heads are trained round-robin in alternating manner. For inference, only the classification head of the primary task is used, here the *Bias aspect* $\#k$ head.

TwitterAAE Corpus We train and evaluate AAE classification on the TwitterAAE corpus (Blodgett et al., 2016). The corpus contains about 59 million tweets from 2013, of which around 1 million are labeled as AAE dialect (dataset statistics are reported in Appendix D). The annotation was done semi-supervised, partially based on geolocation and user demographics. Since Blodgett et al. (2016) do not provide data splits, we randomly seperate 80% as training and validation set, and 20% as test set, in a stratified way to preserve dialect label ratios (using a seed for reproducability).

Social Bias Inference Corpus The Social Bias Inference Corpus (SBIC) (Sap et al., 2020) consists of about 45k English posts from online communities. Humans annotated each post for different aspects of biased language, which include the following five classification labels:[5]

1. *Offensiveness.* Whether or not a text is rude, disrespectful or shows toxicity
2. *Intent.* If a text is offensive, whether this offensiveness was intentional or not
3. *Lewdness.* Whether or not a text contains obscene or sexual references
4. *Target Group.* Whether or not a text is directed towards a specific social group
5. *Ingroup.* Whether or notthe author of the text is part of the targeted social group

We use the aggregated version of SBIC, for deduplicated texts and preprocessed labels (dataset

[5]Additionally, the corpus includes two free-text annotations describing the specific social group being targeted and the implied statement of a text. Since we focus on classification fairness, we disregard the free-text labels in this work.

statistics in Appendix D). For preprocessing, we lowercase texts, and remove mentions, retweet markers, URLs, and non-English characters. Following Sap et al. (2020), we binarize all labels.

4.2 African-American English Classification

To augment the SBIC with AAE dialect annotations, we develop a classifier to identify AAE texts. As mentioned in Section 2, we explicitly refrain from distinguishing AAE and SAE, as our goal is to seperate AAE from non-AAE texts.

Approach We fine-tune DeBERTa-v3-base (He et al., 2023) with a classification head on the TwitterAAE corpus. While bigger models exist, BERT-based text encoders still show state-of-the-art performance in various downstream tasks (He et al., 2023) and remain competitive for text-only classification tasks (Chen et al., 2023). Furthermore, the DeBERTa-v3-large variant did not show a notable increase in performance in preliminary tests.

Due to the strong imbalance in the TwitterAAE corpus, we evaluate two training methods:[6]

- *Subsampling (AAE_{smp}).* We randomly sample non-AAE texts in the training data (using a seed for reproducability) to match the number of AAE texts and create a balanced dataset. This method aims to equalize the importance of AAE and non-AAE labels.
- *Loss weighting (AAE_{wgh}).* Subsampling removes a potentially large number of training instances (nearly 57 million in this case). Instead, this method weighs the loss of each label, relative to the label distribution. For the given data, wrongly (or correctly) classifying AAE texts has, therefore, a higher impact on the model weights during backpropagation.

Baselines We compare our approach to the *TwitterAAE* dialect classifier presented by Blodgett et al. (2016). To verify that the models show a learning effect, we also report *majority* and *random* classifiers. Since we aim to reliably find AAE texts in particular, we emphasize the recall for this class.

Measures For all models, we report per-class and macro-averaged precision, recall, and F_1-scores.

4.3 Social Bias Detection

Now, we detail our proposed multitask learning approach to social bias detection, ablations to further investigate fairness, and the baselines we compare to. All proposed models are based on the DeBERTa-v3-base encoder (He et al., 2023).[7]

Approach We train a joint-weight multitask learning model (MULTITASK$_{+AAE}$) that consists of a shared encoder and separate classification heads for the five bias aspects. Moreover, we add a classification head to detect AAE dialect. We, therefore, train a model with six classification heads in total. As detailed above, the additional dialect detection task should help the model to better differentiate between dialects and biased language properties. Theoretically, this could be extended to further dialects, but we chose to restrict this study to AAE as an example.

Ablations To examine the effect of the auxiliary tasks, we evaluate two ablations of the approach. First, we train the multitask learning model on the five classification tasks *without* AAE dialect modeling (MULTITASK). This allows us to analyze the auxiliary task's influence on the multitask learning model. Second, we train single-task models *with* (SINGLETASK$_{+AAE}$) and *without* (SINGLETASK) AAE detection to better understand its influence on each task. To do so, we use a similar joint-weights learning setup as above.[8]

Baselines We compare our approaches and ablations to two baselines from related works for overall performance. First, we use the results of the best approach of Sap et al. (2020). The authors employ and fine-tune a *GPT-2* model, formulating the problem as an auto-regressive generation task. Second, we use scores of the overall best approach reported by Prabhumoye et al. (2022). Like Sap et al. (2020), the authors formulate the task as a generation task, but do so in a Q&A format. Instead of fine-tuning, however, Prabhumoye et al. (2022) employ a *few-shot* learning setup, providing the model with in-context examples during inference. Since neither the code, the model, nor the predictions per text are available, we compare to the scores reported in the respective papers.

Measures Following Sap et al. (2020) and Prabhumoye et al. (2022), we report the positive-class F_1-score for each task to assess classification per-

[6] In early tests, just fine-tuning led to a majority classifier.

[7] For results with RoBERTa-base as encoder model, see Appendix B.

[8] While the SINGLETASK$_{+AAE}$ is, in fact, also a multitask learning model, we refer to it as single-task model with AAE for clarity and better differentiation to the proposed approach.

	Precision ↑			Recall ↑			F_1 ↑		
Model	**Pos**	**Neg**	**Mac**	**Pos**	**Neg**	**Mac**	**Pos**	**Neg**	**Mac**
Majority	.00	.50	.25	.00	**1.0**	.50	.00	.67	.33
Random	.50	.50	.50	.50	.50	.50	.50	.50	.50
Twi.AAE	.73	.68	.71	.64	.77	.70	.69	.72	.70
AAE_{wgh}	‡**.80**	‡.76	‡**.78**	‡.74	‡.81	‡**.78**	‡**.77**	‡**.78**	‡**.78**
AAE_{smp}	‡.77	‡**.78**	‡.77	‡**.78**	.76	‡.77	‡**.77**	‡.77	‡.77

Table 1: African-American English dialect classification results: postive class (*Pos*), negative class (*Neg*) and macro-averaged (*Mac*). While AAE_{smp} seems better in finding dialect texts (higher recall for *Pos*), AAE_{wgh} performs better overall. Gains of both approaches over the TwitterAAE baseline are significant (‡ for $p < .01$).

Model	**Offens.**	**Intent**	**Lewdn.**	**Target**	**Ingroup**
Majority	.732	.694	.000	.000	.000
Random	.529	.504	.165	.456	.038
GPT-2	.788	.786	**.807**	.699	.000
Few-shot	.822	.798	.411	.737	–
STL	‡.875	‡.861	.744	‡.832	.000
STL_{+AAE}	‡.875	‡.861	.755	‡**.833**	.108
MTL	**‡**.882**	**‡**.864**	**.757	‡.832	**‡**.235**
MTL_{+AAE}	*‡.879	‡**.864**	.751	‡.831	**‡.227

Table 2: Bias classification results (positive-class F_1, averaged over five random seeds) for each aspect: offensiveness, intent, lewdness, target group, and ingroup. The additional dialect modeling in MULTITASK$_{+AAE}$ (MTL_{+AAE}) improves over single-task approaches and baselines in most cases. Most gains over the strongest baseline per bias aspect are significant (‡ for $p < .01$). Significant gains of multitask approaches over single-task variants are marked by * ($p < .05$) or ** ($p < .01$).

formance.[9] To study potential disparities and improvements, we further evaluate the classification performance per dialect (i.e., AAE and non-AAE). This allows us to test our hypothesis that classifiers work better for non-AAE texts than AAE texts and if the proposed approach improves upon this. Lastly, we also consider two common fairness metrics (Garg et al., 2020): *Predictive parity* describes the delta between both groups' precision scores (in this context referred to as the positive predictive value). Predictive parity is said to be satisfied if it is 0. *Equalized odds*, on the other hand, describes fairness based on recall (in this context referred to as true positive rate) and the false positive rate. It is said to be satisfied if deltas between the dialect groups are 0.

5 Results and Discussion

We first discuss the results of dialect classification, before we look at its interplay with social bias.

5.1 African-American English Classification

The results of the African-American English (AAE) dialect classification are reported in Table 1. Due to the heavy imbalance of the test dataset (only 2% are labeled as AAE dialect), we report metric scores on a randomly subsampled test set that balances AAE and non-AAE texts (using a seed for reproducability), with around 230k samples per class. For completeness, results on the full test set are reported in Appendix B.

Both our approaches, AAE_{wgh} and AAE_{smp}, outperform the previous state-of-the-art approach TwitterAAE in nearly all evaluations significantly. Especially, the gains on positive class precision (.80 vs. .73 for AAE_{wgh}) and recall (.74 vs. .64 for AAE_{wgh}) are noteworthy, as they allow us to identify more actual AAE texts more reliably in our main analysis. While scores increased less substantial over the baseline in negative class recall (i.e., identifying more non-AAE texts correctly, with .81 vs. .77 for AAE_{wgh}), increases in negative class precision are similarly noteworthy (.76 vs. .68 for AAE_{wgh}).

Overall, AAE_{wgh} not only performs better than TwitterAAE in all metrics, but also improves over AAE_{smp}, except for positive class recall and negative class precision. While the recall of the positive class is important in this task, AAE_{smp} would likely introduce more noise through false predictions, as indicated by its lower recall for the negative class, which also does not improve over the baseline. Therefore, we use AAE_{wgh} to augment the SBIC data with AAE dialect annotations.

5.2 Social Bias Detection

Table 2 presents the results of predicting the five bias aspects. Following Sap et al. (2020) and Prabhumoye et al. (2022), we report the F_1-score of the positive class for each aspect (for negative and macro F_1, see Appendix B).

Fine-tuning on single labels seems to work notably better than using a generative approach: SINGLETASK outperforms the two baselines (GPT-2 and few-shot learning) on three bias aspects significantly. We observe a strong F_1-score gain of 9.6 points over the best baseline on the target group aspect (.833 vs. .737). A better dialect language understanding (SINGLETASK$_{+AAE}$) further improves

[9] For completeness, we also report negative class and macro-averaged results in Appendix B.

	Offensiveness		Intent		Lewdness		Target Group		Ingroup	
Model	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**
Majority	.361	.397	.344	.368	.476	.467	.372	.360	.497	.482
Random	.495	.500	.492	.487	.396	.444	.499	.499	.342	.370
SINGLETASK	.854	.787	.854	.786	.861	.842	.863	.755	.497	.482
SINGLETASK$_{+AAE}$	.851	‡.808	.853	**.790**	†.869	.836	**.867**	**.760**	.530	.550
MULTITASK	****.860**	****.816**	***.856**	.784	****.870**	.840	**.867**	.756	** **.569**	**.630
MULTITASK$_{+AAE}$	*.856	.806	.855	.783	.865	*‡**.846**	**.867**	.749	.553	**.639**

Table 3: Bias classification results (macro F_1, averaged over five random seeds) for texts with and without AAE dialect (see Appendix B for precision, recall, and F_1-scores per class). While multitask learning has a notable impact, the AAE modeling especially improves the performance of singletask models. Significant gains are marked for multitask models over single-task variants (* $p < .05$, ** $p < .01$) and for AAE models over those without AAE modeling († $p < .05$, ‡ $p < .01$).

performance on lewdness from .744 to .755, while it seems to not influence the performance on offensiveness and intent. These results indicate that AAE dialect texts are most impacted by wrong predictions on the lewdness aspect when finetuning on a single task. In a qualitative analysis, we find that particularly for lewdness, better knowledge of AAE dialect patterns is helpful (cf. Section 5.4). Finally, all approaches that do not explicitly model AAE dialect predict the majority label for ingroup.

Both multitask approaches further improve upon SINGLETASK, showing performance increases in most aspects except target group. The biggest gains are achieved for the offensiveness and ingroup aspects. For ingroup, seemingly the most challenging aspect, MULTITASK and MULTITASK$_{+AAE}$ are among the only three evaluated models that show a learning effect, with a significantly improved F_1-score of .235 and .227 respectively. This may be due to a better ability to detect non-offensive contexts containing terms usually used offensively, and an increased awareness of impossible label combinations, such as predicting that an author is part of the target group, while also predicting that no group was targeted (cf. Section 5.4). These results indicate that multitask learning helps to detect biased aspects of language. Moreover, especially for more complex and implicit signals, such as ingroup, considering several aspects can help.

Modeling dialects seems to improve results most when finetuning on a single task. This is most visible for lewdness and ingroup, where the scores of SINGLETASK$_{+AAE}$ increase by .011 and .108 over SINGLETASK, respectively. A reason may be that jointly modeling the task and dialect disentangles dialect language from bias aspects and explicit word use, resulting in a model with better conceptual representations of the respective aspect (cf. Section 5.4). A model that considers only the aspect might not be complex enough to correctly interpret subtle changes in language introduced by dialects and thus maybe more prone misclassifications. Interestingly, MULTITASK$_{+AAE}$ does not benefit in the same of from modeling the dialect in addition to multiple aspects.

In conclusion, we find that, while supervised classification shows a notable performance increase over generative approaches in detecting bias aspects, considering multiple aspects of biased language jointly, clearly improves the reliability of the predictions further. Modeling dialect in addition to bias aspects has the most impact on models that consider only a single aspects otherwise.

5.3 Fairness in Social Bias Detection

Table 3 shows the results of bias aspect detection for AAE and non-AAE texts. As hypothesized in Section 1, the simple supervised fine-tuning of SINGLETASK performs indeed better for non-AAE texts, showing a difference of up to 10 points (.863 vs. .755 for target group). Such disparities could severely impact fairness if deployed in real-world applications. For example, if a system automatically flags offensive posts for removal, posts by dialect speakers would be falsely removed more often due to the decreased ability of the classifier to model dialect language.

The disparities between AAE and non-AAE performance further suggest that the SINGLETASK approach still partially relies on word usage rather than meaning to identify certain bias aspects, showing limited awareness of the AAE dialect (see Section 5.4). This interpretation is further supported by the fact that the SINGLETASK$_{+AAE}$ ablation im-

proves the performance on AAE texts for most aspects, and especially offensiveness and intent. While the gains come with decreases on non-AAE texts for selected aspects, dialect modeling still reduces the performance gap between AAE and non-AAE most of the time. It may thus be seen as a worthy trade-off, depending on the application.

Interestingly, however, SINGLETASK$_{+AAE}$ shows the opposite effect for two aspects: For lewdness, the performance on AAE texts drops from .842 to .836, but increases for non-AAE texts from .861 to .869. Similarly, an increase for non-AAE texts is visible for the target group and ingroup aspects. These results indicate that awareness of dialect language helps improves results for texts written with dialect, but also for those without.

The gains of multitask learning over the single-task approaches on AAE texts are more consistent. MULTITASK and MULTITASK$_{+AAE}$ reach the most notable gains on AAE texts for offensiveness and ingroup, improving upon SINGLETASK (.787) to .816 and .806, respectively. In terms of fairness, AAE modeling shows a similar pattern for multitask and single-task learning. MULTITASK$_{+AAE}$ improves the performance on AAE texts over MULTITASK and SINGLETASK on selected aspects. Unlike for SINGLETASK$_{+AAE}$, the gain in performance is only visible for AAE, while often slightly decreases for non-AAE. On lewdness, for example, the score increases significantly from .840 (MULTITASK) to .846 (MULTITASK$_{+AAE}$) for AAE texts, but decreases from .870 (MULTITASK) to .865 (MULTITASK$_{+AAE}$) for non-AAE texts. Similarly, the scores improve from .630 to .639 for ingroup on AAE texts. This supports our hypothesis that modeling dialects can improve internal concept representations of bias aspects (cf. Section 5.4), potentially being more important for singletask than for multitask approaches.

Equalized Odds and Predictive Parity MULTITASK$_{+AAE}$ has the biggest impact for equalized odds and predictive parity. Table 4 shows the results exemplarily for *offensiveness*.[10] For SINGLETASK$_{+AAE}$, the dialect modeling improves classification performance for AAE dialect most of the time, and also lowers the performance gap between AAE and non-AAE texts compared to SINGLETASK. While not all gains are substantial, they are, again, consistent across metrics and bias aspects. We hence see these results as further evidence for a positive impact of dialect modeling, especially in singletask learning architectures.

[10]Other aspects show similar patterns (see Appendix B).

Offens.	TPR ↑		FPR ↓		PPV ↑	
Model	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**
STL	.893	.918	.190	.372	.860	.826
STL$_{+AAE}$	.891	‡**.938**	.193	‡.354	.857	‡.836
MTL	**.896**	*.934	**.181**	**.331	***.866**	****.845**
MTL$_{+AAE}$	**.896**	.935	.189	.353	.861	.836

Table 4: Fairness results per approach in terms of true positive rate (*TPR*), false positive rate (*FPR*), and positive predictive value (*PPV*) for the offensiveness aspect (averaged over five random seeds; other aspects in Appendix B). Multitask learning improves results for AAE and reduces some differences to non-AAE (¬*AAE*). MULTITASK is best in most regards. Significant improvements are marked for multitask models over single-task variants (* $p < .05$, ** $p < .01$) and for AAE over non-AAE models († $p < .05$, ‡ $p < .01$).

Overall, both multitask learning and dialect modeling seem to improve the performance of texts written in a given dialect, as we displayed for AAE dialect in this work. The evaluation also suggests that the proposed models consistently make fairer predictions. While we focus on AAE dialect in this study, the dialect classification and bias detection component can be adapted to other dialects. Since the two components are independent, there are no constraints on the chosen dialect classification approach. We therefore expect that modeling dialects may also improve language understanding and performance for dialects other than AAE and encourage future work to consider this direction.

5.4 Qualitative Analysis

To further investigate potential improvements of modeling dialect as an auxiliary task, we conducted a qualitative analysis. Here, we summarize the main results only. A more exhaustive version of the analysis, including specific examples from the data, can be found in Appendix A.

Latent Bias Concepts Generally, modeling AAE dialect seems to benefit offensiveness and lewdness classification by improving the internal concept representations of the respective aspects. These improved representations seem to help the model abstract from word use towards relying more on contextual information. The better abstraction further seems to improve the interpretation of the whole context to decide whether a text is written by an ingroup member, as also demonstrated in Table 3.

Implicit Label Dependencies Due to the conceptual dependency of some labels in the SBIC corpus, certain label-value combinations are nonsensical and should not be predicted. For example, while a text can be intentionally offensive (thus being labeled as *offensive* and *intentional*, by definition of the aspects, it is impossible for a text to be intentional but not offensive. To model and identify this connection, however, it is necessary to consider both aspects simultaneously. While the SINGLETASK model predicts these impossible combinations for *offensive* and *intentional* only 59 times, both multitask learning variants, MULTITASK and MULTITASK$_{+AAE}$, eliminate the issue and never predict such wrong combinations. A similar effect can be observed for the *target group* and *ingroup* labels. These insights highlight the benefit of considering multiple bias aspects together. Future work might further investigate this effect.

6 Conclusion

In this work, we have studied the fairness of social bias detection with respect to dialects, presenting a multitask learning approach to model dialect as an auxiliary task. The approach aims to mitigate disparities in the bias detection performance across dialects. For data availability reasons, our experiments have focused on African-American English (AAE) texts and non-AAE texts. We have obtained state-of-the-art performance in predicting five different aspects of social bias. Moreover, modeling AAE dialect as an auxiliary task narrows selected performance disparities learning setup, thus making results fairer for AAE dialect texts.

This work, therefore, provides empirical evidence that explicitly encoding dialect language patterns into models can have a positive impact on fairness for dialect speakers. Especially when coupled with a reliable dialect classification model, we expect a notable effect. In future work, we aim to investigate how to model dialect language to improve our data augmentation methods, i.e., using as automatic dialect translation methods (Ziems et al., 2022) and counterfactual data generation (Zmigrod et al., 2019; Stahl et al., 2022). Lastly, collecting dialect data from more diverse sources should help to scale our approach to further text genres.

We hope this paper contributes towards fairer bias classification, and encourage others to consider dialects more broadly for NLP applications.

Limitations

To draw conclusions about dialect and standard language, it is essential to not only consider a single dialect. We, therefore, aimed to be careful in our work to point out that our results are limited to AAE vs. non-AAE language. A potential improvement could be to introduce a three-class classifier that can classify "AAE," "Standard," and "Other" language. More preferably, though, one could incorporate more dialects than just AAE. However, given that no part other than the dialect classifier is specific to AAE, we think that our experiments could be repeated for other dialects.

A second limitation concerns the AAE dialect annotations themselves. Since SBIC does not have such annotations, we have labeled them automatically using our classifier. Due to the fact that the fairness evaluation relies on the quality of the AAE dialect classifier, our results might be less conclusive than if humans had annotated the data. The same applies to all approaches incorporating the AAE dialect classifier for their predictions. While the AAE dialect classifier is far from perfect, we still consider it rather reliable based on our evaluation and think it provides a reasonable basis for our analysis, even if it cannot be conclusive.

Lastly, our results might be limited by the fact that not only our final evaluation dataset, SBIC, but also the AAE dialect dataset considers texts exclusively from online platforms (Twitter and Reddit) and in the English language. As mentioned earlier, dialects appear and vary in regional and social communities. Our evaluations therefore investigate only a sub-group of AAE speakers. Additionally, texts from online platforms usually show language patterns very different from other text forms, such as books or news articles (Nguyen et al., 2020). However, we consider this to be only a minor limitation since texts from online platforms are often used as resources for many kinds of NLP models. Reliably and fairly identifying social bias in such data is thus important and necessary. One just has to be aware that the approaches in this work might not apply to other forms of text in the same way.

Ethical Considerations

With our work, we try to improve selected ethical aspects of NLP applications. Namely, we consider the case of social bias detection with a specific focus on fairness for texts written in dialect language. If the developed approaches work as in-

tended, they should make overall predictions on social bias detection fairer for members of dialect-speaking social groups. In our specific case, those are members of the African-American community that choose to write in the AAE dialect. However, since the approaches and evaluation in this work focus on AAE dialect, they disregard other dialects and also Standard American English to a certain degree. Also, since we base our AAE dialect classifier on the dataset of (Blodgett et al., 2016), we limit the classifier ability to detect dialect language patterns present in the data. We acknowledge that other variants and of AAE dialect exist for regional and social communities that were not included in the data, i.e., because they do not make (extensive) use of Twitter. In both cases, the classification performance might be implacted. While developing our approaches, we aimed to make approaches agnostic to specific dialects, and given that data exists, we think they might be adaptable to other dialects.

Another aspect that requires consideration is the problem of bias. We identify two main areas of bias in our work: the data we used and our own bias as researchers. In the former case, our work relies on the assumption that annotations are not biased. However, they are made by humans with personal worldviews and biases which, intentionally or not, might have mislabeled the data (Sap et al., 2022). Especially for our scenario of AAE texts, human annotators who were not part of the African-American community might have confused dialect use with, e.g., offensiveness (Widawski, 2015). Such wrong labels, that we assume to be correct, may influence models and evaluations. Similarly, we, the authors of the paper, might have introduced personal biases by applying our specific world views to the problem or unintentionally making false assumptions, leading to potential oversights.

Finally, we conceive that our approaches might be misused in situations where it is helpful for actors to label a product or approach "debiased." Since none of the presented approaches is perfect for detecting social bias or being completely fair for all dialects, as stated above, they are also not ready for production use. Even unintentionally, actors might apply our approaches to their data, wrongly assuming that it identifies all possible cases of social bias. This might, however, rather be an issue regarding the communication of our work to the more general public, as we assume that this will not be problematic for everyone that takes the time to read this paper entirely.

Acknowledgments

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/1 2021 – 438445824. We thank the anonymous reviewers for their helpful feedback and suggestions.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *Machine Bias*.

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In *The World Wide Web Conference*, WWW '19, pages 49–59, New York, NY, USA. Association for Computing Machinery.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation.

Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The Digital Landscape of Nudging: A Systematic Literature Review of Empirical Research on Digital Nudges. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, pages 1–16, New York, NY, USA. Association for Computing Machinery.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. ArXiv:1707.00061 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei.

2020. Language Models are Few-Shot Learners. ArXiv: 2005.14165 version: 3.

Rich Caruana. 1998. Multitask Learning. In Sebastian Thrun and Lorien Pratt, editors, *Learning to Learn*, pages 95–133. Springer US, Boston, MA.

Zhihong Chen, Guiming Chen, Shizhe Diao, Xiang Wan, and Benyou Wang. 2023. On the Difference of BERT-style and CLIP-style Text Encoders. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13710–13721, Toronto, Canada. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 160–167, New York, NY, USA. Association for Computing Machinery.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Natasha Duarte, Emma Llanso, and Anna Loup. 2017. Mixed messages? The limits of automated social media content analysis. Report, Center for Democracy and Technology.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Guglielmo Faggioli, Laura Dietz, Charles L. A. Clarke, Gianluca Demartini, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kanoulas, Martin Potthast, Benno Stein, and Henning Wachsmuth. 2023. Perspectives on Large Language Models for Relevance Judgment. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, pages 39–50, New York, NY, USA. Association for Computing Machinery.

Susan T. Fiske. 1998. Stereotyping, prejudice, and discrimination. In *The handbook of social psychology*, 4 edition, volume 1-2, pages 357–411. McGraw-Hill, New York, NY, US.

Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness Metrics: A Comparative Analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in Transformer-Based Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Matan Halevy, Camille Harris, Amy Bruckman, Diyi Yang, and Ayanna Howard. 2021. Mitigating Racial Biases in Toxic Language Detection with an Equity-Based Ensemble Framework. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21, pages 1–11, New York, NY, USA. Association for Computing Machinery.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. Kigali, Rwanda.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural Language Processing for Dialects of a Language: A Survey.

Dan Jurafsky and James H. Martin. 2021. *Speech and Language Processing*, 3rd edition draft edition. Online.

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky. 2017. Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57, Vancouver, Canada. Association for Computational Linguistics.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the Dialect Gap and its Correlates Across Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PLOS ONE*, 15(8).

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, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Edwin Simpson, and Iryna Gurevych. 2020. Low Resource Multi-Task Sequence Tagging – Revisiting Dynamic Conditional Random Fields. ArXiv:2005.00250 [cs].

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks.

Shrimai Prabhumoye, Rafal Kocielnik, Mohammad Shoeybi, Anima Anandkumar, and Bryan Catanzaro. 2022. Few-shot Instruction Prompts for Pretrained Language Models to Detect Social Biases.

Jacquelyn Rahman. 2012. The N Word: Its History and Use in the African American Community. *Journal of English Linguistics*, 40(2):137–171.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: Memory Optimizations Toward Training Trillion Parameter Models.

Guilherme H. Resende, Luiz F. Nery, Fabrício Benevenuto, Savvas Zannettou, and Flavio Figueiredo. 2024. A Comprehensive View of the Biases of Toxicity and Sentiment Analysis Methods Towards Utterances with African American English Expressions.

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, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned Language Models are Continual Learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vered Shwartz and Yejin Choi. 2020. Do Neural Language Models Overcome Reporting Bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maximilian Spliethöver and Henning Wachsmuth. 2020. Argument from Old Man's View: Assessing Social Bias in Argumentation. In *Proceedings of the 7th Workshop on Argument Mining*, pages 76–87, Online. Association for Computational Linguistics.

Maja Stahl, Maximilian Spliethöver, and Henning Wachsmuth. 2022. To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 39–51, Abu Dhabi, UAE. Association for Computational Linguistics.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In Jennifer Golbeck, editor, *Online Harassment*, Human–Computer Interaction Series, pages 29–55. Springer International Publishing, Cham.

Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Songül Tolan. 2018. Fair and Unbiased Algorithmic Decision Making: Current State and Future Challenges. Technical Report, European Commission.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and Efficient Foundation Language Models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Celine Wald and Lukas Pfahler. 2023. Exposing Bias in Online Communities through Large-Scale Language Models.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding.

Maciej Widawski. 2015. *African American slang: a linguistic description*. Cambridge University Press, Cambridge, United Kingdom.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Fangsheng Wu, Mengnan Du, Chao Fan, Ruixiang Tang, Yang Yang, Ali Mostafavi, and Xia Hu. 2022. Understanding Social Biases Behind Location Names in Contextual Word Embedding Models. *IEEE Transactions on Computational Social Systems*, 9(2):458–468. Conference Name: IEEE Transactions on Computational Social Systems.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting Racial Bias in Hate Speech Detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Ruoyu Zhang, Yanzeng Li, Yongliang Ma, Ming Zhou, and Lei Zou. 2023. LLMaAA: Making Large Language Models as Active Annotators. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13088–13103, Singapore. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, and Diyi Yang. 2022. VALUE: Understanding Dialect Disparity in NLU. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3701–3720, Dublin, Ireland. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

A Qualitative Analysis

Selected samples that showcase properties of modeling multiple bias aspects (multitask learning) and/or additionally modeling the AAE dialect simultaneously are shown in Table 5. In the remainder of this section, we shortly discuss some of the properties we observed by manually looking at the data and analyzing selected classification results. Each example will be referred to by its index, presented in the first column of the table (i.e., "Example 2" refers to the example in the second row).

A.1 Modeling AAE Improves Latent Concepts

For selected samples, modeling AAE dialect seems to benefit the offensiveness and lewdness classifications by improving the internal concept representations of the respective aspects. The improved representations seem to help the model abstract from word use towards relying more on contextual information.

The improved representations are shown anecdotally by Examples 4, 11, 14, 22, 23, and 24, all classified as lewd by the SINGLETASK model but not by the SINGLETASK$_{+AAE}$ model. For offensiveness, Examples 1, 2, 3, 10, 12, 15, 17, 18, and 21 showcase the behavior. While all listed examples include words that often indicate offensive or lewd statements, they can also be used outside of said contexts, i.e., in statements that are simply obscene.

As mentioned above, one potential theory might be that, due to the additional AAE dialect modeling and the resulting awareness of dialect language, the model is forced to learn better internal concepts of lewdness and offensiveness less strongly based on word use.

Furthermore, Examples 27 and 28 showcase the positive effect of modeling dialect on the ingroup label. While both examples include the N-word, the use of the AAE dialect version ("a"-ending) and other AAE dialect properties, such as dropping "have", suggest that the authors of both texts are ingroup members using ingroup language (this is confirmed by the original annotation). Both examples were classified wrongly by the SINGLETASK and MULTITASK models as not ingroup texts, while being classified correctly by the SINGLETASK$_{+AAE}$ and MULTITASK$_{+AAE}$ models.

#	Offens.	Intent	Lewdn.	Target	Ingroup	Post content
1	Yes	Yes	No	No	No	You all make fucking terrible "music"
2	No	No	No	No	No	“This guy tko’s people. when I knock people out they don’t fucking move." - eddie alvarez
3	No	No	No	No	No	She text me "i hate you, i hate you, i hate you, i hate you" but how the fuck can you hate me?
4	No	No	No	No	No	A man already accused of killing two women and suspected in the deaths of as many as 30 more was charged with raping and murdering five other victims.
5	Yes	Yes	No	No	No	Always remember, it was normies that enacted the holocaust, acted in the slave trade, and committed every atrocity in history normies are evil
6	No	No	No	No	No	Amnesty for 3.6m colonizers would/will be the completion of the demographic coup
7	Yes	Yes	No	No	No	Are orphans allowed in a family restaurant. I think you get it.
8	No	No	No	No	No	Asking north korea not to make nuclear weapons is a bit like asking an audi driver not to drive too close to the car in front
9	No	No	No	No	No	Bpd is a genetic condition caused by having two x chromosomes. When a man is diagnosed with bpd it’s just a professional way of saying he’s acting like a cunt.
10	No	No	No	No	No	Draymond kill me with all that shit he be talking to the refs. Please be quiet boy before you get us all in trouble.
11	Yes	Yes	No	Yes	No	Everytime I clean the room my mom tells me "what bitch is coming over this time" lmao what a dick !!
12	No	No	No	No	No	gooooooooooaaallll gooooooooooaaallll gooooooooooaallll pepe pepe pepe!!!! renew his fucking contract!!!
13	No	No	No	No	No	I do not blame you, ottawa, like all of our cities, is a multicultural hell
14	Yes	No	No	No	No	It’s always the people who tell you not to be so obsessed about women and sex who are the ones who can get them anytime they want
15	No	No	No	No	No	Natiis ka nya. It means he doesn’t fucking care anymore
16	No	No	No	No	No	On god I ain’t never wanna plex with lil bro them at all. I ain’t scared to squabble nobody I just refuse to fight a nigga i had love 4
17	No	No	No	No	No	Bad bitches don’t take days off
18	No	No	No	No	No	I ain gone lie. Chuckie fucked w| my head a lil bit
19	Yes	Yes	No	No	No	Sis don’t assume I’m talking about yo nigga and get in yo feelings.. Yo nigga not the only one on my line
20	Yes	No	No	No	No	Some of y’all lives don’t even be bad.. Y’all jus either ungrateful or just don’t know how to put things in perspective
21	No	No	No	No	No	Vince once again fucked over undertaker over a blind push. ’taker is over now, and so is me as a fan. #vincescrewedtaker
22	Yes	Yes	No	Yes	No	Why does the myth that women rarely lie about rape persist?
23	No	No	No	No	No	Just ordered horny goat weed this shit better fix my ed
24	Yes	Yes	No	No	No	Sorry I’m not humble, dick.
25	No	No	No	No	No	Your not gonna change a nigga. He is who he is either you accept it or you dont. You cant turn silver into gold.
26	No	No	No	No	No	That’s some real hood shit, I’ll never forget heading out wit my niggas & realizing that someone stole his gas cap
27	Yes	Yes	No	Yes	Yes	I come a city where niggas will hate on you for anything.
28	Yes	No	No	Yes	Yes	Nigga Andrew and x been telling me since like 11’ everything’s for sale and legit everything is for sale you tell me a great number lol

Table 5: Example posts from the SBIC test set that highlight different properties and capabilities of the evaluated models. Please note that neither the texts themselves nor the labels assigned to them necessarily reflect the authors opinions. Both are retrieved from the Social Bias Inference Corpus (Sap et al., 2020), which collects texts from the internet and annotations from crowdworkers. Some posts were formatted to improve readability.

Note that the listed examples include both, texts with and without properties of AAE dialect. The effect therefore seems to benefit not only texts containing AAE dialect, but also text without AAE dialect. Especially for the lewdness classification, this can also be observed in Table 2 and Table 3. Lastly, modeling AAE can also help to detect offensive statements, which make use of dialect language elements, more reliably (Example 19).

However, since these examples can only anecdotally show this effect, future work might further investigate and attempt to quantify this theory.

It is important to mention that classifications are still not perfect, and selected samples are missed. For instance, Example 19 and 25 both contain AAE dialect elements (such as dropped copula (Ziems et al., 2022) and use of n-word with an "a"-ending (Rahman, 2012)) and are not obviously offensive (and are also not labeled as such in the SBIC). However, all evaluated models classified them as offensive, including MULTITASK and MULTITASK$_{+AAE}$.

A.2 Multitask Learning Improves Label Dependency Modeling

Due to the conceptual dependency of some labels in the SBIC corpus, certain label-value combinations are nonsensical and should not be predicted. For example, while a text can be intentionally offensive (thus being labeled as *offensive* and *intentional*, such as Example 1), by definition of the aspects, the text cannot be intentional but not offensive. Another example is the label combination of *target group* and *ingroup*: Without referencing a target group in a statement, it is also impossible to state whether the author of the text is part of the referenced group (i.e., the *target group* label is false, while the *ingroup* label is true).

To model and identify this connection, however, it is necessary to consider both aspects simultaneously, as it is lost when considering aspects individually. While the SINGLETASK model predicts these impossible combinations for *offensive* and *intentional* only 59 times (e.g., Example 1), both multitask learning variants, MULTITASK and MULTITASK$_{+AAE}$, eliminate the issue and never predict such wrong combinations. While it is not possible to evaluate this behavior on the *target group* and *ingroup* combinations for SINGLETASK and SINGLETASK$_{+AAE}$ (both only predict the majority class for the *ingroup* label), MULTITASK$_{+AAE}$ only makes one such error (Example 26). At the same time, MULTITASK never does.

While this error appeared only for a smaller number of samples, the extreme effect observed for the multitask models highlights the benefit of considering multiple bias aspects simultaneously. We assume that the effect is also present in more subtle ways throughout the rest of the dataset. Future work might further investigate this effect.

A.3 Incorrect Annotations

In a few cases, it seems that most of the evaluated approaches "correct for" wrong annotations in the corpus and predict the, to the authors' perception, correct label value, even though they are mislabeled in the SBIC. Instances of such "correct misclassifications" for the *target group* label are Examples 5, 6, 7, 8, 9, and 13.

One possible explanation might be that, for all those instances, the target groups are not within the focus of the original SBIC study presented by Sap et al. (2020). Therefore, these target groups were also not part of the pre-defined options in the annotation interface (shown in the appendix of Sap et al. (2020)), and required manual user input. While a reasonable design decision for, what we assume to be the focus of the corpus, it might have caused "Friction nudges" (Bergram et al., 2022), causing annotators to disregard or consider fewer other target groups.

While such cases are present, they seem to be very few and do not seem to pose a notable challenge, as models can apparently correct for the slight noise.

B Further classification results

B.1 AAE Classification

Table 6 shows the results of the AAE classification approaches AAE$_{wgh}$ and AAE$_{smp}$ on the full test set. While the precision values for the positive class seem small, it is to be considered that this is a needle-in-the-haystack problem: Only about 2% of the test cases are positive, meaning that the best result (.07 of our approach AAE$_{wgh}$) is 3.5 times better than guessing. An increase of two points over the TwitterAAE baseline (.07 vs. .05) also indicates a notable learning effect, classifying about 40% more correctly.

B.2 Social Bias Classification

Table 7 and Table 8 show the negative class and macro averaged F_1-scores of the social bias classification. Results for the positive class are reported

	Precision ↑			Recall ↑			F_1 ↑		
Model	**Pos**	**Neg**	**Mac**	**Pos**	**Neg**	**Mac**	**Pos**	**Neg**	**Mac**
Majority	.00	.98	.49	.00	**1.0**	.50	.00	**.99**	.50
Random	.02	.98	.50	.50	.50	.50	.04	.66	.35
TwitterAAE	.05	**.99**	.52	.64	.77	.70	.10	.87	.48
AAE_{wgh}	**.07**	**.99**	**.53**	.74	.81	**.78**	**.13**	.89	**.51**
AAE_{smp}	.06	**.99**	**.53**	**.78**	.77	.77	.11	.86	.49

Table 6: African-American English dialect classification results on the full test set: Postive class (*Pos*), negative class (*Neg*) and macro averaged (*Mac*). Bold values highlights the best result in each column. While AAE_{smp} seems better in finding dialect texts (higher recall for the positive class), AAE_{wgh} performs notably better overall.

Model	**Offens.**	**Intent**	**Lewdn.**	**Target**	**Ingroup**
Majority	.000	.000	.949	.741	**.990**
Random	.463	.479	.639	.543	.654
GPT-2	–	–	–	–	–
Few-shot	–	–	–	–	–
SINGLETASK	.820	**.834**	.973	.869	**.990**
$SINGLETASK_{+AAE}$	.820	.833	.974	**.876**	**.990**
MULTITASK	**.830**	.833	**.975**	**.876**	.989
$MULTITASK_{+AAE}$	.824	.832	.974	.874	**.990**

Table 7: Bias classification results (negative class F_1-scores, averaged over five random seeds) for each aspect: offensiveness, intent, lewdness, target group, and ingroup.

Model	**Offens.**	**Intent**	**Lewdn.**	**Target**	**Ingroup**
Majority	.366	.347	.475	.370	.495
Random	.496	.492	.402	.499	.346
GPT-2	–	–	–	–	–
Few-shot	–	–	–	–	–
SINGLETASK	.848	.847	.859	.851	.495
$SINGLETASK_{+AAE}$	.847	.847	.864	.854	.549
MULTITASK	**.856**	**.849**	**.866**	**.854**	**.612**
$MULTITASK_{+AAE}$	.852	.848	.863	.853	.608

Table 8: Bias classification results (macro F_1-scores, averaged over five random seeds) for each aspect: offensiveness, intent, lewdness, target group, and ingroup.

Model	**Offens.**	**Intent**	**Lewdn.**	**Target**	**Ingroup**
Majority	.732	.694	.000	.000	.000
Random	.529	.504	.165	.456	.038
GPT-2	.788	.786	**.807**	.699	.000
Few-shot	.822	.798	.411	.737	–
SINGLETASK	.875	**.859**	.722	.826	.000
$SINGLETASK_{+AAE}$	.875	.857	.752	**.833**	.281
MULTITASK	.875	.857	.754	.826	**.290**
$MULTITASK_{+AAE}$	**.877**	**.859**	.751	.828	.213

Table 9: Bias classification results (positive-class F_1, averaged over five random seeds) for each aspect: offensiveness, intent, lewdness, target group, and ingroup. Results obtained using a RoBERTa-base encoder model, as comapred to Table 2, where DeBERTa-v3-base is used as encoder model.

in Table 2, as part of Section 5.

B.3 Encoder model

To verify that our results are not specific to the chosen DeBERTa-v3-base encoder model, Table 9 shows the results for the positive F_1-scores, matching Table 2. The results highlight that, while the RoBERTa-base models do not seem to benefit to the same degree the DeBERTa-v3-base models do, most advantages of the multitask learning setup and dialect modeling discussed in Section 5 also hold in this setup.

B.4 Social Bias Classification Scores per Dialect

Table 10, Table 11, Table 12, Table 13 and Table 14 show the per-class precision, recall and F_1-scores for the social bias classifications. Macro-averaged F_1 scores are reported in Table 3, as part of Section 5.

B.5 Fairness in Social Bias Classification

Table 15, Table 16, Table 17, and Table 18 show True Positive Rate, False Positive Rate and Positive Predictive Value scores on the positive class for the bias aspects *intent*, *lewdness*, *target group*, and *ingroup*, respectively. Results for the *offensiveness* aspect are reported in Table 4, as part of Section 5.

C Experimental Details

C.1 AAE Classification

Models for the AAE classification (AAE_{wgh} and AAE_{smp}) were fine-tuned for three epochs on three A100-SXM4-80GB GPUs and a batch size of 270. To keep training time reasonable, given the size of the dataset, we fine-tune the model with bf16 mixed precision using the DeepSpeed (Rajbhandari et al., 2020) integration of the Huggingface library (Wolf et al., 2020). With this setup, fine-tuning takes around 70 hours for AAE_{wgh}, and around 17 hours for AAE_{smp}.

For all models, we report results for a single training and inference run.

Offensiveness	Precision						Recall						F_1					
	Positive		Negative		Macro		Positive		Negative		Macro		Positive		Negative		Macro	
Model	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**
Majority	.566	.658	.000	.000	.283	.329	**1.00**	**1.00**	.000	.000	.500	.500	.723	.794	.000	.000	.361	.397
Random	.564	.670	.432	.354	.498	.512	.485	.514	.510	.513	.498	.513	.522	.582	.468	.419	.495	.500
SINGLETASK	.860	.826	.853	.800	.857	.813	.893	.918	.810	.628	.852	.773	.876	.870	.831	.704	.854	.787
SINGLETASK$_{+AAE}$	.857	.836	.851	**.844**	.854	.840	.891	.938	.807	.646	.849	.792	.874	.884	.828	.732	.851	.808
MULTITASK	**.866**	**.845**	**.859**	.842	**.862**	**.843**	.896	.934	**.819**	**.669**	**.858**	**.802**	**.881**	**.887**	**.838**	**.745**	**.860**	**.816**
MULTITASK$_{+AAE}$	.861	.836	.857	.838	.859	.837	.896	.935	.811	.647	.854	.791	.878	.883	.833	.730	.856	.806

Table 10: Offensiveness classification results on texts written with (*AA.*) and without (*NA.*) AAE dialect for the positive (*Positive*) and negative class (*Negative*), and macro averaged (*Macro*). All scores are averaged over five random seeds. Bold indicates best results per column.

Intent	Precision						Recall						F_1					
	Positive		Negative		Macro		Positive		Negative		Macro		Positive		Negative		Macro	
Model	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**
Majority	.525	.583	.000	.000	.262	.292	**1.00**	**1.00**	.000	.000	.500	.500	.688	.737	.000	.000	.344	.368
Random	.517	.577	.468	.411	.493	.494	.490	.460	.496	.528	.493	.494	.503	.512	.482	.462	.492	.487
SINGLETASK	**.851**	**.783**	.859	.837	.855	.810	.877	.910	**.831**	**.648**	.854	.779	.864	.842	**.845**	.731	.854	.786
SINGLETASK$_{+AAE}$	.847	.781	.862	**.862**	.855	**.822**	.880	.927	.825	.637	.853	**.782**	.863	**.848**	.843	**.732**	.853	**.790**
MULTITASK	.843	.777	**.875**	.856	**.859**	.816	.894	.924	.816	.628	**.855**	.776	**.868**	.844	.844	.724	**.856**	.784
MULTITASK$_{+AAE}$	.842	.776	.874	.856	.858	.816	.894	.925	.814	.626	.854	.775	.867	.844	.843	.723	.855	.783

Table 11: Intent classification results on texts written with (*AA.*) and without (*NA.*) AAE dialect for the positive (*Positive*) and negative class (*Negative*), and macro averaged (*Macro*). All scores are averaged over five random seeds. Bold indicates best results per column.

Lewdness	Precision						Recall						F_1					
	Positive		Negative		Macro		Positive		Negative		Macro		Positive		Negative		Macro	
Model	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**
Majority	.000	.000	.907	.877	.454	.438	.000	.000	**1.00**	**1.00**	.500	.500	.000	.000	.951	.934	.476	.467
Random	.092	.143	.907	.895	.500	.519	.509	.565	.490	.522	.500	.544	.156	.228	.636	.660	.396	.444
SINGLETASK	.757	**.783**	.974	.954	.865	**.869**	.739	.670	.976	.973	.858	.822	.748	.721	.975	**.964**	.861	.842
SINGLETASK$_{+AAE}$	.756	.748	**.976**	.955	.866	.852	**.768**	.678	.975	.968	**.871**	.823	.762	.711	**.976**	.961	.869	.836
MULTITASK	**.779**	.760	.975	.955	**.877**	.858	.749	.678	.978	.970	.864	.824	**.764**	.717	**.976**	.963	**.870**	.840
MULTITASK$_{+AAE}$	.771	.781	.974	**.956**	.872	.868	.740	**.681**	.978	.973	.859	**.827**	.755	**.728**	**.976**	**.964**	.865	**.846**

Table 12: Lewdness classification results on texts written with (*AA.*) and without (*NA.*) AAE dialect for the positive (*Positive*) and negative class (*Negative*), and macro averaged (*Macro*). All scores are averaged over five random seeds. Bold indicates best results per column.

Group	Precision						Recall						F_1					
	Positive		Negative		Macro		Positive		Negative		Macro		Positive		Negative		Macro	
Model	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**
Majority	.000	.000	.592	.562	.296	.281	.000	.000	**1.00**	**1.00**	.500	.500	.000	.000	.743	.719	.372	.360
Random	.412	.440	.595	.564	.503	.502	.504	.510	.503	.494	.503	.502	.453	.473	.545	.526	.499	.499
SINGLETASK	.806	.674	**.916**	**.862**	.861	**.768**	**.887**	**.859**	.852	.674	.869	**.766**	.844	**.754**	.883	.755	.863	.755
SINGLETASK$_{+AAE}$	.821	.699	.908	.824	**.865**	.761	.873	.800	.869	.731	**.871**	.765	**.846**	.746	.888	**.774**	**.867**	**.760**
MULTITASK	**.826**	**.700**	.905	.814	**.865**	.757	.866	.784	.874	.737	.870	.760	.845	.739	**.889**	.773	**.867**	.756
MULTITASK$_{+AAE}$	.822	.687	.908	.813	**.865**	.750	.872	.788	.869	.720	**.871**	.754	**.846**	.734	.888	.764	**.867**	.749

Table 13: Group classification results on texts written with (*AA.*) and without (*NA.*) AAE dialect for the positive (*Positive*) and negative class (*Negative*), and macro averaged (*Macro*). All scores are averaged over five random seeds. Bold indicates best results per column.

In-group	Precision						Recall						F_1					
	Positive		Negative		Macro		Positive		Negative		Macro		Positive		Negative		Macro	
Model	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**	**NA.**	**AA.**
Majority	.000	.000	.988	.932	.494	.466	.000	.000	**1.00**	**1.00**	.500	.500	.000	.000	**.994**	**.965**	.497	.482
Random	.014	.062	**.989**	.926	.501	.494	**.569**	**.474**	.492	.478	.530	.476	.027	.110	.657	.630	.342	.370
SINGLETASK	.000	.000	.988	.932	.494	.466	.000	.000	.000	.000	.500	.500	.000	.000	**.994**	**.965**	.497	.482
SINGLETASK$_{+AAE}$	.233	.140	.988	.941	.611	.540	.039	.147	.000	.980	.519	.564	.066	.141	**.994**	.960	.530	.550
MULTITASK	.578	.331	**.989**	**.949**	.783	.640	.082	.295	.999	.955	**.541**	**.625**	**.143**	.309	**.994**	.952	**.569**	.630
MULTITASK$_{+AAE}$	**.653**	**.375**	.988	**.949**	**.821**	**.662**	.063	.284	.999	.966	.531	**.625**	.113	**.322**	**.994**	.957	.553	**.639**

Table 14: In-group classification results on texts written with (*AA.*) and without (*NA.*) AAE dialect for the positive (*Positive*) and negative class (*Negative*), and macro averaged (*Macro*). All scores are averaged over five random seeds. Bold indicates best results per column.

Intent	True Positive Rate ↑			False Positive Rate ↓			Positive Predictive Value ↑		
Model	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**
SINGLETASK	.877	.910	.033	**.169**	**.352**	**.183**	**.851**	**.783**	.068
SINGLETASK$_{+AAE}$	.880	**.927**	.047	.175	.363	.188	.847	.781	**.066**
MULTITASK	**.894**	.924	**.030**	.184	.372	.188	.843	.777	**.066**
MULTITASK$_{+AAE}$	**.894**	.925	.031	.186	.374	.188	.842	.776	**.066**

Table 15: Results for the *intent* aspect per approach for True Positive Rates, False Positive Rates, and Positive Predictive Value, the elements of the fairness metrics *Equalized odds* and *Predictive parity*. All scores are averaged over five random seeds. Bold indicates best results in each column, arrows indicate whether higher (↑) or lower (↓) scores are better.

Lewdness	True Positive Rate ↑			False Positive Rate ↓			Positive Predictive Value ↑		
Model	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**
SINGLETASK	.739	.670	.069	.024	**.027**	**.003**	.757	**.783**	.026
SINGLETASK$_{+AAE}$	**.768**	.678	.090	.025	.032	.007	.756	.748	**.008**
MULTITASK	.749	.678	.071	**.022**	.030	.008	**.779**	.760	.019
MULTITASK$_{+AAE}$	.740	**.681**	**.059**	**.022**	**.027**	.005	.771	.781	.010

Table 16: Results for the *lewdness* aspect per approach for True Positive Rates, False Positive Rates, and Positive Predictive Value, the elements of the fairness metrics *Equalized odds* and *Predictive parity*. All scores are averaged over five random seeds. Bold indicates best results in each column, arrows indicate whether higher (↑) or lower (↓) scores are better.

Group	True Positive Rate ↑			False Positive Rate ↓			Positive Predictive Value ↑		
Model	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**
SINGLETASK	**.887**	**.859**	**.028**	.148	.326	.178	.806	.674	.132
SINGLETASK$_{+AAE}$	.873	.800	.073	.131	.269	.138	.821	.699	**.122**
MULTITASK	.866	.784	.082	**.126**	**.263**	**.137**	**.826**	**.700**	.126
MULTITASK$_{+AAE}$	.872	.788	.084	.131	.280	.149	.822	.687	.135

Table 17: Results for the *group* aspect per approach for True Positive Rates, False Positive Rates, and Positive Predictive Value, the elements of the fairness metrics *Equalized odds* and *Predictive parity*. All scores are averaged over five random seeds. Bold indicates best results in each column, arrows indicate whether higher (↑) or lower (↓) scores are better.

In-group	True Positive Rate ↑			False Positive Rate ↓			Positive Predictive Value ↑		
Model	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**	**¬AAE**	**AAE**	**Delta**
SINGLETASK	.000	.000	**.000**	**.000**	**.000**	**.000**	.000	.000	**.000**
SINGLETASK$_{+AAE}$	.039	.147	.108	**.000**	.020	.020	.233	.140	.093
MULTITASK	**.082**	**.295**	.213	.001	.045	.044	.578	.331	.247
MULTITASK$_{+AAE}$	.063	.284	.221	.001	.034	.033	**.653**	**.375**	.278

Table 18: Results for the *in-group* aspect per approach for True Positive Rates, False Positive Rates, and Positive Predictive Value, the elements of the fairness metrics *Equalized odds* and *Predictive parity*. All scores are averaged over five random seeds. Bold indicates best results in each column, arrows indicate whether higher (↑) or lower (↓) scores are better.

C.2 Social Bias Classification

The SINGLETASK model for the social bias classification was fine-tuned for three epochs on two A100-SXM4-80GB GPUs, using a batch size of 64. To increase training speed, we fine-tune the model using the DeepSpeed (Rajbhandari et al., 2020) integration of the Huggingface library (Wolf et al., 2020). With this setup, fine-tuning the model for a single aspect takes around 15 minutes.

The models trained with a multitask objective (SINGLETASK$_{+AAE}$, MULTITASK and MULTITASK$_{+AAE}$) were fine-tuned for three epochs on a single A100-SXM4-80GB GPU, usiing a batch size of 64. With this setup, fine-tuning takes around 50 minutes for SINGLETASK$_{+AAE}$, around 2 hours for MULTITASK, and around 3 hours for MULTITASK$_{+AAE}$.

We base our implementation of the multitask learning models on `https://github.com/shahrukhx01/multitask-learning-transformers`, as we found this to work notably better than alternative libraries.

For all models, we report results for a single training and inference run.

C.3 Significance Tests

Due to varying experimental settings, we employ different techniques to test for significance. Below, we describe and justify the applied testing methodology for each setting.

AAE Classification For the AAE classification presented in Table 1, we compare the results of the proposed classifiers to the TwitterAAE approach proposed by Blodgett et al. (2016). Since the code for the baseline is available, we are also able to retrieve per-sample predictions on the test dataset. Therefore, we calculate significance levels using a one-sided independent t-test, marked with † for $p < 0.05$ and ‡ for $p < 0.01$. Here, we employ a one-sided dependent paired t-test if the scores seem to be drawn from a normal distribution, and the Wilcoxon signed-rank test otherwise (as suggested in Dror et al. (2018), we test for normality using the Shapiro-Wilk test with $\alpha = 0.05$). To do so, we split test set of the TwitterAAE corpus (cf. Section 4.1) into ten random subsets (for the *Test smp* set described in Section 5 and Table 19, this results in 45,991 instances per subset), calculate precision, recall and F_1-score for each subset and use the score distribution as input to the t-test.

Dialect	Train	Validation	Test	Test smp
No-AAE	37,171,287	9,292,822	11,616,028	229,955
AAE	735,856	183,964	229,955	229,955
Total	37,907,143	9,476,786	11,845,983	459,910

Table 19: The number of instances per split in the TwitterAAE dataset. The *Test smp* column describes the numbers for the sampled test data used to evaluate the approaches, as presented in Section 5.

Overall Social Bias Classification For the overall bias classification presented in Table 2, we calculate two significance levels. First, we compare the results of the evaluated approaches to baselines from the literature, marked with † for $p < 0.05$ and ‡ for $p < 0.01$. As neither, the code nor per-sample predictions are available for either baseline at the time of writing, we employ a one-sample t-test. Since baseline scores are not available for the negative class and macro averaged F_1-scores, we cannot compute the significance over the baselines for results presented in Table 7 and Table 8.

Second, we compare the multitask approaches to their respective single-task variants, marked with * for $p < 0.05$ and ** for $p < 0.01$. Since we train five models with five different random seeds, we calculate the F_1-score for each model seed and use

	Train				Validation				Test			
	Positive		Negative		Positive		Negative		Positive		Negative	
Label	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**	**¬AAE**	**AAE**
Offensiveness	16294	2432	15286	1492	2281	331	1874	187	2342	368	1797	191
Intent	14795	2181	16785	1743	2109	306	2046	212	2171	326	1968	233
Lewdness	3092	497	28488	3427	365	56	3790	462	383	69	3756	490
Target Group	10624	1556	20956	2368	1581	234	2574	284	1690	245	2449	314
Ingroup	647	321	30933	3603	56	41	4099	477	4088	38	51	521

Table 20: The number of instances per split, label, and dialect in the Social Bias Inference Corpus (Sap et al., 2020). The dialect labels were inferred automatically using the approach presented in Section 3.

the score distribution as input to the significance test.

Per Dialect Social Bias Classification For the classification results per dialect presented in Table 3, we calculate two significance levels. First, we compare the multitask approaches to their respective single-task variants, marked with * for $p < 0.05$ and ** for $p < 0.01$. Second, we compare the results of the approaches that model AAE dialect to their respective non-AAE variants with † for $p < 0.05$ and ‡ for $p < 0.01$.

In both scenarios, we employ a one-sided dependent paired t-test. Since we train five models with five different random seeds, we calculate the F_1-score for each model seed and use the score distribution as input to the significance test.

D Dataset Details

D.1 TwitterAAE

Detailed dataset statistics of the TwitterAAE Corpus (Blodgett et al., 2016) are reported in Table 19.

D.2 Social Bias Inference Corpus

Detailed dataset statistics of the Social Bias Inference Corpus (Sap et al., 2020) are reported in Table 20. Dialect labeles are inferred automatically using the AAE_{wgh} approach presented in Section 3.