Characterization of Stigmatizing Language in Medical Records

Keith Harrigian[†], **Ayah Zirikly**[†], **Brant Chee**[⋆‡], **Alya Ahmad**[⋆],
Anne R. Links[⋆], **Somnath Saha**[⋆], **Mary Catherine Beach**[⋆], **Mark Dredze**[†]
[†]Department of Computer Science, [‡]Applied Physics Laboratory, [⋆]School of Medicine
Johns Hopkins University
Baltimore, MD
{kharrig5,azirikl1,aahmad24}@jhu.edu, Brant.Chee@jhuapl.edu
{alinks1,ssaha9,mcbeach}@jhmi.edu, mdredze@cs.jhu.edu

Abstract

Widespread disparities in clinical outcomes exist between different demographic groups in the United States. A new line of work in medical sociology has demonstrated physicians often use stigmatizing language in electronic medical records within certain groups, such as black patients, which may exacerbate disparities. In this study, we characterize these instances at scale using a series of domain-informed NLP techniques. We highlight important differences between this task and analogous bias-related tasks studied within the NLP community (e.g., classifying microaggressions). Our study establishes a foundation for NLP researchers to contribute timely insights to a problem domain brought to the forefront by recent legislation regarding clinical documentation transparency. We release data, code, and models.[1]

1 Introduction

Widespread and well-documented disparities in healthcare outcomes between demographic groups exist within the United States (Baciu et al., 2017; Zavala et al., 2021). The sources of these disparities are diverse and complex, with numerous interacting factors contributing to worse outcomes for minority patients (Bell and Lee, 2011; Williams et al., 2019). One source of disparities may stem from latent biases of healthcare providers (Hall et al., 2015). Multiple studies have highlighted the tendency for providers to prescribe different treatment plans to black patients compared to white patients despite having similar clinical dispositions (Nelson, 2002; Green et al., 2007; Hoffman et al., 2016). Elevated implicit bias scores have been associated with these decisions and have been further linked with decreased levels of patient-provider communication (Van Ryn et al., 2011; Cooper et al., 2012). A major challenge with these biases is that they are invoked unconsciously.

A new line of work in medical sociology has explored this issue through the lens of clinical documentation (Beach et al., 2021), in which bias may be exhibited in how medical providers describe and document patient interactions in the medical record. In particular, studies have shown physicians often use language that has subtle, stigmatizing connotations (Wolsiefer et al., 2021). This documentation practice may not only negatively frame patients to future providers and thus influence their quality of care, but also discourage patients from seeking treatment altogether (Goddu et al., 2018; Werder et al., 2022). The latter is especially pertinent given the passage of the 21st Century Cures Act that mandates clinical notes are freely accessible by patients in the US (Blease et al., 2021; Harris et al., 2022).

How is stigmatizing language in medical records different from other forms of abusive language? Prior studies of stigmatizing language in clinical notes have relied on qualitative methods (Park et al., 2021) or refrained from analyzing computational nuances of the problem domain (Sun et al., 2022). Modeling tasks such as hate-speech detection (Jahan and Oussalah, 2021; Garg et al., 2022) and analyses of social bias encoded within language models (Liang et al., 2021) share many similarities with characterizing stigmatizing language in medical records. However, it is not clear *a priori* where the task of characterizing stigmatizing language in medical records falls within the broader abusive language landscape.

In this paper, we demonstrate that characterization of stigmatizing language in medical records most strongly parallels the characterization of linguistic microaggressions (Sue et al., 2007). However, unlike traditional microaggressions, biased language in the clinical domain is concentrated in unremarkable phrases and lacks any indication of the targeted identity group. Our analysis establishes a foundation for a novel task that has high importance to both patients and clinicians.

[1] github.com/kharrigian/ehr-stigma

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 312–329
July 9-14, 2023 ©2023 Association for Computational Linguistics

2 Stigmatizing Language in Medical Records as Abusive Language

Clinical stigmatizing language lies in the *implicit* and *directed* quadrant of the typology of abusive language introduced by Waseem et al. (2017). Physicians generally use a vocabulary of commonplace terms and phrases which have negative implications only when interpreted in certain contexts or by other physicians (Valdez, 2021; Beach et al., 2021). This language almost always places the patient as the target of the stigma, even if they are not the intended recipient (Ho et al., 2014).

Stigmatizing language in medical records shares many similarities with linguistic microaggressions. Both reflect an unconscious bias internalized by the speaker and materialized through thinly veiled innuendo (Sue et al., 2007; Raney et al., 2021). This innuendo is not necessarily negative in affect (Glick and Fiske, 2001; McMahon and Kahn, 2016).

One major difference between stigmatizing language in the clinical domain and other forms of abusive language is the notion of *necessity*. Whereas most abusive language is better left unsaid, clinicians have a responsibility to document their interaction with patients (Shanley et al., 2009). Often, this requires that they characterize socially-stigmatized circumstances (e.g., substance use disorders) and medically-relevant patient eccentricities (e.g., unfounded social histories). Minor differences in phrasing may have a large impact on whether a statement is stigmatizing to patients.

The idea of stigmatizing language in medical records is relatively new, with Goddu et al. (2018) providing the first qualitative evidence of negative language in the medical record. Using word counts, Beach et al. (2021) and Himmelstein et al. (2022) later identified a higher prevalence of implicit bias within records of black patients than white patients.

Sun et al. (2022) was the first to use machine learning to analyze stigmatizing language in medical records. The authors identified sentences with possible bias using a manually-curated word list and then annotated whether each match was positive, negative, or out-of-context. A logistic regression classifier trained on a bag-of-words representation of the text achieved good performance (F1 of 0.935). Unfortunately, the authors did not provide a baseline to indicate how valuable context around the seed terms is for classification.

The more general task of identifying biased and abusive language in text has garnered much attention from researchers in recent years (Schmidt and Wiegand, 2017; Yin and Zubiaga, 2021). Breitfeller et al. (2019) was the first to computationally analyze microaggressions. The majority of microaggression research published thereafter has remained confined to using web data (Lees et al., 2021; Sabri et al., 2021). Our study provides an analysis of stigma in an important linguistic domain that differs dramatically from those currently studied in the covert bias research space.

3 Data

We consider two clinical datasets. In addition to covering different clinical specialties, they also feature different demographic compositions.

JHM We retrospectively acquired a dataset of 128,343 English-language progress notes written by physicians across 5 clinical specialties within the Johns Hopkins Medicine (JHM) hospital system — Internal Medicine, Emergency Medicine, Pediatrics, OB-GYN, and Surgery. Notes were processed in accordance with our institution's privacy policy after approval by our Institutional Review Board (IRB). Because the notes contain sensitive identifiable information, they are unable to be shared beyond our study team.

MIMIC To encourage future research, we also include in our study the publicly-accessible MIMIC-IV-Note dataset (v2.2) (Johnson et al., 2023). This recently released extension of the widely-adopted MIMIC-III dataset (Johnson et al., 2016) consists of deidentified free-text clinical notes for patients admitted to an intensive care unit (ICU) or the emergency department at Beth Israel Deaconess Medical Center in Boston, MA. We focus on the 331,794 available discharge summaries, having found minimal evidence of stigmatizing language in the associated radiology reports.

3.1 Annotation

Like Sun et al. (2022), we develop a two-stage process to detect and characterize stigmatizing language in clinical notes. Possible instances of bias are first identified using *anchor* n-grams and then classified using a machine learning classifier. We take the union of n-grams curated by Beach et al. (2021) and Sun et al. (2022) as our anchor set.

Unlike the single, sentiment-like classification task considered by Sun et al. (2022), we formulate three independent classification tasks that discriminate between instances of bias based on impact.

1. **Credibility & Obstinacy** (Disbelief, Difficult, Exclude): insinuation of doubt regarding a patient's testimony or describes the patient as obstinate.
2. **Compliance** (Negative, Neutral, Positive): patient does not appear to follow medical advice.
3. **Descriptors** (Negative, Neutral, Positive, Exclude): evaluates descriptions of patient behavior and demeanor.

We ran our anchor list against both datasets, caching each match and up to 10 words to the left and right which make up its context. A team of annotators (research assistant and physician coauthors) labeled a random sample of 5,201 and 5,043 instances from the JHM and MIMIC datasets, respectively. All instances in the JHM dataset and the majority of the instances in the MIMIC dataset were labeled independently by at least two annotators.[2] We include annotator agreement measures, the label distribution, and full task taxonomy with examples in Appendix A.

4 Characterizing Stigmatizing Language

4.1 What role does context play in characterizing stigmatizing language?

Some forms of abusive language are stigmatizing in isolation, while others critically depend on context to invoke meaning (Waseem et al., 2017). Prior work has not provided insight regarding where stigmatizing language in medical records lies on this spectrum (Sun et al., 2022). We hypothesize that context around a stigmatizing instance is necessary, but insufficient, for characterizing the utterance.

Methods We test our hypothesis by varying feature representations such that they encode different degrees of the stigmatizing anchor term and its surrounding context. We consider 3 classes of models. The first two classes allow us to understand the interaction between context and the anchor n-grams in an additive manner. The third class captures more complex dynamics between anchor n-grams and their context. Additional training and evaluation details are included in Appendix B.

1. **Majority**: Majority class and majority class conditioned on anchor n-gram.
2. **Logistic Regression (LR)**: TF-IDF representations. One version with the anchor n-gram and one without.
3. **BERT**: One version trained on web data (Devlin et al., 2018) and one version trained on clinical notes (Alsentzer et al., 2019).

We also compare four methods of pooling BERT's final hidden layer for input into the task classification head.

1. **Anchor Mean**: Arithmetic mean of tokens (subwords) composing the anchor n-gram.
2. **CLS**: Embedding for the classification token.
3. **Sentence Mean**: Arithmetic mean of all tokens in the instance, excluding special tokens.
4. **BERT Pooler**: Weighted pooling of all tokens; weights learned at training time.

Results The final four rows in Table 1 show clinical BERT's test-set macro F1-score for each pooling method across the three classification tasks; the web version of BERT performs similarly. Although not always statistically significant, the anchored pooling method consistently outperforms the alternative pooling approaches across all tasks and datasets. Under this setting, the classification head lacks direct access to information in each anchor's context window. Classification performance can be thought of as a measure of how well the closed set of anchor n-grams are separated in semantic space. That the anchor pooling approach outperforms the alternative methods suggests characterizing stigmatizing language in medical records can be thought of as a word-sense-disambiguation task more than a sequence classification task.

The majority and logistic regression model outcomes (first four rows of Table 1) lend additional support to this claim. We see that anchors used as classification criteria in isolation provide a significant improvement over the majority overall model in all cases. The context window used in isolation provides a relatively smaller increase in performance over the majority overall model. Jointly modeling the anchors and their context achieves the largest improvement over the majority overall model in 4 of 6 tasks. This outcome suggests that both subsets of text provide different, but complementary, information.

The BERT models effectively capture the interaction between anchors and their surrounding context. Fine-tuning both BERT models significantly increases macro F1 over the best non-BERT model in all settings. Interestingly, the difference in performance between the web and clinical BERT models

[2] A small number of instances from MIMIC were labeled by a single annotator after observing high agreement scores.

Model	Credibility & Obstinacy		Compliance		Descriptors	
	JHM	MIMIC	JHM	MIMIC	JHM	MIMIC
Majority Overall	0.21 ± 0.00	0.17 ± 0.00	0.29 ± 0.00	0.24 ± 0.00	0.16 ± 0.00	0.19 ± 0.00
Majority Per Anchor	0.67 ± 0.10	0.55 ± 0.04	0.68 ± 0.04	0.73 ± 0.01	0.82 ± 0.01	0.83 ± 0.00
LR (Context)	0.60 ± 0.05	0.58 ± 0.04	0.55 ± 0.01	0.68 ± 0.02	0.74 ± 0.03	0.60 ± 0.04
LR (Context + Anchor)	0.69 ± 0.02	0.65 ± 0.03	0.68 ± 0.04	0.80 ± 0.02	0.86 ± 0.02	0.76 ± 0.05
Bert (Web)	0.85 ± 0.04	0.76 ± 0.02	$\mathbf{0.86 \pm 0.01}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$	$\mathbf{0.86 \pm 0.01}$
Bert (Clinical)	$\mathbf{0.89 \pm 0.03}$	$\mathbf{0.78 \pm 0.03}$	0.85 ± 0.02	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.86 \pm 0.01}$
– CLS Token	0.89 ± 0.04	$\underline{0.69 \pm 0.03}$	0.84 ± 0.03	0.92 ± 0.01	$\underline{0.90 \pm 0.01}$	0.84 ± 0.03
– Sentence Mean	0.85 ± 0.06	$\underline{0.69 \pm 0.06}$	0.84 ± 0.03	0.92 ± 0.01	$\underline{0.91 \pm 0.01}$	$\underline{0.84 \pm 0.02}$
– BERT Pooler	0.83 ± 0.08	0.70 ± 0.07	0.84 ± 0.02	0.91 ± 0.02	$\underline{0.89 \pm 0.03}$	$\underline{0.80 \pm 0.03}$

Table 1: Test macro F1 ($\mu \pm \sigma$) for each classification task. Underlining indicates a pooling method is significantly worse than anchor mean pooling (paired t-test $p < 0.05$). The best model(s) for each classification task are bolded.

is not significant. We hypothesize that understanding social bias may be more important than understanding clinical jargon for our tasks, but leave this as an open question for future work.

4.2 Is stigma conveyed in the same manner about different demographic groups?

The majority of bias-related tasks in NLP examine language which, while covert, contains some indication of the targeted demographic of identity group (e.g., racial slurs, sexist microaggressions) (Sue, 2010; Waseem et al., 2017). Here, we show that stigmatizing language in medical records uniquely *does not* target any racial group or sex.

Methods Results from §4.1 verify that our BERT encoders learn semantic representations of the anchor n-grams which are informative for the downstream stigma characterization tasks. If language is used differently for different demographic groups, we expect the encoders to reflect this (Adam et al., 2022). We can test our hypothesis by attempting to infer a patient's self-reported race and sex using each anchor n-gram's BERT representation.

Because our datasets represent a concatenation of notes from multiple clinical specialties which each have a unique demographic pool, it's possible to conflate the encoding of specialty knowledge with demographic knowledge. Additionally, any differences in the prevalence of our anchor n-grams or their associated labels between demographic groups could be exploited by a classifier. For this reason, we ground inference performance against baselines which model one-hot-encoded representations of the anchor n-gram, clinical speciality, and the primary classification task label. We also consider a version of the anchor embeddings generated after replacing gender-indicative pronouns (e.g., himself, her) and other identifiers with non-uniform gender associations (e.g., woman, husband) with gender-neutral alternatives. As before, additional experimental details are included in the appendix.

Results We present demographic inference results for the JHM dataset in Table 2 and report MIMIC results in the appendix. Across all but one experimental setting, inference performance achieved using the gender-neutral version of the embeddings is not significantly different from what is achieved by the metadata-only baselines. This trend suggests that the learned embeddings encode little to no information about a patient's race or sex that cannot be explained by underlying differences in prevalence between patient populations. Future work is necessary to understand whether there exist semantic differences along other axes (e.g., socioeconomic status, substance use, obesity) (Healy et al., 2022).

4.3 Is stigma conveyed in the same manner across different patient populations?

Machine learning models trained on one distribution often experience a loss in performance when evaluated on a different distribution (Blitzer et al., 2006; Harrigian et al., 2020). Understanding the causes of this loss is necessary for ensuring systems do not exacerbate existing social disparities (Bender et al., 2021). Here, we identify speciality-specific nuances in stigmatizing language and highlight limitations of anchor-focused modeling.

Methods We evaluate models trained using the JHM dataset in §4.1 on the test set of the MIMIC dataset, and vice-versa. We also conduct a qualitative error analysis to understand how stigmatizing language differs between the two datasets.

Model	Credibility & Obstinacy		Compliance		Descriptors	
	Sex	Race	Sex	Race	Sex	Race
Majority Baseline	0.37 ± 0.01	0.26 ± 0.02	0.37 ± 0.02	0.29 ± 0.01	0.35 ± 0.02	0.26 ± 0.01
Anchor	0.50 ± 0.04	0.31 ± 0.05	0.42 ± 0.02	0.29 ± 0.01	0.50 ± 0.02	0.30 ± 0.03
Label	0.37 ± 0.01	0.27 ± 0.03	0.37 ± 0.02	0.29 ± 0.01	0.46 ± 0.07	0.26 ± 0.01
Specialty	0.44 ± 0.04	0.36 ± 0.04	0.53 ± 0.04	0.29 ± 0.01	0.58 ± 0.03	0.32 ± 0.03
Anchor × Label	0.50 ± 0.03	0.31 ± 0.05	0.46 ± 0.03	0.30 ± 0.01	0.53 ± 0.02	0.32 ± 0.04
Anchor × Speciality	0.51 ± 0.04	0.38 ± 0.03	0.54 ± 0.02	0.35 ± 0.03	0.56 ± 0.04	0.34 ± 0.02
Label × Speciality	0.47 ± 0.04	0.38 ± 0.04	0.53 ± 0.05	0.32 ± 0.02	0.58 ± 0.04	0.32 ± 0.03
Anchor × Label × Speciality	0.54 ± 0.01	0.35 ± 0.03	0.54 ± 0.03	0.36 ± 0.02	0.55 ± 0.03	0.36 ± 0.02
Embedding	0.76 ± 0.02	0.34 ± 0.02	0.57 ± 0.01	0.36 ± 0.02	0.61 ± 0.04	0.34 ± 0.03
Embedding (Gender Neutral)	$\mathbf{0.59 \pm 0.02}$	0.34 ± 0.06	0.52 ± 0.01	0.35 ± 0.01	0.52 ± 0.03	0.34 ± 0.02

Table 2: Average held-out macro F1-score ($\mu \pm \sigma$) for the demographic inference tasks in the JHM dataset. Inference performance using the gender-neutral embeddings is only significantly different from the baselines in one setting.

		Credibility & Obstinacy		Compliance		Descriptors	
	Target →	JHM	MIMIC	JHM	MIMIC	JHM	MIMIC
Source ↕	JHM	$\mathbf{0.89 \pm 0.03}$	0.70 ± 0.01	$\mathbf{0.85 \pm 0.02}$	0.86 ± 0.03	$\mathbf{0.93 \pm 0.02}$	0.81 ± 0.03
	MIMIC	0.81 ± 0.03	$\mathbf{0.78 \pm 0.03}$	0.82 ± 0.02	$\mathbf{0.92 \pm 0.02}$	0.89 ± 0.03	$\mathbf{0.86 \pm 0.01}$

Table 3: Average test macro F1-score ($\mu \pm \sigma$) when transferring between datasets. There exists a statistically significant loss in performance (paired t-test $p < 0.05$) within all transfer settings (columns).

Results We observe consistent drops in performance when models are evaluated in a different domain than which they were trained (i.e., Table 3). This performance loss is significant in all 6 transfer settings. What causes this loss? Are there spurious artifacts to which our models overfit (Wang et al., 2022)? Or does each dataset contain unique stigmatizing language that arises disproportionately across patient populations?

Although many transfer errors can be attributed to differences in each dataset's joint anchor-label distribution, some special cases emerge. For example, models trained on the JHM dataset incorrectly characterize instances in MIMIC which describe parties secondary to the patient (e.g., family). This situation is more common in the MIMIC dataset due to ICU patients often being incapacitated. Models trained on the JHM dataset also struggle with statements in MIMIC from Psych ICU notes, where patients frequently describe their own behavior.

One on hand, these shortcomings appear to be a consequence of covariate shift (Sugiyama et al., 2007), for which many general mitigation strategies exist (Ramponi and Plank, 2020). On the other hand, each of the errors we observe presents a unique linguistic challenge that may be better handled using targeted interventions. Few-shot word sense disambiguation techniques may improve transfer for low-volume anchor-label pairs (Kumar et al., 2019; Scarlini et al., 2020), while augmented annotations may reduce speaker/receiver confusion (Rashkin et al., 2016; Hovy and Yang, 2021).

5 Discussion

The covert, highly contextual, and non-demographically aligned nature of stigmatizing language in medical records places it in a unique area of the abusive language research landscape. The current reliance on domain experts to identify possible instances of bias using anchor terms is limiting given the adversarial relationship been abusive language and speakers (Nobata et al., 2016). It also does not address abstract forms of stigma (Kopera et al., 2015) or stigmatizing pragmatics (Beach and Saha, 2021).

Methods for discovering stigmatizing language in medical records are poised to be highly impactful (Field and Tsvetkov, 2020). Counterfactual analyses may be instrumental for better characterizing the nuance between stigmatizing and non-stigmatizing clinical language (Kaushik et al., 2019). Whether these nuances are uniform across patient populations (e.g., hospital systems, regions) and providers (e.g., nurses, resident physicians) remains an open question not answerable from our datasets alone. Likewise, future work is necessary to understand whether clinical knowledge is necessary for models in this domain (Roberts, 2016).

Ethics Statement

Our datasets were collected from real patients, contain protected health information (PHI), and are subject to HIPAA regulations. As a result, we took the utmost care to maintain data integrity and privacy. First, we obtained IRB approval to access and process the data. Second, we obtained permission and approval for all applications and libraries used to process the data. Third, data storage and computational experimentation was done on IRB-approved platforms.

Limitations

In our work we faced numerous types of limitations that fall under different categories.

Data Our relatively small dataset size limits our analysis, especially with the use of language models. Furthermore, the label distribution is skewed across the different specialties (domains), which affects model performance, robustness and generalizability. The differences in distribution might be the result of how the data was collected, which was not in light of the anchor words, or due to the domain's nature and/or the medical providers' language of that specialty. Furthermore, the time frame that the data was sampled from might manifest certain biases that are different from other time frames. Finally, our datasets are only representative of a small number of specialties from two medical institutions. Patient populations and providers may vary greatly across medical fields and additional institutions.

Task The formulation of the labels for our task imposes limitations and challenges. Stigmatizing language is subjective and can vary between the perspective of the patient and the medical provider. As a result, we are aware that our medical experts' annotations might impose a bias. Additionally, the negative connotations of language might be ambiguous and can change depending on a medical expert's identity, background and specialty, which creates a bias that is hard to mitigate.

Computational Resources We only used IRB-approved servers to access the dataset and perform the experiments. Because these platforms had limited computational capacity and lacked the specifications required to build more complex neural models, we were not able to include more recent language models in our experiments that might have yielded better performance. In the future, we hope to have access to machines that support more recent and state-of-the-art models.

Acknowledgements

This work was supported by the National Institute on Minority Health and Health Disparities under grant number R01 MD017048. The content is solely the responsibility of the authors and does not necessarily represent the official views of NIMHD, NIH, or Johns Hopkins University.

References

Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini Soares, Charles Senteio, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write it like you see it: Detectable differences in clinical notes by race lead to differential model recommendations. In *AAAI/ACM Conference on AI, Ethics, and Society*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Alina Baciu, Yamrot Negussie, Amy Geller, James N Weinstein, National Academies of Sciences, Engineering, and Medicine, et al. 2017. The state of health disparities in the united states. In *Communities in action: Pathways to health equity*. National Academies Press (US).

Mary Catherine Beach and Somnath Saha. 2021. Quoting patients in clinical notes: First, do no harm. *Annals of internal medicine*, 174(10):1454–1455.

Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A Cooper, and Brant Chee. 2021. Testimonial injustice: linguistic bias in the medical records of black patients and women. *Journal of general internal medicine*, 36(6):1708–1714.

Judith Bell and Mary M Lee. 2011. Why place and race matter: Impacting health through a focus on race and place. *Oakland, CA: PolicyLink*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Charlotte Blease, Jan Walker, Catherine M DesRoches, and Tom Delbanco. 2021. New us law mandates access to clinical notes: implications for patients and clinicians.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1664–1674.

Lisa A Cooper, Debra L Roter, Kathryn A Carson, Mary Catherine Beach, Janice A Sabin, Anthony G Greenwald, and Thomas S Inui. 2012. The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care. *American journal of public health*, 102(5):979–987.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438. American Medical Informatics Association.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2022. Handling bias in toxic speech detection: A survey. *arXiv preprint arXiv:2202.00126*.

Peter Glick and Susan T Fiske. 2001. An ambivalent alliance: Hostile and benevolent sexism as complementary justifications for gender inequality. *American psychologist*, 56(2):109.

Anna P Goddu, Katie J O'Conor, Sophie Lanzkron, Mustapha O Saheed, Somnath Saha, Monica E Peek, Carlton Haywood, and Mary Catherine Beach. 2018. Do words matter? stigmatizing language and the transmission of bias in the medical record. *Journal of general internal medicine*, 33(5):685–691.

Alexander R Green, Dana R Carney, Daniel J Pallin, Long H Ngo, Kristal L Raymond, Lisa I Iezzoni, and Mahzarin R Banaji. 2007. Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients. *Journal of general internal medicine*, 22(9):1231–1238.

William J Hall, Mimi V Chapman, Kent M Lee, Yesenia M Merino, Tainayah W Thomas, B Keith Payne, Eugenia Eng, Steven H Day, and Tamera Coyne-Beasley. 2015. Implicit racial/ethnic bias among health care professionals and its influence on health care outcomes: a systematic review. *American journal of public health*, 105(12):e60–e76.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. 2020. Do models of mental health based on social media data generalize? In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3774–3788.

Jennifer Huang Harris, Nomi C Levy-Carrick, and Ashwini Nadkarni. 2022. Opennotes: transparency versus stigma in patient care. *The Lancet Psychiatry*, 9(6):426–428.

Megan Healy, Alison Richard, and Khameer Kidia. 2022. How to reduce stigma and bias in clinical communication: a narrative review. *Journal of General Internal Medicine*, pages 1–8.

Gracie Himmelstein, David Bates, and Li Zhou. 2022. Examination of stigmatizing language in the electronic health record. *JAMA network open*, 5(1):e2144967–e2144967.

Y-X Ho, CS Gadd, KL Kohorst, and ST Rosenbloom. 2014. A qualitative analysis evaluating the purposes and practices of clinical documentation. *Applied Clinical Informatics*, 5(01):153–168.

Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, 113(16):4296–4301.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

Md Saroar Jahan and Mourad Oussalah. 2021. A systematic review of hate speech automatic detection using natural language processing. *arXiv preprint arXiv:2106.00742*.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. Mimic-iv-note: Deidentified free-text clinical notes.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Maciej Kopera, Hubert Suszek, Erin Bonar, Maciej Myszka, Bartłomiej Gmaj, Mark Ilgen, and Marcin Wojnar. 2015. Evaluating explicit and implicit stigma of mental illness in mental health professionals and medical students. *Community mental health journal*, 51(5):628–634.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

Alyssa Lees, Daniel Borkan, Ian Kivlichan, Jorge Nario, and Tesh Goyal. 2021. Capturing covertly toxic speech via crowdsourcing. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 14–20.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jean M McMahon and Kimberly Barsamian Kahn. 2016. Benevolent racism? the impact of target race on ambivalent sexism. *Group Processes & Intergroup Relations*, 19(2):169–183.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Alan Nelson. 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the national medical association*, 94(8):666.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach. 2021. Physician use of stigmatizing language in patient medical records. *JAMA Network Open*, 4(7):e2117052–e2117052.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855.

Julia Raney, Ria Pal, Tiffany Lee, Samuel Ricardo Saenz, Devika Bhushan, Peter Leahy, Carrie Johnson, Cynthia Kapphahn, Michael A Gisondi, and Kim Hoang. 2021. Words matter: an antibias workshop for health care professionals to reduce stigmatizing language. *MedEdPORTAL*, 17:11115.

Hannah Rashkin, Sameer Singh, and Yejin Choi. 2016. Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 311–321.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Kirk Roberts. 2016. Assessing the corpus size vs. similarity trade-off for word embeddings in clinical nlp. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 54–63.

Nazanin Sabri, Valerio Basile, Tommaso Caselli, et al. 2021. Leveraging bias in pre-trained word embeddings for unsupervised microaggression detection. In *CLiC-it*.

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. 2020. Sensembert: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8758–8765.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Jenelle R Shanley, Deborah Shropshire, and Barbara L Bonner. 2009. To report or not report: A physician's dilemma. *AMA Journal of Ethics*, 11(2):141–145.

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Derald Wing Sue, Christina M Capodilupo, Gina C Torino, Jennifer M Bucceri, Aisha Holder, Kevin L Nadal, and Marta Esquilin. 2007. Racial microaggressions in everyday life: implications for clinical practice. *American psychologist*, 62(4):271.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5).

Michael Sun, Tomasz Oliwa, Monica E Peek, and Elizabeth L Tung. 2022. Negative patient descriptors: Documenting racial bias in the electronic health record: Study examines racial bias in the patient descriptors used in the electronic health record. *Health Affairs*, 41(2):203–211.

Anna Valdez. 2021. Words matter: Labelling, bias and stigma in nursing. *Journal of Advanced Nursing*, 77(11):e33–e35.

Michelle Van Ryn, Diana J Burgess, John F Dovidio, Sean M Phelan, Somnath Saha, Jennifer Malat, Joan M Griffin, Steven S Fu, and Sylvia Perry. 2011. The impact of racism on clinician cognition, behavior, and clinical decision making. *Du Bois review: social science research on race*, 8(1):199–218.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in nlp models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1719–1729.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. *arXiv preprint arXiv:1705.09899*.

Karen Werder, Alexa Curtis, Stephanie Reynolds, and Jason Satterfield. 2022. Addressing bias and stigma in the language we use with persons with opioid use disorder: A narrative review. *Journal of the American Psychiatric Nurses Association*, 28(1):9–22.

David R Williams, Jourdyn A Lawrence, and Brigette A Davis. 2019. Racism and health: evidence and needed research. *Annual review of public health*, 40:105–125.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Katherine J Wolsiefer, Matthias Mehl, Gordon B Moskowitz, Colleen K Cagno, Colin A Zestcott, Alma Tejeda-Padron, and Jeff Stone. 2021. Investigating the relationship between resident physician implicit bias and language use during a clinical encounter with hispanic patients. *Health Communication*, pages 1–9.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

Valentina A Zavala, Paige M Bracci, John M Carethers, Luis Carvajal-Carmona, Nicole B Coggins, Marcia R Cruz-Correa, Melissa Davis, Adam J de Smith, Julie Dutil, Jane C Figueiredo, et al. 2021. Cancer health disparities in racial/ethnic minorities in the united states. *British journal of cancer*, 124(2):315–332.

Ciyou Zhu, Richard H Byrd, Peihuang Lu, and Jorge Nocedal. 1997. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on mathematical software (TOMS)*, 23(4):550–560.

A Data

A.1 Task Taxonomy

We present the task taxonomy developed for this study in Table 4, along with de-identified examples for each of the stigmatizing language classes. The taxonomy was developed by clinicians on our team, drawing upon previous literature (Beach et al., 2021; Sun et al., 2022). We plan to expand our current anchor n-gram list in future work using context-aware keyword discovery.

A.2 Anchor & Label Distribution

We provide the distribution of labels for each task in Table 5. This distribution is further broken down by anchor n-gram in Figure 1. Each task contains a subset of anchors with extreme class imbalance.

Task	Class	JHM	MIMIC
Credibility & Obstinacy	Difficult	413	526
	Disbelief	438	609
	Exclude	77	115
Compliance	Negative	1,578	893
	Neutral	283	439
	Positive	357	271
Descriptors	Exclude	430	496
	Negative	843	1,221
	Neutral	233	96
	Positive	549	377

Table 5: Label distribution for each task.

A.3 Annotator Agreement

Three annotators were responsible for labeling all data used in our study – one clinician C1 and two research assistants R1, R2. We present agreement matrices in Figure 2 for the MIMIC and JHM datasets. Each instance in the JHM dataset was labeled by at least two annotators, with a subset labeled by three. A subset of instances in the MIMIC dataset were labeled by two annotators, with the remainder labeled by a single annotator. Annotators labeled

Stigma Type	Class	Definition	Examples
Credibility & Obstinacy	Disbelief	Insinuates doubt about a patient's stated testimony.	*adamant* he doesn't smoke; *claims* to see a therapist
	Difficult	Describes patient (or patient's family) perspective as inflexible/difficult/entrenched, typically with respect to their intentions.	*insists* on being admitted; *adamantly* opposed to limiting fruit intake
	Exclude	Word/phrase is not used to characterize the patient or describe the patient's behavior; may refer to medical condition or treatment or to another person or context.	patient's friend *insisted* she go to the hospital; test *claims* submitted to insurance
Compliance	Negative	Patient not, unlikely to, or questionably following medical advice.	*adherence* to therapeutic medication is unclear; mother *declines* vaccines; struggles with medication and follow-up *compliance*
	Neutral	Not used to describe whether the patient is not following medical advice or rejecting treatment; often used to describe generically some future plan involving a hypothetical. Alternatively, see Exclude (Credibility & Obstinacy).	discussed the medication *compliance*; school *refuses* to provide adequate accommodations; feels that her parents' health has *declined*
	Positive	Patient following medical advice.	continues to be *compliant* with aspirin regimen; reports excellent *adherence*
Descriptors	Negative	Patient's demeanor or behavior is cast in a negative light; insinuates the patients is not being forthright or transparent; patient may be falsifying symptoms to get something they want.	*drug-seeking* behavior; concern for *secondary gain*; *unwilling* to meet with case manager; unfortunately a poor *historian*
	Neutral	Negation of negative descriptors; insinuates the patient was expected to have a negative demeanor or be difficult to interact with.	his mother is the primary *historian*; interactive and *cooperative*; not *combative* or *belligerent*; dad seems *angry* with patient at times
	Positive	Patient's demeanor or behavior is described in a positive light; patient is easy to interact with.	*lovely* 80 year old woman; *well-groomed* and holds good eye contact; *pleasant* and appropriate interaction with staff
	Exclude	Patient self-description or description of another individual. Alternatively, see Exclude (Credibility & Obstinacy).	does not want providers to think she's *malingering*; reports feeling *angry* before her period; lives on *pleasant* avenue downtown

Table 4: Taxonomy of stigmatizing language. Complete anchor sets for each task can be found in Figure 1. Annotators were provided a comprehensive guide with general examples and edge cases for each anchor n-gram in our taxonomy.

the data independently and then met with the larger team to resolve disagreements and discuss ambiguous cases.

Agreement scores prior to resolution were quite high, suggesting 1) the annotation taxonomy was clear and 2) the stigmatizing language we considered was generally not ambiguous in its impact. We observed similar agreement trends for both datasets; the Descriptors task had the highest agreement, while the Credibility & Obstinacy task had the lowest agreement. The former consists of several highly polar anchor n-grams (e.g., pleasantly, unkempt), while the latter requires a higher degree of personal interpretation.

A.4 Preprocessing

All clinical free text in our datasets was case-normalized and converted to an ASCII encoding prior to additional processing. The MIMIC dataset was de-identified before we obtained access to it. The JHM dataset, however, was not subject to any de-identification procedures because it is protected within a secure cloud environment and we are not distributing assets derived from it.

Anchor terms are identified using regular expressions implemented in Python's `re` package. Up to 10 words to the left and 10 words to the right of the matched spans (based on whitespace) are maintained for annotation and modeling. Context sizes were specified *a priori* based on guidance from our clinical collaborators; future work may consider evaluating the effect this choice has on annotation and modeling outcomes.

For the logistic regression models, we use a custom pipeline to transform the raw text into feature space. The text instances are first tokenized using a clinical domain tokenizer implemented in the medspaCy library (Eyre et al., 2021). Tokens are recursively merged together to form phrases based on the bi-gram scoring function introduced by Mikolov et al. (2013) and implemented in Gensim (Řehůřek and Sojka, 2010). We use a scoring threshold of 10, minimum vocabulary frequency of 5, and recurse twice to identify 1-4 grams.

B The Role of Context (§4.1)

B.1 Experimental Design

The annotated dataset is split into training, development, and test subsets at a 70/20/10 ratio. Instances are assigned randomly into each subset, using their associated patient identifiers as stratification criteria to limit data leakage. The training and development subsets are further split at random to facilitate 5-fold cross-validation.

B.2 Models

The Majority Per Anchor baseline outputs the following class probabilities given an input anchor n-gram w:

$$p(y \mid w) = \frac{C(w, y) + \alpha}{\sum_{y' \in \mathcal{Y}} C(w, y') + |\mathcal{Y}|\alpha}$$

where $C(w, y)$ is the number of examples with anchor w having class y in the training data, $\mathcal{Y}$ is the set of possible classes y, and α is a smoothing hyperparameter. We use $\alpha = 1$ for all of our experiments.

The logistic regression baselines use scikit-learn (Pedregosa et al., 2011) for data transformations and classifier training. For the TF-IDF representations, we use an ℓ_2 row-wise norm. As a classifier, we use multinomial logistic regression optimized using lbfgs (Zhu et al., 1997). We balance class weights and perform a grid search over the following ℓ_2 regularization parameters: 0.01, 0.03, 0.1, 0.3, 1, 3, 5, 10. The model which maximizes macro F1-score in each training split's associated development set is chosen for application on the test set.

We use Hugging Face's transformers library (Wolf et al., 2019) to initialize all BERT models and fine-tune them using code written in PyTorch (Paszke et al., 2019). We train all models using a batch size of 16, a fixed learning rate of 5e-05, a dropout probability of 0.1, and class-balanced cross-entropy loss. As an optimizer, we use AdamW (Loshchilov and Hutter, 2017). We evaluate the model every 50 updates and save the model which maximizes macro F1-score on the training split's associated development data. Due to compute limitations in our HIPAA-compliant environment (i.e., limited GPU access), we do an initial exploration of the ℓ_2 regularization strength on one split of the data for each classification task. We find the regularization strength to have minimal effect on performance for decay values of 1e-5, 1e-4, and 1e-3; we set a decay weight of 1e-5 for all remaining experiments.

Readers should keep in mind that the clinical BERT models (Alsentzer et al., 2019) were pretrained on MIMIC-III (Johnson et al., 2016), which may have a small amount of note and/or patient overlap with our MIMIC-IV discharge summary sample. Despite this potential leakage, the clinical BERT models do not consistently outperform the BERT models pretrained using general web data (Devlin et al., 2018). Understanding whether clinical knowledge is necessary to fully understand stigmatizing language in the context of a medical record is left as an open question for future research. Provided sufficient data privacy protections, we also see opportunities to leverage larger generative models.

All experiments were run in a HIPAA-compliant remote computing environment secured with OS-level group permissions. We used servers outfitted with NVIDIA Tesla M60 GPUs (2 x 8 GB VRAM) and Intel Xeon E5-3698 CPUs (2.20 GHz).

C Demographic Differences in Stigmatizing Language (§4.2)

C.1 Experimental Design

We train new clinical BERT models for each of the three classification tasks. This time, we forego cross-validation and instead use a single training, development, and test split. We detach each task's classification head and pass the anchor n-grams through their respective models to extract their internal mean-pooled representation.

Maintaining separation between the three classification tasks, we randomly split the subset of patients whose data was used for training the BERT models into 5 non-overlapping groups and use these groups as folds for cross-validation. Using 4 of the groups for training, an unregularized logistic regression classifier is fit to independently predict race and sex from the internal semantic representations. We evaluate separation using data from the held-out group. This process is repeated 5 times until each patient group has been used as the held-out test group.

The joint race and sex distribution of instances is

JHM		Black or African American		White or Caucasian		Other	
Task	**Class**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Credibility & Obstinacy	Difficult	159 (129)	94 (76)	87 (60)	40 (29)	11 (10)	17 (13)
	Disbelief	160 (133)	142 (117)	59 (46)	47 (39)	8 (7)	18 (17)
	Exclude	20 (18)	20 (17)	20 (13)	11 (9)	3 (3)	3 (2)
Compliance	Negative	714 (499)	480 (324)	187 (146)	104 (81)	22 (20)	41 (29)
	Neutral	107 (102)	87 (81)	43 (37)	31 (29)	4 (4)	4 (3)
	Positive	146 (135)	105 (93)	50 (45)	35 (31)	9 (7)	6 (5)
Descriptors	Exclude	146 (132)	132 (108)	68 (55)	58 (56)	8 (8)	11 (9)
	Negative	253 (172)	254 (189)	134 (72)	144 (89)	17 (11)	34 (18)
	Neutral	78 (69)	51 (50)	54 (52)	32 (29)	6 (5)	8 (8)
	Positive	232 (185)	117 (98)	111 (91)	59 (48)	19 (16)	9 (9)

MIMIC		Black or African American		White or Caucasian		Other	
Task	**Class**	**Female**	**Male**	**Female**	**Male**	**Female**	**Male**
Credibility & Obstinacy	Difficult	35 (32)	48 (47)	177 (167)	189 (177)	31 (29)	32 (31)
	Disbelief	64 (64)	56 (55)	209 (198)	191 (179)	31 (30)	41 (41)
	Exclude	13 (13)	8 (8)	36 (36)	43 (43)	7 (6)	7 (7)
Compliance	Negative	127 (121)	109 (93)	232 (219)	277 (258)	64 (61)	56 (54)
	Neutral	30 (30)	26 (25)	146 (140)	160 (157)	23 (23)	35 (33)
	Positive	23 (23)	23 (22)	81 (79)	93 (90)	23 (22)	21 (19)
Descriptors	Exclude	50 (49)	36 (35)	161 (157)	171 (162)	29 (29)	19 (19)
	Negative	106 (84)	126 (112)	341 (309)	514 (419)	49 (44)	49 (46)
	Neutral	4 (4)	10 (9)	38 (38)	29 (29)	5 (5)	6 (6)
	Positive	33 (33)	13 (13)	157 (152)	105 (104)	37 (35)	20 (20)

Table 6: Joint sex, race, and label distribution for the JHM and MIMIC datasets. The format is "# Examples (# Patients)". These distributions are insufficient for characterizing the extent to which demographic disparities are replicated within our dataset. A more thorough statistical analysis which controls for differences in anchor term usage, repeated measures, underlying conditions, and clinical specialty is necessary to make any substantive claims.

provided in Table 6. Note that we ignore instances in which a patient either declined to report or did not self-report their race or sex. After this exclusion, we are left with 5,129 of the original 5,201 instances for the JHM dataset, and 4,875 of the original 5,043 instances for the MIMIC dataset.

C.2 Baselines

Our clinical datasets represent a concatenation of notes from different specialties. Each speciality has a unique patient demographic pool and thus invites the possibility of conflating the encoding of specialty-specific knowledge with demographic-specific knowledge. For example, OB-GYN notes come specifically from female patients and our sample of JHM pediatric notes come from a population which is 95% black. Encoding the speciality would naturally allow inference of patient demographics.

Additionally, any differences in prevalence of our anchor n-grams between demographic groups may be exploited by the linear classifier. The latter is expected given the extant literature which highlights demographic disparities in usage of stigmatizing language (Beach and Saha, 2021; Beach et al., 2021).

For these reasons, we ground the predictive performance achieved using the semantic representations against simple logistic regression baselines which model one-hot-encoded representations of the anchor n-gram, clinical speciality, and the primary stigmatizing language classification label. A qualitative review of instances in both datasets suggest there are likely additional auxiliary attributes not accounted for here (e.g., diagnoses) that would further explain the encoding of race and sex in the embeddings. For the MIMIC dataset, we consider the service which wrote the discharge summary (e.g., `SURG`, `GYN`, `PSYCH`) to be the speciality.

In Table 7, we include our ability to infer each of these baseline attributes considered within the experiment. The anchor n-grams, task label, and speciality are all predictable from the BERT embeddings, confirming the necessity of the baselines.

Credibility & Obstinacy	JHM					MIMIC				
	Anchor	Label	Speciality	Sex	Race	Anchor	Label	Speciality	Sex	Race
Majority Baseline	0.03 ± 0.00	0.20 ± 0.02	0.11 ± 0.01	0.37 ± 0.01	0.26 ± 0.02	0.02 ± 0.00	0.22 ± 0.01	0.06 ± 0.01	0.33 ± 0.01	0.27 ± 0.01
Anchor	–	0.51 ± 0.05	0.14 ± 0.03	0.50 ± 0.04	0.31 ± 0.05	–	0.51 ± 0.02	0.07 ± 0.01	0.52 ± 0.04	0.27 ± 0.01
Label	0.08 ± 0.01	–	0.11 ± 0.01	0.37 ± 0.01	0.27 ± 0.03	0.09 ± 0.01	–	0.06 ± 0.01	0.51 ± 0.05	0.27 ± 0.01
Speciality	0.07 ± 0.02	0.31 ± 0.02	–	0.44 ± 0.04	0.36 ± 0.04	0.05 ± 0.01	0.32 ± 0.03	–	0.55 ± 0.05	0.28 ± 0.02
Anchor × Label	–	–	0.18 ± 0.05	0.50 ± 0.03	0.31 ± 0.05	–	–	0.07 ± 0.01	0.49 ± 0.04	0.28 ± 0.02
Anchor × Speciality	–	0.52 ± 0.06	–	0.51 ± 0.04	0.38 ± 0.03	–	0.60 ± 0.07	–	0.51 ± 0.02	0.28 ± 0.02
Label × Speciality	0.11 ± 0.02	–	–	0.47 ± 0.04	0.38 ± 0.04	0.10 ± 0.02	–	–	0.54 ± 0.04	0.27 ± 0.01
Anchor × Label × Speciality	–	–	–	0.54 ± 0.01	0.35 ± 0.03	–	–	–	0.51 ± 0.02	0.29 ± 0.02
Embedding	0.76 ± 0.05	0.95 ± 0.03	0.24 ± 0.03	0.76 ± 0.02	0.34 ± 0.02	0.92 ± 0.02	0.87 ± 0.03	0.11 ± 0.01	0.75 ± 0.02	0.30 ± 0.03
Embedding (Gender Neutral)	0.77 ± 0.06	0.93 ± 0.02	0.25 ± 0.04	0.59 ± 0.02	0.34 ± 0.06	0.92 ± 0.01	0.86 ± 0.06	0.10 ± 0.01	0.49 ± 0.03	0.33 ± 0.02

Compliance	JHM					MIMIC				
	Anchor	Label	Speciality	Sex	Race	Anchor	Label	Speciality	Sex	Race
Majority Baseline	0.01 ± 0.00	0.28 ± 0.01	0.08 ± 0.00	0.37 ± 0.02	0.29 ± 0.01	0.01 ± 0.00	0.24 ± 0.01	0.05 ± 0.00	0.33 ± 0.01	0.26 ± 0.01
Anchor	–	0.59 ± 0.02	0.18 ± 0.04	0.42 ± 0.02	0.29 ± 0.01	–	0.66 ± 0.02	0.05 ± 0.00	0.54 ± 0.02	0.27 ± 0.02
Label	0.03 ± 0.00	–	0.14 ± 0.01	0.37 ± 0.02	0.29 ± 0.01	0.03 ± 0.01	–	0.05 ± 0.00	0.47 ± 0.03	0.26 ± 0.01
Speciality	0.03 ± 0.00	0.28 ± 0.01	–	0.53 ± 0.04	0.29 ± 0.01	0.02 ± 0.01	0.34 ± 0.03	–	0.55 ± 0.02	0.26 ± 0.01
Anchor × Label	–	–	0.27 ± 0.02	0.46 ± 0.03	0.30 ± 0.01	–	–	0.07 ± 0.01	0.52 ± 0.03	0.31 ± 0.02
Anchor × Speciality	–	0.62 ± 0.05	–	0.54 ± 0.02	0.35 ± 0.03	–	0.67 ± 0.03	–	0.56 ± 0.02	0.30 ± 0.02
Label × Speciality	0.08 ± 0.01	–	–	0.53 ± 0.05	0.32 ± 0.02	0.08 ± 0.02	–	–	0.56 ± 0.03	0.28 ± 0.01
Anchor × Label × Speciality	–	–	–	0.54 ± 0.03	0.36 ± 0.02	–	–	–	0.54 ± 0.01	0.30 ± 0.02
Embedding	0.77 ± 0.04	1.00 ± 0.00	0.38 ± 0.05	0.57 ± 0.01	0.36 ± 0.02	0.86 ± 0.04	1.00 ± 0.00	0.13 ± 0.04	0.56 ± 0.03	0.33 ± 0.02
Embedding (Gender Neutral)	0.74 ± 0.04	1.00 ± 0.00	0.39 ± 0.05	0.52 ± 0.01	0.35 ± 0.01	0.85 ± 0.03	1.00 ± 0.00	0.12 ± 0.02	0.50 ± 0.04	0.34 ± 0.02

Descriptors	JHM					MIMIC				
	Anchor	Label	Speciality	Sex	Race	Anchor	Label	Speciality	Sex	Race
Majority Baseline	0.01 ± 0.00	0.14 ± 0.01	0.10 ± 0.01	0.35 ± 0.02	0.26 ± 0.01	0.00 ± 0.00	0.18 ± 0.00	0.05 ± 0.00	0.34 ± 0.02	0.28 ± 0.00
Anchor	–	0.83 ± 0.03	0.22 ± 0.02	0.50 ± 0.02	0.30 ± 0.03	–	0.87 ± 0.03	0.13 ± 0.01	0.56 ± 0.03	0.28 ± 0.01
Label	0.07 ± 0.00	–	0.10 ± 0.01	0.46 ± 0.07	0.26 ± 0.01	0.03 ± 0.00	–	0.06 ± 0.01	0.58 ± 0.03	0.28 ± 0.00
Speciality	0.01 ± 0.00	0.28 ± 0.03	–	0.58 ± 0.03	0.32 ± 0.03	0.03 ± 0.00	0.27 ± 0.03	–	0.44 ± 0.03	0.28 ± 0.00
Anchor × Label	–	–	0.30 ± 0.02	0.53 ± 0.02	0.32 ± 0.04	–	–	0.13 ± 0.01	0.56 ± 0.02	0.29 ± 0.01
Anchor × Speciality	–	0.84 ± 0.03	–	0.56 ± 0.04	0.34 ± 0.02	–	0.86 ± 0.02	–	0.57 ± 0.02	0.31 ± 0.02
Label × Speciality	0.09 ± 0.01	–	–	0.58 ± 0.04	0.32 ± 0.03	0.11 ± 0.01	–	–	0.57 ± 0.04	0.28 ± 0.00
Anchor × Label × Speciality	–	–	–	0.55 ± 0.03	0.36 ± 0.02	–	–	–	0.58 ± 0.02	0.30 ± 0.02
Embedding	0.82 ± 0.06	1.00 ± 0.00	0.45 ± 0.02	0.61 ± 0.04	0.34 ± 0.03	0.91 ± 0.02	1.00 ± 0.00	0.24 ± 0.05	0.58 ± 0.02	0.33 ± 0.02
Embedding (Gender Neutral)	0.82 ± 0.07	1.00 ± 0.00	0.44 ± 0.04	0.52 ± 0.03	0.34 ± 0.02	0.90 ± 0.03	1.00 ± 0.00	0.24 ± 0.04	0.54 ± 0.03	0.31 ± 0.02

Table 7: Macro F1 ($\mu \pm \sigma$) for each attribute considered in §4.2. Higher inference performance suggests an attribute is more strongly encoded by (or correlated with) a given feature set. Differences in the prevalence of racial groups and sexes across auxiliary attributes (e.g., speciality, labels) can be exploited when inferring race and sex from the anchor embeddings.

C.3 Demographic-Neutral Substitutions

Sex During an initial run of the experiment, we recognized that patient sex could be easily inferred from the semantic representations due to the cues from gender-specific language. We adopt a naive approach to mitigate the presence of overt gender-informative language affecting conclusions within the demographic inference experiments. We replace gendered pronouns (e.g., he, herself), identifiers of sex (e.g., male, Mrs. Smith), and terms with non-uniform gender associations (e.g., husband, wife). The full mapping of substitutions is provided below in Table 8.

There are two limitations with this approach. First, we do not make substitutions for any patient names in the text. Second, we do not address any grammatical issues that arise after substitution of a gendered word (e.g., “he denies” $\rightarrow$ “they denies”). In practice, the former implies that true amount of the sex-related information encoded in the learned embeddings may be lower than current estimates suggest. This case would only further strengthen our current conclusions. Regarding the latter, we find that any grammatical inconsistencies do not affect our ability to infer the stigma labels associated with each anchor embedding (Table 7).

Race We briefly explored using rules to obfuscate racial identifiers as well (e.g, “43 y.o. Asian”). We found this procedure difficult to perform automatically (e.g., “wearing black T-shift”) and likely to be a low-yield process based on a qualitative review of the instances in both datasets. For this reason, we opted not to include any race-neutral substitutions. Nonetheless, the lack of obfuscation should be noted while interpreting our results.

D Dataset Differences in Stigmatizing Language (§4.3)

D.1 Experimental Design

We use the clinical BERT models trained during the §4.1 experiments to evaluate domain-transfer. That is, we take the clinical BERT models (with anchor pooling) trained within each cross-validation fold and apply them to the test set of the oppo-

Original	Replacement
He, She	They
Him, Her	Them
His, Hers	Their
Himself, Herself	Themselves
Male, Female, Girl, Boy, Man, Woman	Person
Mr. XX, Ms. XX, Mrs. XX, Miss. XX	Patient
Husband, Wife	Partner

Table 8: Gender-informative words and their associated gender-neutral substitutions.

site dataset (JHM $\rightarrow$ MIMIC, MIMIC $\rightarrow$ JHM). We *do not* modify or otherwise tune the existing models to improve transfer performance, with the primary goal being to understand differences in stigmatizing language usage between datasets (not to optimize generalization). To facilitate our qualitative analysis, we cache all test-set predictions and organize them into four groups based on whether the in-domain (source = target) and out-of-domain (source $\neq$ target) models characterized them correctly.

D.2 Error Distribution

Errors made by both the in-domain *and* out-of-domain models are those which appear to be a consequence of task difficulty and model underspecification. Examples include hypothetical statements (e.g., "if the patient declines") and instances containing both positive and negative sentiment (e.g, "disinhibited, but charming").

Errors made by the out-of-domain model, but not the in-domain model, are a consequence of distribution shift. The two notable areas of shift include 1) the prevalence of statements regarding individuals other than the patient (e.g., family), and 2) differences in class priors conditioned on each anchor. The latter is sometimes the result of speciality-specific nuances (e.g., psychiatry notes include more self-descriptions).

Errors made by the in-domain model, but *not* the out-of-domain model, are generally a consequence of the out-of-domain model having seen more training examples containing the test example's anchor.

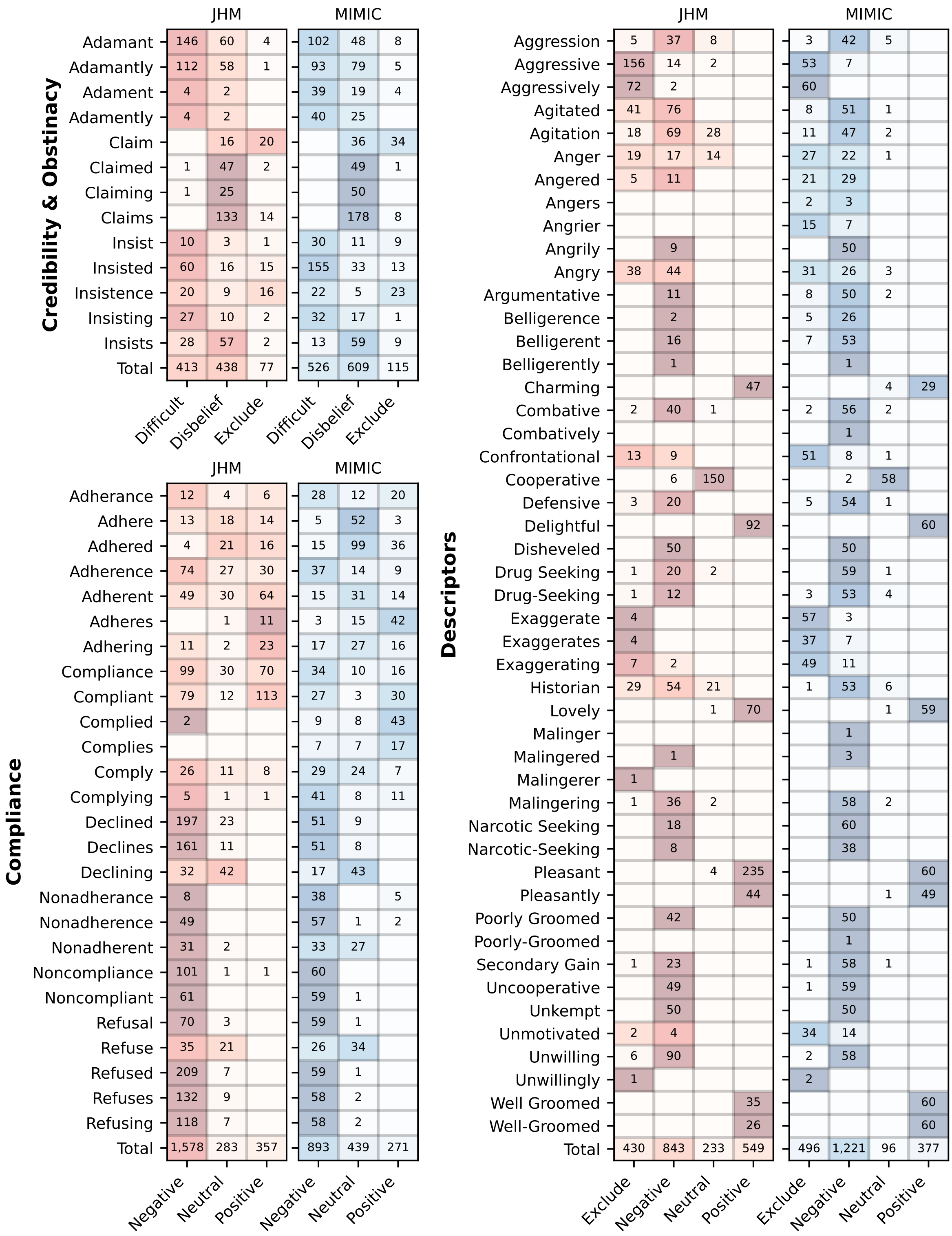

Figure 1: Joint anchor and label distribution for each task.

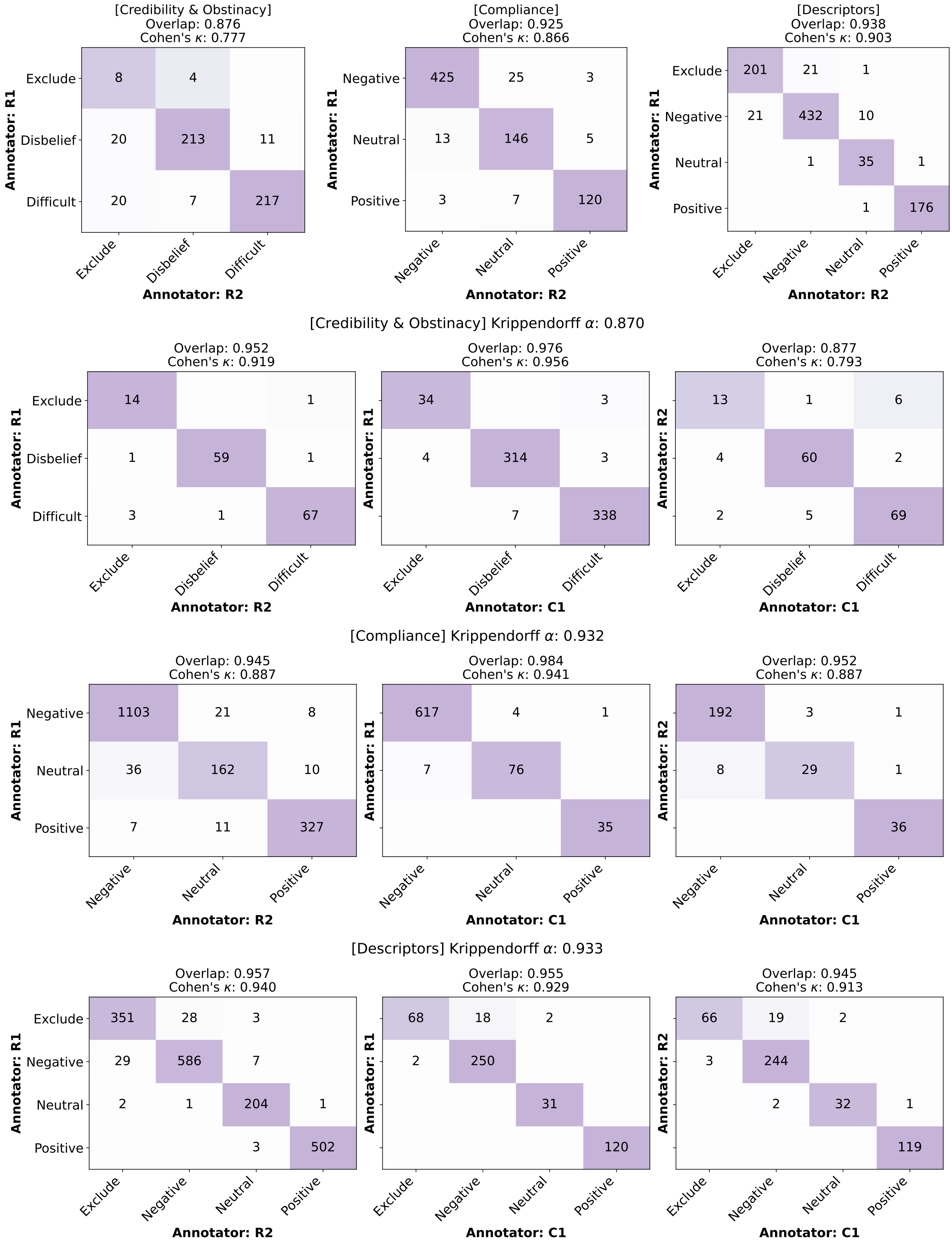

Figure 2: Pairwise interannotator agreement for the MIMIC dataset (first row) and JHM dataset (bottom 3 rows).

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

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

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

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

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

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

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

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

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

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

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

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

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

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

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

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

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.