TIDE: Textual Identity Detection for Evaluating and Augmenting Classification and Language Models

Anonymous ACL submission

Abstract

Machine learning models can perpetuate unintended biases from unfair and imbalanced datasets. Evaluating and debiasing these datasets and models is especially hard in text datasets where sensitive attributes such as race, gender, and sexual orientation may not be available. When these models are deployed into society, they can lead to unfair outcomes for historically underrepresented groups. In this paper, we present a dataset coupled with an approach to improve text fairness in classifiers and language models. We create a new, more comprehensive identity lexicon, TIDAL, which includes 15,123 identity terms and associated sense context across three demographic categories. We leverage TIDAL to develop an identity annotation and augmentation tool that can be used to improve the availability of identity context and the effectiveness of ML fairness techniques. We evaluate our approaches using human contributors, and additionally run experiments focused on dataset and model debiasing. Results show our assistive annotation technique improves the reliability and velocity of human-in-the-loop processes. Our dataset and methods uncover more disparities during evaluation, and also produce more fair models during remediation. These approaches provide a practical path forward for scaling classifier and generative model fairness in real-world settings.

1 Introduction

The growing adoption of machine learning across a variety of applications have reignited concerns about unfair and unintended bias in models. Bias can be introduced throughout the development workflow, for example during problem framing, data sampling and preparation, and even through training algorithm choices (Shah et al., 2020; Saleiro et al., 2018). When models contain biases, they can play an active role in perpetuating societal inequities and unfair outcomes for underrepresented groups (Sweeney, 2013; Abid et al., 2021).

Algorithmic fairness is a rapidly growing field of research with a wide range of definitions, techniques and toolkits available. Fairness is anchored in understanding and mitigating model performance disparities across sensitive and protected attributes. Popular toolkits such as AI Fairness 360 (Bellamy et al., 2018), Fairlearn (Bird et al., 2020), and the Responsible AI toolkit in TensorFlow (Abadi et al., 2015), all assume these attributes are readily available in datasets. In many real-world datasets, attributes are either not available or not reliable. This is due to a myriad of issues like privacy and safety laws around protected attributes, human annotation cost and reliability, and inconsistent taxonomy and attribute coverage (Andrus et al., 2021).

Attempts to address this problem involve techniques to extract attributes from text, through human or computational means. A common one is to create an adhoc list of "identity terms" (Dixon et al., 2018) for token matching. However this approach is limited due to the polysemy of words (e.g. "black" as a color or race), scalability of token matching techniques, and a lack of important contextual information about the terms (Blodgett et al., 2020). Connotation is one such example of missing context: a non-literal meaning of a word informed by one's beliefs and prejudices about its typical usage (e.g. "undocumented workers" and "illegal aliens" have the same lexical denotation but different connotations) (Carpuat, 2015; Allan, 2007; Webson et al., 2020).

Our research goal is to first explore techniques that can improve availability and reliability of identity term annotations by providing context for disambiguation. A second goal is to leverage these annotations to adapt existing fairness techniques in ways that scale for use in real-wold text datasets and throughout the development workflow.

1.1 Related Work

1.1.1 Availability of identity labels.

Gupta et al.; Jung et al. propose methods to leverage proxy attributes in the absence of identity labels, however Tschantz; McLoughney et al. show proxies could be a source of bias and discrimination. When labels exist but are noisy or unreliable, Celis et al. explore techniques to achieve fairness under uncertainty. Lahoti et al. attempt to remove the need for identity labels altogether. Our work follows Andrus and Villeneuve (2022), focusing on addressing the issue earlier in the pipeline by taking a human-in-the-loop approach. We deploy assistive techniques for acquiring high quality annotations from humans faster.

1.1.2 Identity lexicon.

(Eckle-Kohler et al., 2012) show the need for a standardized lexicon, while (Allaway and McKeown, 2021) extend one with contextual dimensions including sentiment and emotional association. Our approach is most closely related to (Smith et al., 2022) who create a similar identity lexicon. We focus on creating an extensible schema that enables multilingual support, and enabling fairness use cases by capturing additional context and increasing the depth of coverage across groups

1.1.3 Identity entity recognition.

Sense disambiguation (Pal and Saha, 2015) has been used to address polysemy, with recent advances in knowledge-based techniques (Agirre et al., 2014). On the other hand (Honnibal and Montani, 2017; Bird et al., 2009) use syntactic and NLP techniques to detect canonical entities like "person", which is too coarse. Our work merges both techniques to build a reusable annotation tool. We specialize in identity detection and optimize for fairness workflows, and additionally adapt for counterfactual generation.

1.1.4 Effectiveness of fairness techniques.

(Dixon et al., 2018) use a keyword list to source new organic data for debiasing datasets, while (Wadhwa et al., 2022) generate counterfactuals using existing datasets as the seed. Our experiments aim to scale up both fairness techniques for use throughout the entire ML workflow. We also leverage identity taxonomy instead of terms to uncover previously missed bias in classifiers and generative models alike.

1.2 Contributions

Our key contributions are summarized below:

- Textual Identity Detection and Augmentation Lexicon (TIDAL)[1]: to the best of our knowledge TIDAL is the largest identity lexical dataset with comprehensive coverage of groups and associated sense context, using a methodology and schema that supports multiple languages.
- A specialized identity annotation tool built with the lexicon and optimized for multiple fairness workflows.
- An assistive technique for human annotation that improves time, cost and reliability of acquiring identity labels.
- Updated fairness techniques that improve coverage of bias detection and result in more effective remediation of datasets and models.

1.3 Preliminaries

1.3.1 Datasets.

We use the CivilComments dataset (Borkan et al., 2019) for most experiments conducted, relying on its human-annotated identity labels as ground truth. We use the C4 dataset (Raffel et al., 2020) as a control.

1.3.2 Data Augmentation.

We generate synthetic datasets using sentence templates from HolisticBias (Smith et al., 2022) and UnintendedBias (Dixon et al., 2018). We additionally generate counterfactuals (Wadhwa et al., 2022) for robustness.

1.3.3 Models.

For generative tasks we use BlenderBot (Roller et al., 2021). For classification we train toxicity models on CivilComments, and additionally use counterfactual logit pairing (CLP) for remediation.

1.3.4 Dataset and model evaluation metrics.

We use slice analysis and deficits to understand class balance in datasets and models (Dixon et al., 2018). We measure model performance using F1, area-under-curve (AUC), and counterfactual flips (Garg et al., 2019) for classifiers, and token likelihood (Smith et al., 2022) for generative models.

[1] Dataset will be made available after review and acceptance

1.3.5 Inter-annotator reliability (IAR).

Following (Lacy et al., 2015), we use simple percent agreement, Krippendorff's alpha (Krippendorff, 1970) and Gwet's AC1 (Gwet, 2014) to measure the degree of agreement on annotations between human annotators. While Krippendorff's alpha penalizes for data scarcity, Gwet's AC1 corrects for the probability that the annotators agree by chance - both cases are likely given our data distribution and task complexity.

1.3.6 Identity terms and sense context.

Multiple descriptors are used throughout the literature to describe words, utterances or context associated with identity, such as "sensitive attributes", "sensitive features", "group labels", "protected attributes" or "identity terms" (Garg et al., 2019; Dixon et al., 2018). In our work we use "identity terms" for the lexicon that appears in text, and "sense context", for the structured contextual data associated with senses of identity terms.

2 Methodology

2.1 TIDAL dataset

The TIDAL dataset consists of lexical entries and their related forms (e.g. black, gay, trans, hindus) that are associated with identity groups. Each head and related form is associated with grammatical properties (e.g. part-of-speech, grammatical gender) and context (or "sense") entries (e.g. identity groups/subgroups, connotation). Although we develop a lexicon, schema and methodology that works for multiple languages, we will focus on English in this paper. In total TIDAL has 1,419 English language head-form identity lexical entries, with over 13,709 related lexical forms and 15,270 context/sense entries.

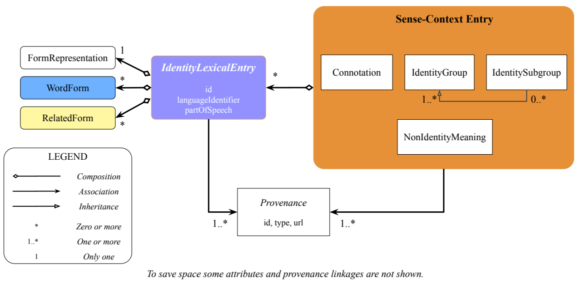

Figure 1: TIDAL: Conceptual model

2.1.1 Schema.

Figure 1 shows the conceptual model of the TIDAL schema and Figure 2 shows a flattened tabular example of TIDAL data. We create an adapted UBY-LMF schema (Eckle-Kohler et al., 2012) which is based on the Lexical Markup Framework (LMF) standard (for Standardization, 2022) for representing NLP lexicons.

IdentityLexicalEntry			WordForm	Sense-Context Entry				RelatedForm	
				type=Identity			type=NonIdentity		
FormRepresentation _writtenForm	language Identifier	partOfSpeech	number	**Connotation**	**Group**	**Subgroup**	HasNonIdentity Meaning	relType	lexicalEntry_ writtenForm
trans	en	NOUN	Singular	**NEUTRAL**	**SOGIESC**	**GenderIdentity > Transgender**	TRUE		
trans folks	en	NOUN + NOUN	Plural	**NEUTRAL**	**SOGIESC**	**GenderIdentity > Transgender**	FALSE	**PersonNoun CombinationOf**	**trans**

Examples created for illustrative purposes only.

Figure 2: TIDAL: Example, flattened tabular format.

Our paper focuses on the following identity groups (IdentityGroup): race, nationality or ethnicity (RNE), sexual orientation, gender identity, gender expression and sex characteristics (SOGIESC) and Religion. We choose RNE as a collective category to be more inclusive since their constituent concepts of race, ancestry, nationality and ethnicity are inconsistent and sometimes redundant across cultures (Morning, 2008). We choose SOGIESC for similar reasons, instead of Gender Identity and Sexual Orientation, LGBTI or SOGI (Trithart, 2021). Although multiple dimensions of connotation like social value, politeness or emotional association have been proposed in prior lexical work (Allaway and McKeown, 2021), our scope is limited to NEUTRAL and PEJORATIVE connotations. PEJORATIVE implies a term can be used to demean or disparage a group of people.

Table 1 shows a comparative analysis of TIDAL with known similar sources such as UnintendedBias (Dixon et al., 2018) used by Perspective API [2], and HolisticBias (Smith et al., 2022). Additional details of our data distribution can be found in Appendix A.3.

2.1.2 Sourcing.

We source the seed set of identity terms for our lexicon from the following public sources:

- **UNdata** (UNSD, 2003): "Population by national and/or ethnic group" and "Population by religion" tables from UNData are used to create RNE and Religion seed sets, respectively.

- **CAMEO** (Gerner et al., 2002): We utilize the CAMEO coding framework, which contains

[2]https://perspectiveapi.com/

approximately 1,500 religions and 650 ethnic groups.

- **GLAAD**: We leverage GLAAD glossary of LGBTQ and transgender terms (GLAAD) for SOGIESC seed sets.
- **HRC**: We use HRC glossary of words and meanings (HRC Foundation) for SOGIESC seed sets.
- **Wikipedia**: We leverage demonyms and adjectivals (Wikipedia contributors, 2023) list for RNE seed sets.

Appendix A.2 provides additional details on seed set data processing.

2.1.3 Curation.

We expand the seed terms to their grammatical and morphological variants using linguistic experts and rule-based lexical expansion tools. Each resulting term is treated as a new lexical entry with reference to the head. Next we curate multiple pools of data contributors to corroborate, correct and expand our data. We leverage a human annotation platform to curate a diverse pool of linguistic experts and create tasks reflecting the following phases:

1. **Expansion**: expand seed terms to grammatical variants, common misspellings and person noun combinations.
2. **Contextualization**: research and associate all possible context for seed terms and expansions, including connotation and identity groups.
3. **Disambiguation**: research and associate context that can help distinguish identity and prevalent non-identity usage of the terms.

Contributors research public sources (such as dictionaries, encyclopedias, and other lexical sources) for unstructured context for identity terms. They also provide citations for the sources they use, their own beliefs about missing context or usage of a term not available in sources. Finally, we anonymize contributor personally-identifiable information before aggregating the assertions and ingesting the data into the lexicon database.

	HolisticBias	UnintendedBias	**TIDAL**
Supported Identity Groups	14	N/A	**3**
Head terms / lexical entries	594	50	**1565**
Variants and expansions	-	-	**14148**
Includes connotation context	No	No	**Yes**
Includes identity groups/subgroups	Yes	No	**Yes**
Includes non-identity context	No	No	**Yes**

Table 1: Comparison of TIDAL to other lexicons datasets.

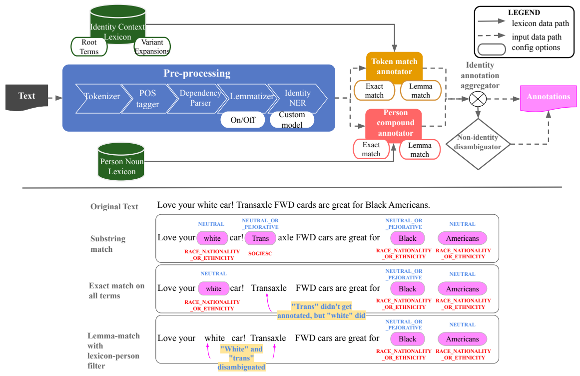

Figure 3: Data flow and system components of the annotation tool, with examples.

2.2 Identity Annotation Tool

To scale the acquisition of identity labels, we build a configurable multi-label multi-class annotation tool that leverages our identity lexicon and lexical properties to label identity terms found in text.

2.2.1 Annotator components.

We first preprocess text using spaCy (Honnibal and Montani, 2017) to tokenize and tag with part-of-speech labels, the dependency tree and morphological properties. We then match tokens with terms in the lexicon, using lemmas and variants. We disambiguate non-identity usage of terms with person-noun detection using i) a lexicon of person nouns from Wiktionary (Wiktionary contributors, 2021) and ii) the NLTK (Bird et al., 2009) wordnet module to compare similarity with person identifiers like “person” and “people” and non-person identifiers like “object” and “thing”. Additionally, spaCy linguistic features (Honnibal and Montani, 2017) is used for person-nouns detection using named entities like “PERSON”, “NORP”, and “GPE”. To disambiguate a potential identity term we use the dependency tree (with support for conjunctions) and part-of-speech tags to include tokens that modify person-nouns and exclude tokens that modify non-person nouns. Finally, we train a custom spaCy NER model. The output of the annotator includes identity groups, subgroups, connotation and possible non-identity usage. Figure 3 shows the annotation flow and example output. Additional design details are specified in Appendix B.1.

3 Acquiring Identity Context at Scale

3.1 Annotation Tool Performance

We measure the performance of our annotation techniques against human annotations available in the CivilComments dataset, and additionally validate performance consistency using the C4 dataset as a control. Our goal is to understand the effectiveness of techniques for a variety of downstream tasks, and whether performance can generalize to new datasets.

3.1.1 Annotation techniques.

We implement substring matching as the baseline technique and configure multiple annotator variants using tokenizers: i) tokenize and match any occurrence in the lexicon, including all term forms and expansions; ii) tokenize and match occurrence of head terms only; iii) a variation of ii) that additionally disambiguates using a person-term lexicon; and iv) a variation of iii) that uses similarity-to-person-term disambiguation. We finally configure the custom NER model as a standalone annotator variant. Across all techniques, only annotations matching lexical entries in the dataset are considered valid. Figure 3 shows examples of annotation output.

3.1.2 F1 scores.

All techniques outperform substring matching, with the custom NER model achieving the highest score of 91.92%, followed by lemma and exact matching (91.13%, 91.11%) in Figure 4. Disambiguation filters result in increased false negatives that impact overall performance. RNE has the lowest performance trend among subgroups while Religion has the most similar performance across techniques. Additional performance details are provided in Appendix B.2.

3.2 Human Annotation Impact

We assess the impact of assistive annotation in human annotation workflows used to acquire identity labels. In addition to time and cost improvements

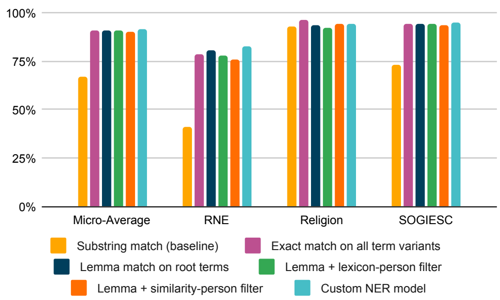

Figure 4: Multi-class F1 scores for the identity annotation tool on CivilComments.

we seek to understand the quality and consistency of human annotations, including potential new biases.

3.2.1 Methodology.

We sample 337 examples from the CivilComment dataset annotated in the previous experiment. This example dataset is balanced across groups and highlights the performance differences between annotator variants. We present these examples in a human computation task for contributors to first identify tokens associated with identity and then provide an appropriate IdentityGroup label (RNE, Religion or SOGIESC). From a pool of more than 1,000 human annotators, at least 5 annotators review each example. We run three variations of this human annotation task, i) the first with an example-only dataset as the baseline, and the others with assistive annotations: ii) using a token-matching annotator without disambiguation, and iii) using a token-matching annotator with disambiguation. We also request an optional satisfaction survey for each task where the human annotators are asked to rate “Ease of Job” and “Pay”. We run the same set of experiments on the C4 dataset as a control. Detailed human annotation job design and guidelines can be found in Appendix B.3.

3.2.2 Inter-annotator reliability (IAR).

Assistive annotations consistently improve the reliability of human annotations as seen in Figure 5. Token-matching achieves an Gwet’s AC1 score of 0.7622, representing a 89.27% increase over the baseline, while additional disambiguation results in a score of 0.6257, a 55.37% increase. Our analysis finds similar improvement trends in percent agreement and Krippendorff’s Alpha metrics. Additional results are available in Appendix B.4.

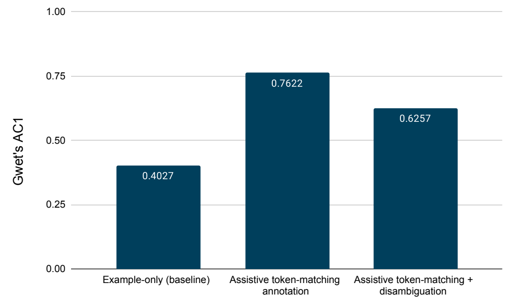

Figure 5: IAR (Gwet's AC1) for human annotations: identity labeling on CivilComments.

3.2.3 F1 scores.

Since IAR doesn't provide a per-class understanding of agreement and quality, we use micro-average F1 scores to understand performance across groups. We use the output of the baseline annotation task (example-only) as ground truth for this comparison. Token-matching achieves the highest overall score of 87.38%, while additional disambiguation performs better only for Religion, seen in Figure 6. Further analysis reveals tradeoffs between false positives and false negatives across the two annotation techniques. More details are in Appendix B.4.

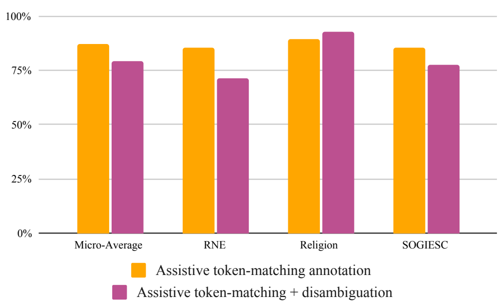

Figure 6: Multi-class F1 scores for human annotations: identity labeling on CivilComments.

3.2.4 Velocity, cost and satisfaction scores.

We use the interquartile mean (IQM) of time taken for a human annotator to complete the tasks as a proxy for completion velocity. To understand cost, we count the total number judgements required to meet the agreement threshold of 0.7. Lastly, the results from a task satisfaction survey inform task completion difficulty. Token-matching performs the best on velocity, taking 44.8% less time than the baseline. Both assistive annotations tasks have similar costs (24-27% better compared to the baseline). While we receive no data on satisfaction for token-matching, contributors find assistive annotations with disambiguation makes tasks 84.4% easier to perform and result in 43.4% better pay to the baseline task. Table 2 provides detailed per task scores.

	Velocity	Cost	Ease of Job	Pay
	Judgement Time (s)	Total Judgements	Scale: 1-5	Scale: 1-5
Example-only (baseline)	82.5	2623	2.25	3
Assistive annotations using token-matching	45.5	1981	-	-
Assistive annotations with disambiguation	64	1905	4.15	4.3

Table 2: Velocity, cost and satisfaction results from human annotation tasks for identity labels

4 Fairness Applications

Our experiments in this section explore opportunities to leverage our lexicon and annotation tool at various points in the ML fairness workflow, from data labeling to model training. We modify and augment existing techniques from the literature in ways that are only enabled by our work. Our goal is to improve overall effectiveness of fairness interventions and demonstrate that it can be done at scale.

4.1 Assistive Context for Ground Truth Labeling

We explore data collection interventions by replicating the toxicity labeling human annotation task[3] for the Perspective API. Figure 7 shows an example of the assistive annotations we provide during human computation to understand the impact of context on annotation quality.

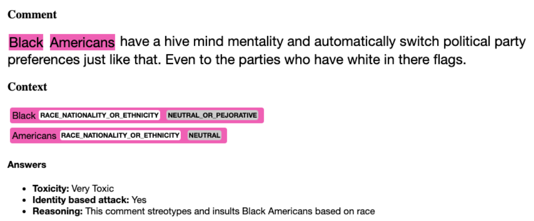

Figure 7: Example of identity context annotation in HCOMP toxicity labeling task.

4.1.1 Methodology.

We modify their human computation setup by excluding all sub-attributes except "Identity based attack", which we show only when the toxicity question is answered with "VERY TOXIC", "TOXIC"

[3] https://github.com/conversationai/conversationai.github.io

or "NOT SURE". We sample 298 examples from the CivilComment dataset annotated in the previous experiment, only including examples where our annotations are an exact match with provided ground truth labels. This example dataset is balanced across groups and is representative of the performance differences between annotator variants. We run three variations of the human evaluation task, i) the first with an example-only dataset as the baseline, and the others with assistive identity context: ii) providing "IdentityGroup" annotations, and iii) providing "IdentityGroup" and "Connotation" annotations. From a pool of more than 1,300 human annotators, at least 10 annotators review each example. Detailed human annotation job design and guidelines are given in Appendix C.3.

4.1.2 Inter-annotator reliability (IAR).

Assistive annotations consistently improve the reliability of human annotations as seen in Figure 8. IdentityGroup+Connotation annotations achieve the highest AC1 score, seeing an 14.04% increase over the baseline, IdentityGroup annotations achieve an 9.96% increase over baseline. Krippendorff's Alpha scores have the lowest trend due to class imbalance - 85% of labels are toxic. Our agreement performance is consistent with prior work ((Ross et al., 2016) and (Wulczyn et al., 2017)), given the subjective nature of toxicity labeling. Additional results are in Appendix C.4.

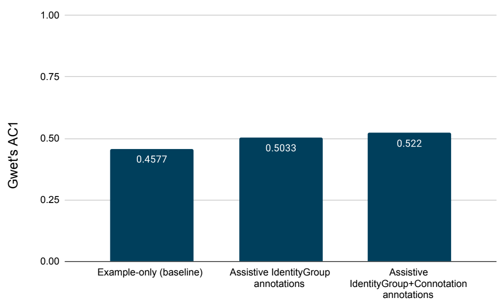

Figure 8: IAR for human annotations: toxicity labeling on CivilComments.

4.2 Counterfactual Logit Pairing

We replicate the experimental setting from the counterfactual logit pairing (CLP) guide[4], and introduce additional counterfactual techniques enabled by our work to evaluate and mitigate classifier bias.

4.2.1 Counterfactual techniques.

We establish a baseline with token ablation using their keyword list. We implement two additional techniques: i) token ablation using subgroup annotations instead of keywords and ii) token replacement using least similar counterfactuals. We train CLP-remediated models for each technique and evaluate flips on the baseline test set. Additional details in Appendix C.2.

4.2.2 Counterfactual flip rates.

The counterfactual flip rate diff metric measures the difference between the flip rate for a counterfactual model and that of the base model on the baseline counterfactual dataset. Results show that using annotations for ablation instead of a keyword list increases the coverage of terms, leading to consistently fewer counterfactual flips in Table 3. We also observe that the counterfactual ablation technique performs better than replacement since ablation creates only one counterfactual compared to multiple generated with replacement technique. Mitigating using counterfactual replacements requires generating multiple counterfactuals for better chances of success, which we'll observe in the next section. The CLP library also only supports generating one counterfactual which limits the coverage of counterfactual evaluation and remediation.

	Overall	Black	Homosexual	GenderIdentity
Keyword ablation (baseline)	0.37%	0.27%	-0.30%	0.32%
Annotation ablation	0.08%	-0.09%	-0.74%	0.00%
Annotation replacement	0.34%	0.36%	-0.30%	0.26%

Table 3: Difference in counterfactual flip rates per technique on CivilComments compared to the original model.

4.3 Dataset Debiasing

We replicate the experimental setting from (Dixon et al., 2018) to evaluate dataset and model bias. We additionally augment their data augmentation techniques and introduce counterfactual generation to improve effectiveness of data debiasing and model remediation.

4.3.1 Data debiasing techniques.

We use their keyword list as a baseline to understand toxicity rates, compute subgroup rate deficits

[4]https://www.tensorflow.org/responsible_ai/model_remediation/counterfactual/guide/counterfactual_keras

and source non-toxic examples from Wikipedia article snippets for debiasing. We implement two additional techniques: i) sourcing using subgroup annotations instead of keywords and ii) generating five least similar counterfactual examples per label. We train a model per augmented dataset and evaluate classification performance on a templated synthetic dataset. Additional details can be found in Appendix C.1.

4.3.2 Dataset toxicity rates and model AUC.

Annotation-driven data sourcing increases the coverage of terms compared to the keyword list, leading to more balanced toxicity rates across subgroups. Counterfactual augmentation increases per-label term diversity, resulting in the highest AUC scores and the most equality across subgroups in Figure 9. Toxicity rate balance from annotations translates to equality in model performance across subgroups, but with lower overall performance.

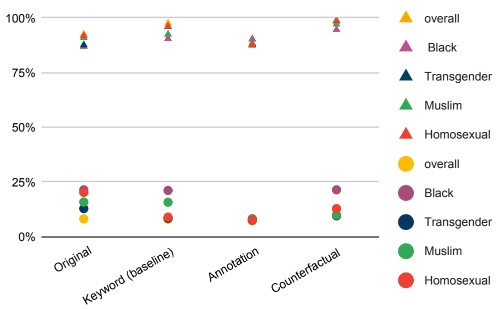

Figure 9: Model AUCs (triangles) and dataset toxicity rates (circles) per debiasing technique on a synthetic dataset. A tighter cluster pattern indicates less bias across subgroups.

4.4 Generative Model Bias

We replicate the experimental setting from (Smith et al., 2022) to evaluate generative model bias, leveraging our lexicon to expand the coverage of bias detection.

4.4.1 Dataset generation.

We create two datasets: i) a baseline dataset using the templates and lexicon from HolisticBias and ii) a new dataset using our lexicon with the same templates. We generate perplexity scores by running evaluations of the 90M-paremeter BlenderBot model on both datasets.

4.4.2 Token likelihood bias.

Our lexicon's deeper coverage of terms reveals a broader bias in token likelihoods for RNE in Figure 10. SOGIESC and Religion have a much smaller vocabulary as seen in Appendix A.3, thus are not as prone to coverage issues.

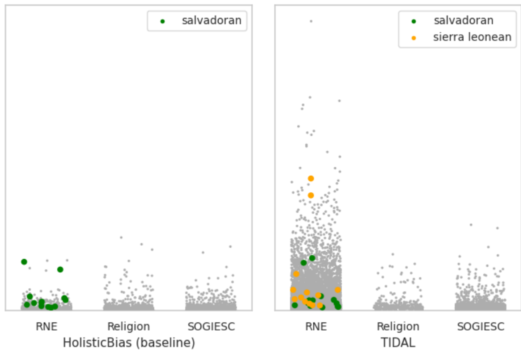

Figure 10: Generative model perplexities on a synthetic dataset, with a max of 6000. Our lexicon shows an example of a previously missed term.

5 Conclusion

We create a new identity lexicon, TIDAL and use it to develop an annotation tool for textual identity detection and augmentation. Through our experiments we demonstrate the effectiveness of our work to scale and improve existing human annotation and fairness techniques.

When coupled with a comprehensive lexicon that includes term forms and expansions, token-matching emerges as the most practical annotation technique given its implementation simplicity and low computational cost. We note that a custom NER model results in computational speed gains, but requires training resources and ground truth annotations. We demonstrate improvements in human annotation reliability and cost, positioning our annotator as an assistive tool for acquiring identity labels from contributors.

To scale fairness in practice, we build on our work to advance techniques used throughout the machine learning workflow. We demonstrate how to increase reliability in human annotations of ground truth, uncover more bias in data than previously known and train more fair models using improved techniques. We find that our approaches can be leveraged across different notions of fairness, ML development stages and model types.

6 Limitations

Our current lexicon is limited in a number of ways due to the scope of the paper. We propose future work to increase the number of represented identity groups and subgroups. The scope of terms can be expanded to include non-literal associative words (e.g. "temple" for Religion), compound phrases that imply an identity group (e.g. "same-sex marriage" for SOGIEC), and prevalent stereotypes (e.g. "kinky hair" for RNE), all the while considering intersectionality. Coverage of contextual dimensions (Appendix A.3) can be improved for balance across groups. Additional sense context can also be added to improve disambiguation, for example by integrating with other lexical-semantic datasets such as WordNet and Wiktionary (Eckle-Kohler et al., 2012) as shown in Appendix A.1

Token-based techniques presented are limited due to complexity of identity, contextual interpretation and fluidity of language. In addition to NLP, advanced knowledge-based approaches (Agirre et al., 2014) need to be explored for disambiguated identity detection. Generative techniques like DataSynth[5] hold a lot of promise for counterfactual generation. All of these require expanding the lexicon to include more "sense context" as mentioned above.

Our results show that trade-offs are required in fairness depending on use case and type of bias, as techniques have different impacts in datasets and models (Goldfarb-Tarrant et al., 2021). While our experiments use techniques independently, we propose future work to examine mixed-method approaches to improve guidelines for practical settings.

Finally, our goal is to incorporate sense context from many perspectives, however crowd-sourcing does not explicitly advance this goal. Contributor diversity, task sensitivity and a lack of benchmarks all impact representation and perceived quality. Future work on identity datasets should explore participatory data collection and governance models to empower groups to not only shape how they're represented, but also where and how their data is used.

7 Ethical Statement

During our research we encounter a variety of questions, including how to collect identity context data ethically, how assistive context could bias human annotations, and what the right compensation for those tasks should be.

We acknowledge that there are a lot more demographic categories and context than we choose to focus on in this paper. This means the work presented does not mitigate bias for everyone. Given our limited scope there is a high risk of misrepresentation and disenfranchisement especially of historically underrepresented groups.

We recommend caution when generalizing our findings to non-English languages or even across different cultures and groups given the subjectivity of identity assertions and toxicity labels.

7.1 Wellness in Human Evaluation

Toxicity labeling has a side-effect of exposing human annotators and researchers to toxic languages, something we experience first-hand during our work. We only select contributors that accept explicit content (Appen, a) on the Appen platform.

We also leverage the Fair Pay plugin (Appen, c) to ensure that each contributor is fairly compensated based on their geographical location, with an extra 50% pay increase over the suggested baseline to account for task complexity.

[5]https://github.com/Tobiadefami/datasynth

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Keith Allan. 2007. The pragmatics of connotation. *Journal of Pragmatics*, 39:1047–1057.

Emily Allaway and Kathleen McKeown. 2021. A unified feature representation for lexical connotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational*

Linguistics: Main Volume, pages 2145–2163, Online. Association for Computational Linguistics.

McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. 2021. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 249–260.

McKane Andrus and Sarah Villeneuve. 2022. Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, page 1709–1721.

Appen. a. Guide to: Contributors channels page. [Online; accessed 28-June-2023].

Appen. b. Guide to: Designating test questions for quiz or work mode with gold pool. [Online; accessed 28-June-2023].

Appen. c. Guide to: Fair pay. [Online; accessed 28-June-2023].

Appen. d. Machine learning assisted text utterance collection. [Online; accessed 28-June-2023].

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *Computing Research Repository*, abs/1810.01943.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491–500.

Marine Carpuat. 2015. Connotation in translation. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, Lisboa, Portugal. Association for Computational Linguistics.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1349–1361. PMLR.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF – a uniform model for standardizing heterogeneous lexical-semantic resources in ISO-LMF. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 275–282, Istanbul, Turkey. European Language Resources Association (ELRA).

International Organization for Standardization. 2022. Language resource management — lexical markup framework (lmf) — part 5: Lexical base exchange (lbx) serialization. ISO Standard No. 24613-5:2022.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226.

Deborah J. Gerner, Rajaa Abu-Jabr, Philip A. Schrodt, and Ömür Yilmaz. 2002. Conflict and mediation event observations (cameo): A new event data framework for the analysis of foreign policy interactions. *International Studies Association*. Paper presented at the 43rd Annual Convention of the International Studies Association, New Orleans, March 2002.

GLAAD. Glossary of terms: Lgbtq - glaad. [Online; accessed 28-June-2023].

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Maya R. Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Lutong Wang. 2018. Proxy fairness. *Computing Research Repository*, abs/1806.11212.

Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

HRC Foundation. Glossary of terms - human rights campaign. [Online; accessed 28-June-2023].

Jigsaw. 2019. Jigsaw unintended bias in toxicity classification. [Online; accessed 28-June-2023].

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. 2022. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10348–10357.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Stephen Lacy, Brendan R. Watson, Daniel Riffe, and Jennette Lovejoy. 2015. Issues and best practices in content analysis. *Journalism & Mass Communication Quarterly*, 92(4):791–811.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. 2020. Fairness without demographics through adversarially reweighted learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 728–740.

Aidan James McLoughney, Jeannie Marie Paterson, Marc Cheong, and Anthony Wirth. 2023. 'emerging proxies' in information-rich machine learning: a threat to fairness? In *2023 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS)*, pages 1–1.

Ann Morning. 2008. Ethnic classification in global perspective: A cross-national survey of the 2000 census round. *Population Research and Policy Review*, 27:239–272.

Alok Ranjan Pal and Diganta Saha. 2015. Word sense disambiguation: A survey. *International Journal of Control Theory and Computer Modeling (IJCTCM)*, 5(3).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9. Ruhr-Universitat Bochum.

Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *Computing Research Repository*, abs/1811.05577.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Albert Trithart. 2021. A un for all?: Un policy and programming on sexual orientation, gender identity and expression, and sex characteristics. Technical report, International Peace Institute.

Michael Carl Tschantz. 2022. What is proxy discrimination? In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 1993–2003, New York, NY, USA. Association for Computing Machinery.

UNSD. 2003. Ethnicity: A review of data collection and dissemination. Technical report, United Nations Statistics Division.

Mohit Wadhwa, Mohan Bhambhani, Ashvini Jindal, Uma Sawant, and Ramanujam Madhavan. 2022. Fairness for text classification tasks with identity information data augmentation methods. *Computing*

Research Repository, abs/2203.03541. Paper presented at the Measures and Best Practices for Responsible AI workshop at the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Virtual/Singapore, August 2021.

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

Wikipedia contributors. 2023. List of adjectival and demonymic forms for countries and nations — Wikipedia, the free encyclopedia. [Online; accessed 28-June-2023].

Wiktionary contributors. 2021. Category - english terms of address — Wiktionary, the free dictionary. [Online; accessed 28-June-2023].

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Appendices

Warning: Following sections may include terms or generated sample text that may be considered offensive or toxic.

A TIDAL Dataset

A.1 Schema Design

We leverage and customize the UBY-LMF model (Eckle-Kohler et al., 2012) for its comprehensiveness and extensibility in supporting lexical, semantic and pragmatic properties of words and phrases. Figure 11 shows a simplified Entity-relationship diagram (ERD) of the lexicon schema using UML notation. The grayed out entities and relationship are not in scope of this paper, but are shown in the diagram to support the extensibility argument for choosing the UBY-LMF model for our schema.

This schema allows us to model lexical information types in detail, including morphology, syntax, semantic and pragmatic arguments. It also enables standard-compliant sense alignments between other lexical sources. We define subclasses for the Context class, allowing us to model the context as a subclass of Sense entry associated with the lexical entry.

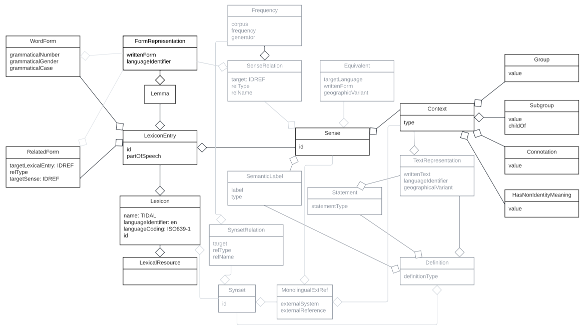

Figure 11: Simplified Entity-relationship diagram of the lexicon schema using UML notation.

A.2 Post-processing of seed set

All seed set sources are cleaned-up by lowercasing the data, removing punctuations, numbers, extra space, hyphens and back/forward slashes. Additionally for RNE category, we use country names from Wikipedia as a filter to remove terms which could be country names from the seed set. All sources are then aggregated and provenance of sourcing and post-processing is stored along with the seed term set. These seed terms then form the lexical entries for our lexicon.

A.3 Data distribution of TIDAL

Table 4 shows the distribution of TIDAL across IdentityGroups while Table 5 shows the distribution across Connotation context.

	Total	RNE	Religion	SOGIESC
All Entries	15123	13762	355	1046
Head Entries	1277	1278	25	121
Person Noun Compound Entries	10090	9256	260	600
Other Related Form Entries	3592	3233	70	299

Table 4: TIDAL: Head Lexical entry and Related form distribution by IdentityGroup.

	Total	RNE	Religion	SOGIESC
All Entries	15123	13762	355	1046
NEUTRAL	15031	13734	355	1054
PEJORATIVE	216	113	34	137
BOTH	124	30	17	60

Table 5: TIDAL: Connotation distribution by IdentityGroup.

B Acquiring Identity Context

B.1 Annotation tool design details

Training and test data preprocessing.

We use the “train” split of the CivilComments dataset because other splits do not have identity annotations. We only include identity and toxicity labels where rater agreement is greater than 0.5. We then partition the dataset using a 3-1 ratio for training (75%) and test (25%) data. The test data partition is then used for evaluation of annotators. For C4, we use the “validation” split for evaluation of annotators.

Custom NER model training.

During qualitative analysis we observe some incorrect human-annotated labels on the CivilComment dataset. To ensure high-quality training data, we first annotate CivilComments using the exact-match annotator. We only use a label set as ground truth when the annotation tool matches human-annotated labels. We train a spaCy pipeline for 11 epochs with a 50% dropout rate.

B.2 Annotation tool results

False Positives/False Negatives.

Analysis in Table 6 shows a false positive, false negative tradeoff between token-matching and token-matching with disambiguation for RNE and SOGIESC. We however observe consistent false negatives for Religion across all annotators except exact-matching.

	RNE		Religion		SOGIESC	
	FP	FN	FP	FN	FP	FN
Substring match (baseline)	24665	10	424	1572	15657	34
Exact match on all term variants	4523	27	206	804	2511	125
Lemma match on head terms	4079	32	197	1298	2270	103
Lemma+lexicon-person filter	2764	586	164	1571	2214	240
Lemma+similarity-person filter	3309	570	173	1126	2456	279
Custom NER model	3421	38	193	1217	2185	131

Table 6: Multi-class false positive (FP) and false negative (FN) counts for the annotation tool on CivilComments

C4 annotation tool performance as control

We corroborate annotation tool performance using a different dataset. We use C4 as the control,

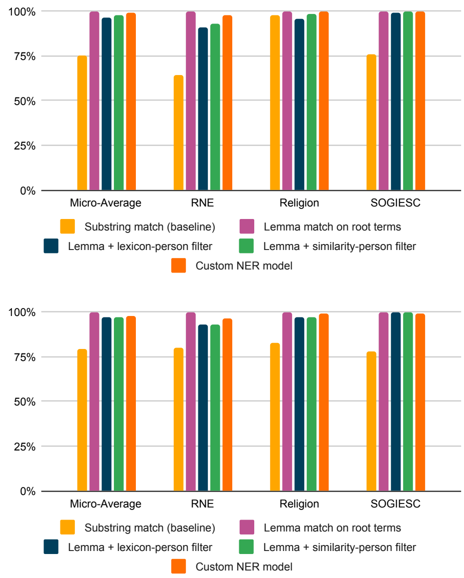

Figure 12: Multi-class F1 scores for the annotation tool on CivilComments (top) and the C4 (bottom).

however it lacks human-annotated labels, so for consistency we treat the exact-match annotator as ground truth for both datasets in this evaluation. Results show similar performance on both datasets. Figure 12 shows overall and per-class performance for the annotators on both datasets.

B.3 Human computation design details

The identity context human computation task is designed to be completed by crowd-sourced contributors on the data annotation platform. In the baseline task, contributors are asked to read the text and identify any tokens that they believe are associated with identity. In the assistive-annotation tasks, contributors are also provided with annotated associated class (RNE, Religion, or SOGIESC) for each identity token. The task consists of three steps:

1. **Review the class labels.** Contributors are asked to review the class labels and definitions before beginning before beginning the actual annotation task. This helps ensure that contributors are familiar with the different types of classes and the criteria for annotation.

2. **Read the text/comment**: We ask contributors to read the text/comment in detail.

3. **Selecting/validating “token” and respective “class”**: In the baseline task, contributors are asked to select any tokens that they believe are associated with identity. In the assistive-annotation task, they are asked to validate the

assistive-annotations and select the ones that were missed.

For each class of identity-related tokens, contributors are provided with specific guidelines and examples. For example, for the RNE class, they are asked to select tokens that refer to race, nationality or ethnicity (e.g. black, white, spaniard, indian) or RNE insults (e.g. wetback, bluegum). They are specifically instructed not to annotate people names (e.g. John, Abdul) or terms that do not describe a specific group's race or ethnicity (e.g. literal terms like racist, race, ethnicity, ethnic group).

Similarly, for the Religion class, contributors are asked to select tokens that refer to religious groups (e.g. islam, muslim, christian, jewish) or religious insults (e.g. kike, raghead). They are specifically instructed not to annotate people names/religious figures (e.g. Jesus, Christ, Mohammad, Bishop) or religious worship terms (e.g. Church, Temple, Mosque).

Finally for the SOGIESC class, contributors are asked to select tokens that refer to particular SOGIESC (e.g. trans, bisexual, cisgender, queer, lgbtq), SOGIESC insults (e.g. fag, poof, bull dyke) or gendered terms (e.g. man, woman). They are specifically instructed not to annotate pronouns (e.g. he, she, him, her, they), a gendered name (e.g. Donald, Margaret) or literal terms (e.g., sex, gender, sexual, sexist).

In addition to selecting and validating tokens, contributors are also asked to provide a brief explanation of why they believe the token is associated with the selected class. To help reduce spam and gibberish in this free-form text field, we use an ML-assisted text utterance tool by Appen on low threshold settings (Appen, d).

Finally, to ensure the quality of the annotations a small subset of the task questions are used in a test run. Questions with high agreement in their answers are then used as new test questions. We use the Gold pool feature by Appen (Appen, b) to select these test questions.

B.4 Human computation results

IAR results.

Table 7 shows all measures we used for human annotation reliability evaluation.

False Positives/False Negatives.

Analysis in Table 8 shows a false positive, false negative trade-off between assistive annotations with token-matching and with disambiguation for all 3 groups, while token-matching with disambiguation additionally had No Classes false positives.

	Percent Agreement	Krippendorff's Alpha	Gwet's AC1
Example-only (baseline)	0.4036	0.404	0.4027
Assistive Identity Group annotations	0.7636	0.763	0.7622
Assistive Identity Group + Connotation annotations	0.6265	0.6316	0.6257

Table 7: IAR for human annotations: identity labeling on CivilComments (All metrics)

	RNE		Religion		SOGIESC		No Classes
	FP	FN	FP	FN	FP	FN	FP
Assistive token-matching annotation	2	1	1	4	5	4	0
Assistive token-matching + disambiguation	1	4	4	1	3	3	3

Table 8: Multi-class false positive (FP) and false negative (FN) counts for identity labeling human computation task on CivilComments

C4 performance as control

We use C4 as control to corroborate the impact of assistive annotation for identity labeling. We run three variations of human annotation tasks, similar to CivilComments. We use the output of the example-only (no assistive annotations) task as the ground truth. Results show similar performance on both datasets. The IAR improvement (Figure 13) of token-matching is more prominent in C4 than in CivilComments, when compared to the baseline. Similarly F1 scores (Figure 14) are consistently better for token-matching across all groups.

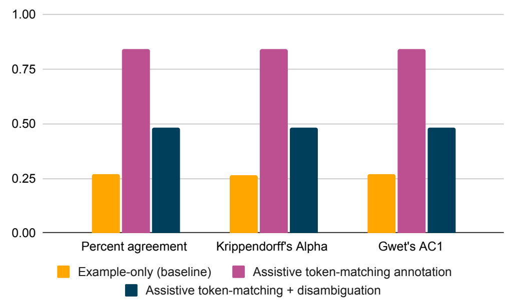

Figure 13: IAR for human annotations: identity labeling on C4.

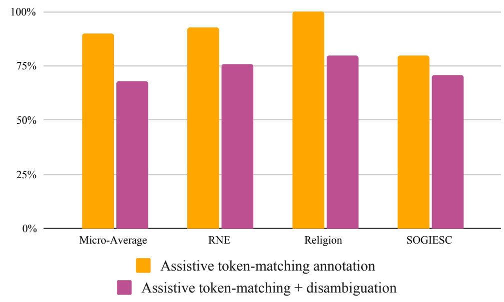

Figure 14: Multi-class F1 scores for human annotations: identity labeling on C4.

C Fairness Applications

C.1 Data and models

Dataset preprocessing.

We use the original splits of the CivilComments dataset (Jigsaw, 2019) for classifier training and evaluation. We only include toxicity labels where rater agreement is greater than 0.5. All input data is lower-cased for annotation.

Model training.

All models are trained for 11 epochs with a dropout rate of 30%, using an early stopping patience window of 3 epochs.

C.2 Counterfactuals

Similarity logic.

We use the nnlm-en-dim128[6] embedding to compute similarity between terms in the lexicon. To create a counterfactual mapping we first generate a subspace of the embedding which constitutes terms for an identity group that exist in its vocabulary. To find the least similar terms, we compute the linear distance from the reflection of the term around the center of the space. The center is the average value of all vectors in the embedding subspace.

Candidate generation.

To generate counterfactuals we first annotate terms with identity groups and subgroups. We then replace all terms in a text with their corresponding counterfactuals. To address cases where identity impacts toxicity, we only generate counterfactuals for labels which are not expected to be influenced by identity i.e. identity attack greater than or equal to 0.5.

C.3 Human computation design details

We adapt the annotation instructions from Perspective API[7] for our toxicity labeling task. Similar to the Perspective API process we discard NOT SURE human annotations and map TOXIC and VERY TOXIC to 1.0, and NOT TOXIC to 0.0.

We ask the human annotators to answer the toxicity question and identity identity-based attack question (enabled only if the answer of to toxicity question was VERY TOXIC, TOXIC or HARD TO SAY). We also ask human annotators to also provide a reason for their selection.

We do not highlight any tokens or provide context for the baseline task. For the assistive-annotation tasks, we highlight the tokens and provide the context associated with them. The test questions for these tasks are created using the strategy in Appendix B.3.

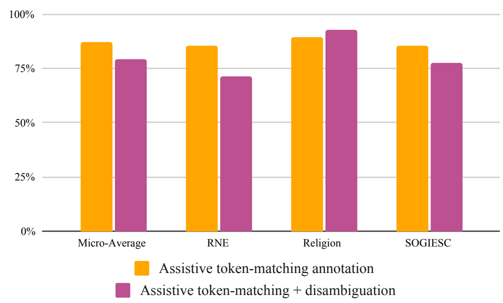

Figure 15: Multi-class F1 scores for human annotations: toxicity labeling on CivilComments.

	Velocity	Cost	Ease of Job	Pay
	Judgement Time (s)	Total Judgements	Scale: 1-5	Scale: 1-5
Example-only (baseline)	33	11987	2.4	3.3
Assistive Identity Group annotations	60	12014	3.6	3.4
Assistive Identity Group + Connotation annotations	46	12140	2.5	3.5

Table 9: Velocity, cost and satisfaction results from human annotation tasks for toxicity labeling

C.4 Human computation results

Pre-processing.

For F1 score computation, we discard NOT SURE from toxicity human annotations, and map TOXIC and VERY TOXIC to 1 and NOT TOXIC to 0 for a binary output. Similarly we discard NOT SURE from identity attack human annotations, and map

[6] https://tfhub.dev/google/nnlm-en-dim128/1

[7] https://github.com/conversationai/conversationai.github.io

YES to 1 and NO to 0. We then use F1 binary average scores to gauge the overall performance and the output of the example only (no assistive annotations) job as ground truth for this comparison.

Qualitative analysis.

The assistive IdentityGroup+Connotation task achieves the highest F1 score for both toxicity and identity-based labeling attack labels. The difference in performance is more pronounced in toxicity labeling (Figure 15).

The human annotation task with no assistive identity context performs the best in terms of velocity, taking 45% and 28.26% less time than the assistive IdentityGroup and IdentityGroup+Connotation tasks, respectively (Table 9). Cost-wise, the baseline task is slightly better than the assistive tasks, although they all perform similarly. In the optional satisfaction survey, human annotators find the IdentityGroup+Connotation task to be easier to perform (33.33%) and have slightly better pay compared to the baseline task.

The assistive IdentityFacet+Connotation annotation improves the IAR in human computation tasks for toxicity labeling compared to the baseline. However, the assistive IdentityFacet annotation leads to higher IAR for the "Identity based attack" question. This could indicate that showing Connotations might bias toxicity labels while showing IdentityGroups might bias identity-based attack labels.

Considering all the above, providing assistive identity context for task labeling should be approached carefully since it may lead to unintended bias in the labels required for model training and testing.