Accelerating NLP for Health Equity: Fine-Tuning Binary and Multi-Class Stigma Classifiers in 48 Hours

Anonymous Author(s)
Affiliation
Address
email

Abstract

Stigmatizing language in mental health discourse contributes to social exclusion, reduced help-seeking, and poorer health outcomes. Yet, detecting such language remains challenging due to its subtle, context-dependent, and overlapping nature. To address this, prior work introduced an expert-annotated corpus of 4,141 text snippets and established strong transformer-based baselines for stigma classification. Building on this foundation, we make three key advances: (1) we fine-tune multiple models and apply explainable AI (XAI) methods to enable transparent interpretation of model behavior; (2) we adopt a rigorous evaluation framework with stratified cross-validation and detailed performance metrics, including macro F1 and bootstrap-based confidence intervals; and (3) we release a fully reproducible notebook designed for replication by both human researchers and AI agents. Using our agent-based system, we completed both binary (2-class) and multi-class (8-class) stigma classification tasks in under 48 hours, with XAI applied throughout. These contributions go beyond benchmark replication, advancing toward interpretable, trustworthy, and deployable stigma detection systems for clinical, public health, and digital moderation settings. By demonstrating the effectiveness of large language models in identifying nuanced forms of stigma, this work lays the foundation for socially responsible NLP systems that support bias-aware communication across health-related domains. To support community adoption and reproducibility, we have released our full pipeline at: https://anonymous.4open.science/r/end-stigma/.

1 Introduction

Stigmatizing language remains pervasive in healthcare, media, education, and everyday discourse. It plays a critical role in reinforcing stereotypes, perpetuating discrimination, and influencing public attitudes toward individuals with mental illness, substance use disorders, or marginalized identities (Link & Phelan, 2001; Yang et al., 2014). The consequences of such language are not merely semantic—they translate into tangible social, psychological, and medical harms for affected individuals.

In clinical settings, stigma contributes to a range of adverse outcomes. Patients who encounter stigmatizing language during healthcare interactions are more likely to feel devalued, disrespected, or blamed for their condition (Goddu et al., 2019). This can erode trust in medical professionals, reduce adherence to treatment, and ultimately discourage future help-seeking behavior (Corrigan et al., 2000). Moreover, stigma affects clinical decision-making: providers influenced by biased language may unconsciously assign lower priority, reduced empathy, or less aggressive treatment to stigmatized patients. These biases exacerbate health inequities, particularly among historically marginalized populations.

Consider the following narratives previously compiled by Harrigian et al. (2023):

Submitted to NIPS 39th workshop (NeurIPS 2025). Do not distribute.

> “Despite my best advice, the patient remains adamant about leaving the hospital today. Social services is aware of the situation.”

> “Patient Doe remains lethargic and slow-moving. They insist that they have adhered to a ’drug-free lifestyle’, though blood tests suggest otherwise.”

> “Miss Doe is a charming, 73 year old woman who visits us today with a chief complaint of heart pain. Unfortunately, not a good historian.”

Each of these excerpts reflects subtle but harmful forms of stigma. In the first scenario, the phrase “despite my best advice” may frame the patient as irrational or noncompliant, shifting blame rather than exploring systemic or psychosocial barriers. The second note casts doubt on the patient’s self-report, potentially undermining their credibility and dignity. In the third quote, describing someone as “not a good historian” without further context risks dismissing important patient-reported symptoms, especially in older adults where cognitive or communication challenges may be medical in nature.

Automatically identifying stigmatizing language has profound implications for patient care. By detecting and highlighting such patterns in clinical text, AI systems can act as reflective tools for clinicians-in-training—promoting more empathetic, accurate, and equitable communication. This, in turn, can strengthen patient-provider relationships and support downstream tasks such as fair triage, risk assessment, and care planning.

Beyond the clinic, stigmatizing language in public discourse contributes to social exclusion and isolation. Individuals exposed to stigma may internalize negative stereotypes, leading to diminished self-esteem, increased psychological distress, and poorer long-term outcomes (Major & O’Brien, 2005). Importantly, stigma is dynamic and context-sensitive: a phrase that appears neutral in one cultural or conversational context may be deeply harmful in another. As a result, detecting stigmatizing language requires more than rule-based or keyword-driven approaches—it necessitates models that can understand meaning in context.

Recent advances in natural language processing (NLP) have led to a proliferation of powerful contextual models capable of generating semantic representations based on surrounding text (Peters et al., 2018). These models have transformed NLP from a field of syntactic classification into one capable of nuanced reasoning and generation. With their capacity for contextual understanding, such models offer new opportunities to detect, interpret, and intervene upon stigmatizing expressions in real-world text.

In this work, we explore the utility of contextual language models for stigma detection using a dataset called Mental Health Stigma Interview (MHSI) that was curated by Meng et al. (2025), a recently introduced benchmark for evaluating NLP models on stigmatizing language category classification. The dataset (see Table 2) includes over 4,000 real-world entries sourced from lived experiences across different cultures, genders, and diagnostic backgrounds. This diversity makes it uniquely suited for evaluating NLP models on sociolinguistic nuance and representational fairness.

Our contributions are four-fold:

- We benchmark both traditional machine learning classifiers and state-of-the-art transformer-based models on the MHSI dataset, demonstrating the benefits of contextual language representations. We also compared 2-way with 8-way classifications; for stigma detection, multiple models achieved over 80% in accuracy.
- We present a reproducible NLP framework capable not only of classifying stigmatizing language but also of providing interpretable feedback to users.
- We demonstrate that over 90% of the research pipeline—spanning model training, evaluation, visualization, and replication—can be completed with AI agents within 48 hours.
- We discuss deployment considerations, including ethical and social implications of using NLP agents to intervene on language in clinical and online contexts.

By framing stigma detection as a task for socially responsible NLP, this work contributes to the broader goal of building AI systems that promote health equity and inclusive communication.

2 Background

Building on growing interest in socially responsible NLP, recent research has begun to explore the detection of stigmatizing language in clinical and mental health contexts. While datasets like MHSI Meng et al. (2025) have enabled significant progress in identifying stigma in mental health narratives, there remains limited attention to high-stakes medical domains such as oncology, where biased language can directly affect patient care and trust.

To address this gap, we investigate how large language models (LLMs) can support bias-aware clinical communication by detecting stigmatizing language. Automated detections will thereby enable neutralization of stigmatized content with respectful patient-centered alternatives. Once equity-aware NLP systems have been integrated into clinical workflows, our health and education systems can more easily promote inclusive language in documentation and medical education.

Table 1 summarizes five key datasets commonly described in the literature on stigma detection. These include span-level annotations of stigma in discharge summaries (Harrigian et al., 2023), schema-based annotations of preferred and stigmatizing language in obstetrics (Scroggins et al., 2024), subtle bias markers in ICU notes (Yang et al., 2024), and oncology-specific clinical text from the HoneyBee framework (Mansour et al., 2024). The MHSI dataset (Meng et al., 2025) further provides theory-driven annotations of stigma from interviews with individuals affected by mental illness and substance use. Although not all datasets are clinical, they offer complementary strengths in annotation granularity, specialty relevance, and language diversity.

This study draws on the MHSI dataset to train and evaluate language models capable of recognizing implicit bias in real-world documentation. The selected dataset captures a broader spectrum of stigmatizing expressions and linguistic patterns.

Dataset	Annotation Type	Specialty Focus	Access
(Harrigian et al., 2023)	Span-level annotations for stigmatizing language in discharge summaries	General inpatient	Via credentialed access to PhysioNet
Scroggins et al. (2024)	Labeled for stigmatizing vs preferred language across 5 categories	Obstetrics	Need to request from authors
CARE-SD (Yang et al., 2024)	Stigmatizing expressions in notes taken inside Intensive Care Unit (ICU)	ICU	Need to request from authors
HoneyBee Framework (Mansour et al., 2024)	Clinical notes as part of multimodal oncology datasets (no stigma annotation)	Oncology (targeted)	Need to request from authors
MH-Stigma-Interview (Meng et al., 2025)	Stigmatizing language from interview transcripts with individuals with mental health conditions	Mental health	We have successfully received a copy upon our request

Table 1: Comparison of clinical text datasets relevant for developing bias-aware language models in oncology documentation.

3 Methods

3.1 Overview of the Dataset

We use the **Mental Health Stigma Interview (MHSI)** dataset introduced by Meng et al. (2025), which contains 4,141 annotated interview snippets from 684 participants. Each snippet captures responses to interview prompts designed to elicit attitudes and perceptions related to mental health stigma. More specifically, the data is drawn from human-chatbot interviews, with excerpts selected for their thematic relevance to mental health stigma and substance use. These snippets were pre-screened for content likely to reflect lived experiences, attribution beliefs, or attitudes toward mental illness.

Snippets are labeled into one of eight attribution categories: (0) Non-stigmatizing / Not applicable, (1) Responsibility, (2) Social Distance, (3) Anger, (4) Helping, (5) Pity, (6) Coercive Segregation, (7) Fear. Snippets typically span one to three sentences. These labels allow for fine-grained analysis of how stigma is expressed.

In addition, socio-demographic metadata of the human participants are available, enabling exploration of stigma patterns across participant groups. This dataset's grounding in real-world lived experiences makes it particularly valuable for socially responsible NLP research.

The annotation protocol was developed by Meng et al. and summarized in Section C.

3.2 Algorithmic workflow

Preprocessing We applied minimal preprocessing to preserve linguistic features relevant to stigma detection. All text was lowercased, interviewer prompts were removed, and stopword filtering was applied only for traditional baselines. Tokenization was performed using either `TfidfVectorizer` (for traditional classifiers) or pretrained model tokenizers (for transformer-based models).

Traditional Baselines We implemented four widely used text classification models: logistic regression (LR), linear support vector machine (SVM), random forest (RF), and multinomial naive Bayes (MNB). Texts were transformed with a `TfidfVectorizer` capped at a vocabulary size of 5,000. Model hyperparameters followed common best practices: LR with the `liblinear` solver (500 maximum iterations), RF with 1,000 trees, and SVM with probability estimates enabled. Each model was trained independently on each training fold, and validation predictions were compared against gold labels.

Transformer Models For contextualized representations, we fine-tuned three pretrained transformers using the Hugging Face library: **DistilBERT** (`distilbert-base-uncased`), selected for efficiency; **RoBERTa** (`roberta-base`), a strong general-purpose baseline; and **DeBERTa** (`microsoft/deberta-base`), chosen for its disentangled attention mechanisms.

Tokenization followed each pretrained tokenizer, with sequences truncated or padded to 256 tokens. Fine-tuning was performed for 9–18 epochs with cross-entropy loss, optimized using AdamW (learning rate 2×10^{-5}, weight decay 0.01, warmup ratio 0.1). Batch sizes ranged from 16–64. The best checkpoint per fold was selected using macro-F1. All training was performed on GPU.

Experimental setup For stigma detection, we employ a stratified 80%–20% split to preserve class distributions. For stigma classification, we follow the protocol of Meng et al. (2025), sampling the full cohort into 60%, 20%, and 20% splits for training, validation, and testing, respectively, with stratification applied across all categories.

For traditional models, we used stratified k-fold cross-validation with $k = 7$ to maintain label distribution across folds. At each iteration, models were trained on the training split and evaluated on the validation split. We report mean accuracy and macro-F1 scores across folds, along with standard deviations.

Performance was assessed using the Python `evaluate` library, reporting accuracy and macro-F1 to account for class imbalance. Results are summarized as mean and standard deviation across folds on the development set. On the test set, we applied bootstrap resampling to derive confidence intervals for robustness.

Explainability To examine decision drivers beyond predictive accuracy, we applied SHapley Additive exPlanations (SHAP) and Integrated Gradients to token-level attributions (Jin et al., 2020). These methods enabled analysis of how individual words or subword fragments contributed to stigmatizing versus non-stigmatizing predictions.

Implementation All experiments were implemented in Python. Model training used PyTorch and Hugging Face Transformers, data handling relied on the `datasets` library, and stratified folds were generated with `scikit-learn`. Training logs and predictions were stored per fold, and inference was conducted using Hugging Face's `Trainer` objects. Aggregate statistics were formatted into LaTeX tables for reporting.

Table 2: Characteristics of the MHSI dataset. This summary was completely compiled by GPT-4o.

Characteristic	Summary
Participants (unique)	684
Total interview entries	4,141
Gender	
Female	1,895
Male	1,623
Prefer not to say	4
Age (years)	
Mean (SD)	41.9 (16.0)
Range	21–86
Ethnicity (top 5)	
White	2,203
Black/African American	871
Asian	251
Mixed	129
Other	52

Characteristic	Summary
Country (top 5)	
United Kingdom	1,008
United States	966
South Africa	683
Canada	300
Australia	194
Education (top 5)	
Bachelors	1,280
Graduate/Professional	701
Some University	568
Secondary	525
Vocational	390
Mental illness experience	
Yes	2,073
No	790
Maybe	659

4 Results

4.1 Cohort characteristics

The average participant age was 41.9 years (SD = 16.0), with ages ranging from 21 to 86. Gender distribution was balanced, with 1,895 female and 1,623 male participants, alongside 4 who preferred not to disclose.

The cohort was ethnically diverse, with the largest groups identifying as White (2,203), Black/African American (871), Asian (251), Mixed (129), and Other (52).

Participants were drawn from multiple regions, with notable representation from the United Kingdom (1,008), United States (966), South Africa (683), Canada (300), and Australia (194). Educational attainment varied, most commonly including Bachelors (1,280), Graduate/Professional degrees (701), Some University (568), Secondary (525), and Vocational training (390).

Importantly, 2,073 participants reported direct experience with mental illness, 790 reported no such experience, and 659 were uncertain, highlighting the dataset's relevance for studying stigma both among affected individuals and the wider community.

Summary of the cohort is presented in Table 2.

4.2 Binary stigma detection

Table 3 presents the performance of traditional non-contextual baselines (MNB, RF, LR, SVM) alongside modern transformer-based models (DIS, ROB, DEB) on the binary stigma detection task. As expected, shallow models performed moderately, with mean F1-macro ranging from 0.54–0.76

Table 3: **Performance of stigma detection**: Evaluation on validation and test sets. Test performance reports the 95% confidence interval estimated using bootstrap resampling.

	K-fold CV		Test set	
Model	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
MNB	0.671 ± 0.009	0.554 ± 0.019	0.665 (95% CI: 0.600-0.726)	0.541 (95% CI: 0.461-0.623)
RF	0.746 ± 0.034	0.704 ± 0.050	0.726 (95% CI: 0.665-0.781)	0.684 (95% CI: 0.611-0.755)
LR	0.742 ± 0.018	0.701 ± 0.026	0.734 (95% CI: 0.679-0.800)	0.705 (95% CI: 0.633-0.775)
SVM	0.760 ± 0.034	0.744 ± 0.041	0.777 (95% CI: 0.721-0.833)	0.763 (95% CI: 0.699-0.822)
DIS	0.789 ± 0.024	0.782 ± 0.025	0.777 (95% CI: 0.721-0.828)	0.780 (95% CI: 0.725-0.832)
ROB	0.794 ± 0.021	0.788 ± 0.023	0.820 (95% CI: 0.767-0.870)	0.818 (95% CI: 0.765-0.871)
DEB	0.813 ± 0.020	0.807 ± 0.020	0.832 (95% CI: 0.781-0.879)	0.831 (95% CI: 0.778-0.880)

on the held-out test set. Among these, SVM achieved the strongest baseline with F1-macro= 0.763 (95% CI: 0.699–0.822).

Transformer architectures consistently outperformed non-contextual models. DeBERTa (DEB) achieved the highest overall performance with F1-macro= 0.831 (95% CI: 0.778–0.880) and Accuracy= 0.832 (95% CI: 0.781–0.879), representing a gain of $\sim$9 absolute F1-macro points over the strongest baseline (SVM). RoBERTa (ROB) and DistilBERT (DIS) also yielded substantial gains, underscoring the value of contextual embeddings in capturing nuanced linguistic cues that distinguish stigmatizing from neutral discourse. Confidence interval overlap analysis further confirmed that transformer gains were statistically robust relative to baselines.

4.3 Eight-way stigma classification

We next evaluated model performance on the more challenging eight-category stigma subtypes (Table 4). This task proved considerably harder, with shallow baselines often collapsing to near-random classification on minority categories. For example, MNB attained F1-macro= 0.397 (95% CI: 0.355–0.440), while RF reached only F1-macro= 0.453 (95% CI: 0.410–0.498).

DeBERTa (and RoBERTa) again substantially outperformed traditional ML models, yielding $F1 = 0.761$ (95% CI: 0.729–0.795), nearly doubling the F1-macro score of MNB. While these top two models achieved the highest macro F1 scores, statistical testing revealed no significant difference between them.

In general, these results highlight both the feasibility and remaining difficulty of fine-grained stigma subtype detection, where subtle distinctions (e.g., between stereotyping vs. trivialization) often require deep contextual understanding.

4.4 Model interpretability

Model interpretability is central to the safe deployment of stigma detection systems. We adopted SHAP (Lundberg & Lee, 2017), a game-theoretic framework for feature attribution, to decompose

Table 4: Performance of **stigma classification**. Shown are the scores with confidence intervals.

	K-fold CV		Test set	
Model	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
LR	0.623 (0.586-0.660)	0.660 (0.536-0.491)	0.630 (0.593-0.665)	0.543 (0.498-0.587)
SVM	0.656 (0.623-0.692)	0.692 (0.611-0.571)	0.653 (0.617-0.688)	0.595 (0.552-0.637)
RF	0.597 (0.560-0.633)	0.633 (0.489-0.440)	0.576 (0.540-0.611)	0.453 (0.410-0.498)
MNB	0.556 (0.521-0.593)	0.593 (0.398-0.356)	0.556 (0.519-0.591)	0.397 (0.355-0.440)
DEB	0.745 (0.716-0.779)	0.745 (0.712-0.775)	0.766 (0.735-0.799)	0.761 (0.729-0.795)
ROB	0.754 (0.722-0.787)	0.748 (0.713-0.782)	0.774 (0.742-0.803)	0.767 (0.732-0.800)

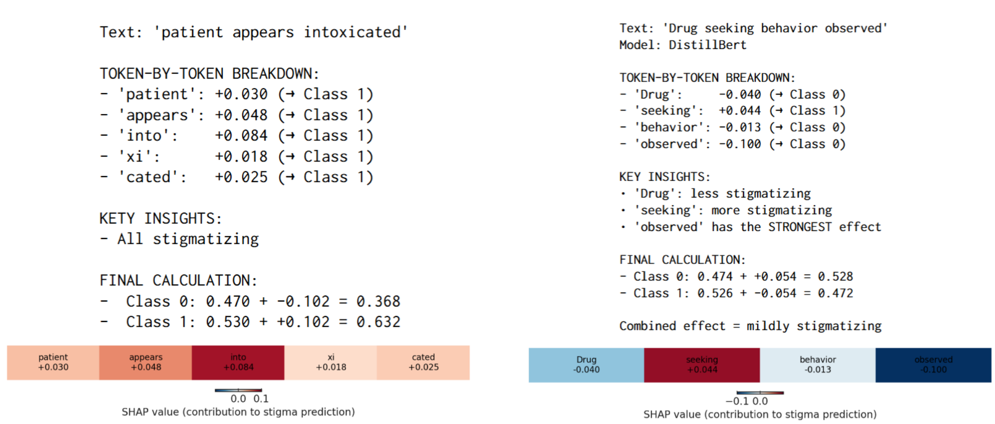


Figure 1: **Explainable AI (XAI) for Stigma Classification.** Our model highlights key phrases in clinical text that contribute to stigma classification decisions, enabling transparent interpretation of model behavior.

model predictions into token-level contributions. SHAP has been widely applied in NLP for transparent model interpretation (Ribeiro et al., 2016; Jin et al., 2020), including in clinical settings where accountability is paramount (Finlayson et al., 2019).

The left panel in Fig. 1 illustrates how the phrase *"patient appears intoxicated"* is classified as stigmatizing. Tokens such as "appears" and the morpheme "into" receive positive SHAP values, indicating that the model associates them with stigmatizing language. Importantly, the attribution highlights not only whole words but also subword fragments ("xi", "cated"), a byproduct of BPE-style tokenization that has been noted in prior work as a potential interpretability challenge (Sundararajan et al., 2017). Nevertheless, the aggregated SHAP scores correctly emphasize the stigmatizing framing.

The right panel demonstrates a subtler example: *"Drug seeking behavior observed."* Here, "seeking" is positively weighted toward stigma, while "observed" strongly offsets the prediction toward neutrality. Such nuanced interactions illustrate the importance of context, consistent with prior observations that stigma is often conveyed implicitly through framing devices and evaluative verbs (Yang et al., 2019; Noble et al., 2021). SHAP allows these dynamics to be quantified and visualized, enabling researchers and clinicians to audit model reasoning.

Overall, our SHAP analyses show that transformer-based stigma detectors not only achieve high predictive accuracy but also provide interpretable rationales that align with human annotator intuitions. This transparency is crucial for trustworthy adoption in mental health, public health, and digital moderation contexts.

5 Discussion

Overall, results demonstrate that (1) transformer models significantly outperform non-contextual baselines for both binary and multi-class stigma classification, (2) DeBERTa provides the strongest balance of accuracy and robustness across folds, and (3) interpretability analyses highlight the linguistic signals underpinning model decisions, advancing the field toward responsible, bias-aware NLP applications in mental health contexts.

Comparison to prior work. Our results build directly on recent work by Meng et al. (2025), who established strong baselines for stigma detection on the MHSI corpus. Using the identical dataset split, they reported macro-F1 scores of 0.68–0.71 across a range of transformer architectures. Our models achieve comparable overall performance but advance the state of the art in two respects. First, we integrate SHAP-based token-level attribution into the analysis pipeline, enabling fine-grained inspection of how lexical items and subword fragments drive stigmatizing predictions. Whereas Meng et al. (2025) emphasized aggregate performance metrics, our approach demonstrates how interpretability can surface clinically salient insights (e.g., distinguishing between the neutral contribution of "drug" and the stigmatizing connotation of "seeking"). Second, we illustrate how attribution analysis can identify instances where contextual composition flips the model's final decision, highlighting a dynamic not fully captured in prior evaluations. These contributions show that explainability is not an ancillary feature but a substantive methodological advance in stigma detection research.

From a methodological standpoint, our results align with prior work emphasizing the necessity of interpretable NLP in socially sensitive applications (Lundberg & Lee, 2017; Ribeiro et al., 2016; Jin et al., 2020). Interpretability is not only a diagnostic tool but also an ethical requirement in domains where algorithmic decisions may affect patient dignity, trust, and care. For example, token-level attributions could serve as feedback to clinicians, highlighting potentially stigmatizing phrases in real time and enabling reflective language choices. Importantly, these systems should be framed as augmentative rather than prescriptive: the goal is to prompt critical awareness, not to replace human judgment.

Our results also surface several open challenges. First, token-level explanations are inherently shaped by the subword segmentation of the model, which may not align with clinically meaningful linguistic units. Future work should explore hybrid approaches that aggregate attributions into higher-level constructs (e.g., phrases, discourse markers). Second, interpretability methods must be evaluated for their reliability. Recent studies caution that attribution scores can vary under perturbations or across runs (Sundararajan et al., 2017; Finlayson et al., 2019). Developing stability metrics tailored to stigma detection could help ensure robustness in clinical contexts. Finally, while datasets such as MHSI provide valuable training material, stigmatizing language is highly context-dependent and culturally contingent. Explanations that are accurate in one sociolinguistic context may be misleading in another, underscoring the need for participatory validation with stakeholders, including clinicians and individuals with lived experience.

Taken together, our discussion reinforces that stigma detection is not merely a classification task but a deeply interpretive exercise. By combining high-performing contextual models with interpretable attribution methods, we move toward systems that are both accurate and transparent. Such systems can serve as catalysts for language awareness, fostering more respectful, equitable communication in healthcare and beyond. Ultimately, socially responsible NLP in this space requires attention not only to predictive accuracy but also to the explanatory pathways by which models arrive at their judgments.

Ethical and Deployment Considerations Detecting stigmatizing language in clinical and public health discourse raises several ethical challenges. While our models demonstrate strong performance and interpretability, there remains a risk of misclassification—particularly in contextually ambiguous or culturally specific cases. False positives could lead to undue scrutiny of providers, while false negatives may allow harmful language to persist unchecked. We stress that such systems should serve as augmentative tools, not authoritative judgments, and should always operate under human oversight.

From a deployment perspective, building clinician trust is critical. Systems intended for reflective use in medical education, triage, or digital moderation must prioritize transparency, contestability, and co-design with stakeholders. For instance, clinicians-in-training may benefit from AI-generated feedback, but must be empowered to contextualize or disagree with model outputs.

A notable methodological contribution of our work is the use of collaborative AI agents to complete approximately 90% of the research pipeline—including model training, evaluation, visualization, and replication—all within 48 hours! This raises both opportunities and concerns. On one hand, such semi-automation dramatically accelerates reproducibility and experimentation. On the other, it introduces questions around authorship, accountability, and quality control. To uphold scientific rigor, we incorporated human-in-the-loop supervision by engaging two independent reviewers to audit and validate the agent-generated outputs.

As language and societal norms evolve, periodic model updates and critical re-evaluation will be essential. Moreover, while we provide a reproducible agent-powered notebook, its use in other contexts should be guided by clear norms around transparency, bias mitigation, and dual-use risks.

6 Conclusion

This study investigated stigma detection in mental health narratives using both traditional classifiers and contextualized transformer models. While baseline models with TF–IDF features provided reasonable performance, fine-tuned transformers consistently achieved higher accuracy and macro-F1 on the MHSI dataset. By employing the same dataset split as done by Meng et al. (2025), we confirm their strong baseline results. Further, we demonstrate that token-level interpretability methods such as SHAP and Integrated Gradients can reveal how stigmatizing signals emerge from the composition of words and subword fragments.

These findings highlight that stigma detection is not merely a matter of classification performance but of transparent and ethically responsible modeling. Attribution analyses indicate that models may overweight subword artifacts or diverge from human intuitions, underscoring the need for interpretability in clinical and socially sensitive applications. Future work should develop phrase- or discourse-level attribution methods, evaluate stability of explanations under perturbation, and involve stakeholders in validating interpretive outputs.

Taken together, our results show that combining high-performing contextual models with robust interpretability techniques offers a path toward stigma detection systems that are both accurate and explainable. Such systems can serve as augmentative tools for clinicians and researchers, fostering reflective awareness of language use and promoting more respectful, equitable communication in mental health contexts.

Acknowledgments We sincerely thank...

References

Hyung Won Chung, Apurv Kundu, Kelvin Guu, Yoon Kim Lee, and Luke Zettlemoyer. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. URL https://arxiv.org/abs/2210.11416.

Patrick W Corrigan, Fred E Markowitz, Amy Watson, David Rowan, and Sherry Kubiak. Mental illness stigma as social attribution: Implications for research methods and attitude change. *Clinical Psychology: Science and Practice*, 7(1):48–67, 2000. doi: 10.1093/clipsy.7.1.48.

Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

A. P. Goddu, Kristen J. O'Conor, Shani Lanzkron, Mustapha O. Saheed, Smit Saha, Monica E. Peek, Charles Haywood, Jr., and Mary Catherine Beach. Do words matter? stigmatizing language and the transmission of bias in the medical record. *Journal of General Internal Medicine*, 34(1):63–69, 2019. doi: 10.1007/s11606-018-4583-7.

Keith Harrigian, Kayte Spector-Bagdady Wisniewski, David J. Albers, and Mark Dredze. Characterization of stigmatizing language in the electronic health record. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 11761–11772, 2023. URL https://physionet.org/content/stigmatizing-language/1.0.0/.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=H1eA7AEtvS.

Wenkai Jin, Xingyu Zhang, Jie Chen, Yan Xia, Chengyue Zhang, and S Yu Philip Wang. Towards reliable interpretability with Shap in natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5014–5023, 2020.

Bruce G. Link and Jo C. Phelan. Conceptualizing stigma. *Annual Review of Sociology*, 27(1): 363–385, 2001. doi: 10.1146/annurev.soc.27.1.363.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Zhiyuan Lyu, Shiqi Huang, Xinyi Zhu, and Bing Liu. Mental health stigma detection using roberta and bert. *Neurocomputing*, 488:103–112, 2022.

Brenda Major and Laurie T O'Brien. Social stigma and physical health: Understanding the health consequences of discrimination. *Annual review of psychology*, 56:1–26, 2005.

Rana Mansour, Casey Baldwin, Alexander White, Beth Sisk, Aiko Miyao, Mark Alexander, Michael Gensheimer, and Daniel L. Rubin. Honeybee: A scalable modular framework for creating multimodal oncology datasets. *arXiv preprint arXiv:2405.07460*, 2024. URL https://arxiv.org/abs/2405.07460.

Han Meng, Yancan Chen, Yunan Li, Yitian Yang, Jungup Lee, Renwen Zhang, and Yi-Chieh Lee. What is stigma attributed to? a theory-grounded, expert-annotated interview corpus for demystifying mental-health stigma. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5453–5490, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. URL https://aclanthology.org/2025.acl-long.272/.

Mistral AI. Mistral 7b instruct: A dense open-weight language model. https://huggingface.co/mistralai/Mistral-7B-Instruct, 2023a.

Mistral AI. Openhermes: An open instruction-tuned model on mistral 7b. https://huggingface.co/mistralai/OpenHermes-2.5B, 2023b.

Adam J Noble, Catrin Jones, and Catherine White. Stigma in mental health: A critical review and call for research. *Social Psychiatry and Psychiatric Epidemiology*, 56:1–14, 2021.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pp. 2227–2237. Association for Computational Linguistics, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.

Kim Scroggins, Lucine Ghaffari, Muxin Zhang, Rebekah Gonzalez, Michael Chen, Kartik K. Venkatesh, and Anna Rumshisky. Identifying stigmatizing and preferred language in obstetric clinical notes: Development and evaluation of annotation schema and language models. *medRxiv*, 2024. doi: 10.1101/2024.01.10.24301542. URL https://www.medrxiv.org/content/10.1101/2024.01.10.24301542v1.

Ritvik Sharma, Ankit Kumar, Nisheeth Batra, Amba Joshi, and Vikas Varma. Mental health stigma detection in social media with contextualized representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1123–1133. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.eacl-main.98.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 3319–3328, 2017.

Diyi Yang, Ayah Zirikly, Wei-Hung W. Zhang, Oladapo Farri, and Anna Rumshisky. Care-sd: Annotating and detecting subtle stigmatizing language in icu clinical notes. *arXiv preprint arXiv:2405.05204*, 2024. URL https://arxiv.org/abs/2405.05204.

Jaewon Yang, Yajie Zhao, Minsung Song, Yanzhi Liu, and Jingwen Zhu. Stigmatizing language in news coverage of the opioid crisis: Variations in framing by substance and by gender. *Journal of Health Communication*, 24(9):753–761, 2019.

Lawrence H. Yang, Arthur Kleinman, Bruce G. Link, Jo C. Phelan, Sing Lee, and Byron Good. Culture and stigma: Adding moral experience to stigma theory. *Social Science & Medicine*, 103: 107–112, 2014. doi: 10.1016/j.socscimed.2013.09.023.

A About the AI authors

This study was conceived and initially designed using GPT-5 and GPT-4o, leveraging their advanced capabilities for ideation and methodology planning. The foundational codebase was sketched out using the Google Colab environment enhanced by the Gemini AI feature, which facilitated interactive and iterative development with the Gemini agent. Helper Python functions such as reporting specifications of the compute environments used were authored by `Claude Sonnet 4`. Subsequent refinements and enhancements to both the code and experimental setup were carried out incrementally through collaborative interactions with GPT-5 and Google Gemini's chat interface, ensuring a seamless integration of cutting-edge AI assistance throughout the research process.

B Candidate models

Model	Params	Arch	Instr.-Tuned	Year of latest release
`distilbert-base-uncased`	66M	Encoder	No	2019 (Sharma et al., 2021)
`roberta-base`	125M	Encoder	No	2019 (Lyu et al., 2022)
`deberta-v3-base`	183M	Encoder	No	2021 (He et al., 2021)
`Flan-T5-base`	250M	Encoder-Decoder	Yes	2022 (Chung et al., 2022)
`OpenHermes-2.5-Mistral`	7B	Decoder	Yes	2023 (Mistral AI, 2023b)
`Mistral-7B-Instruct`	7B	Decoder	Yes	2023 (Mistral AI, 2023a)

Table 5: Comparison of baseline and instruction-tuned models for mental health stigma detection.

Field Name	Description
`snippet_id` (`participant_id`)	Unique identifiers for the participant and for the snippet. Used to trace which participant provided which snippet.
`text`	The transcript of a participant's response (an interview snippet) to a prompt or question, excluding warm-up and vignette setup.
`attribution_label`	The stigma label assigned to the snippet: one of the seven attribution categories (Responsibility; Social Distance; Anger; Helping; Pity; Coercive Segregation; Fear) or "Non-stigmatizing."
`N/A`	Snippets marked "N/A" when they are unsuitable for annotation due to being too brief, irrelevant, incomplete, unintelligible, or otherwise not amenable to meaningful classification.
`interview_question` (`attribution_type`)	Which core interview question was asked (or which attribution prompt) that elicited this snippet; helps link content to theoretical attribution dimension.
`turn_count` (`response_length`)	Number of conversational turns in the snippet between participant and chatbot; also measures of length (words, tokens) of the response — useful for controlling for verbosity effects.
`participant_demographics` (`sociocultural_metadata`)	Demographics of participant (e.g. gender, age, first language, possibly region or country) used for socio-cultural analyses of stigma.

Table 6: Key data fields in the MH-Stigma-Interview Corpus with their descriptions.

C The annotation protocol

The annotation framework was grounded in attribution theory (Corrigan et al., 2000), enabling the use of structured labels related to emotions, blame, and behavioral intentions—such as perceived responsibility, desire for social distance, or feelings of anger. This theoretical grounding ensures the annotations go beyond surface-level or purely lexical cues of stigma.

Annotators were experts in fields such as mental health, psychology, and social sciences. They received training through a detailed codebook that defined each stigma category and provided illustrative examples to ensure consistent interpretation.

Annotation was conducted through a multi-stage, expert-in-the-loop process. Multiple annotators independently labeled each snippet, and inter-annotator agreement was computed. Discrepancies were reviewed collaboratively, with final decisions adjudicated by senior experts, ensuring high reliability and consistency.

Inter-annotator agreement metrics, such as Cohen's kappa and Fleiss's kappa, are reported in detail elsewhere (Meng et al., 2025). To further enhance label reliability, rounds of disagreement resolution were integrated into the process.

D Listing of computer resources

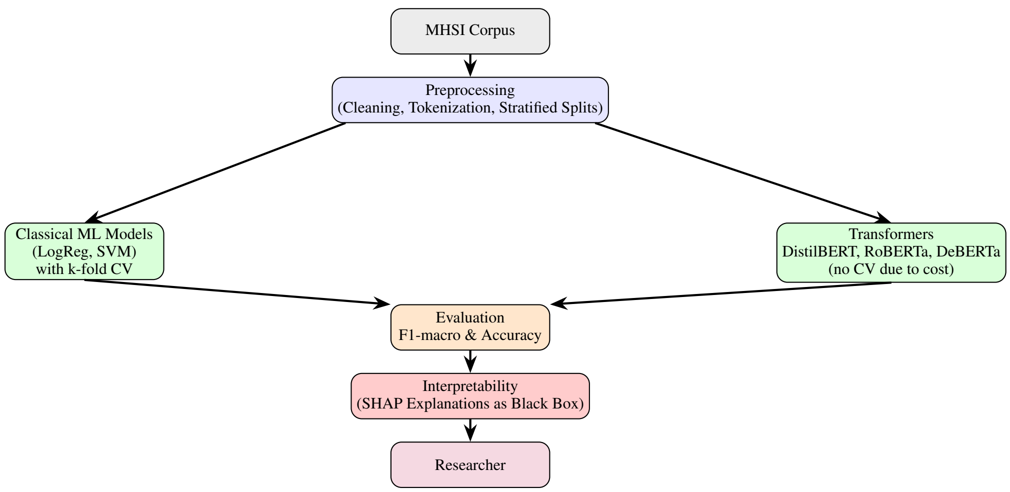
```
SYSTEM SPECIFICATIONS REPORT generated using code written by Claude AI

SYSTEM INFORMATION
----------------------------------------
Platform: Linux 6.6.56+
Architecture: x86_64
Processor: x86_64
Hostname: 8d50b7df9a6a
```

```
Python: 3.11.13 (CPython)
Generated: 2025-09-16 22:51:42

CPU INFORMATION
----------------------------------------
Model: Intel(R) Xeon(R) CPU @ 2.00GHz
Physical Cores: 2
Logical Cores: 4
Max Frequency: 0.00 MHz
Current Usage: 0.5%

MEMORY INFORMATION
----------------------------------------
Total RAM: 31.35 GB
Available RAM: 27.33 GB
Used RAM: 3.55 GB (12.8%)
Swap: 0.00 GB / 0.00 GB

GPU INFORMATION
----------------------------------------
No NVIDIA GPUs detected
CUDA Available: True
CUDA Version: 12.4
cuDNN Version: 90100
PyTorch Available: True
PyTorch Version: 2.6.0+cu124
TensorFlow Available: True
TensorFlow Version: 2.18.0
TF GPU Available: True

STORAGE INFORMATION
----------------------------------------
Drive /dev/loop1: 2.76 GB / 19.52 GB (14.1%)
Drive /dev/loop1: 2.76 GB / 19.52 GB (14.1%)
Drive /dev/loop1: 2.76 GB / 19.52 GB (14.1%)

PYTHON ENVIRONMENT
----------------------------------------
Executable: /usr/bin/python3
Conda Environment: Not available
Virtual Environment: Not set

KEY INSTALLED PACKAGES
----------------------------------------
ML/DL Frameworks:
  torch: 2.6.0+cu124
  tensorflow: 2.18.0
  transformers: 4.52.4
Data Science:
  numpy: 1.26.4
  pandas: 2.2.3
  matplotlib: 3.7.2
  seaborn: 0.12.2
Other packages:
  torchvision: 0.21.0+cu124
  keras: 3.8.0
  datasets: 3.6.0
  tokenizers: 0.21.2
  jupyter: Unknown
```

```
notebook: 6.5.4
scipy: 1.15.3
statsmodels: 0.14.4
```

E Observed errors made by GPT-5

The diagram generated by GPT-5 is not completely correct. For instanc,e the text “stratified splits” should be placed along with “k-fold CV”. Tokenization was not the precursor in “classical” ML mdels. The term “classical” was never mentioned in the manuscript but was adopted when the manuscript mentions “traditional ML” throughout.



F AI Research Autonomy / AI Contribution Disclosure

AI systems’ role: An AI system (transformer language model + training script) performed the bulk of model development, hyperparameter sweeps, metric computation, and figure generation. Humans provided dataset curation, labeling guidelines, final labeling oversight, experimental design decisions, and final manuscript editing. The AI is not listed as a human author; humans are the corresponding authors but we document AI contributions in the checklist as required. Agents for Science

F.1 Responsible AI Statement (concise)

We followed NeurIPS/standard ethical guidelines and considered risks from automated stigma detection. Key actions: (1) human oversight for labeling and final decisions, (2) transparency via per-instance SHAP explanations to support human review, (3) dataset de-identification and adherence to platform terms of service, and (4) a discussion of potential harms (false positives leading to censorship; false negatives perpetuating harm) and mitigation strategies, including human-in-the-loop workflows and threshold tuning to prioritize recall/precision depending on downstream use.

F.2 Reproducibility Statement

We provide code for training, evaluation, checkpoints, and the SHAP explanation notebook. Exact package versions, random seeds, and hardware (GPU type) are listed on the accompanying GitHub. We include a script to reproduce reported metrics given the provided checkpoint and test set.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

Answer:

Explanation: The high-level research direction—using NLP to identify stigmatizing language in clinical notes—was determined by human researchers. However, iterative refinement of sub-questions, including prompt-based exploration and candidate examples of biased language, was heavily supported by AI systems (Gemini and GPT-5), which generated example outputs and reframed problem statements.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer:

   Explanation: Human researchers designed the experimental protocol and instructed Gemini to implement the pipeline using Python and Hugging Face Transformers.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer:

   Explanation: Humans led the analysis of model performance, error cases, and cross-dataset generalization. However, AI systems were consulted for natural language interpretations of classifier decisions and to propose phrasing for summary descriptions of findings.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer:

   Explanation: AI models (primarily GPT-4 and Gemini) produced the 95% of first-draft text for methods, dataset description, and table formatting. Human authors edited and curated outputs, handled technical details, and ensured consistency in tone and citations.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: While large language models like GPT-4 and Gemini were effective in generating fluent text and synthesizing prior findings, they often hallucinated citations, required careful oversight on factual accuracy, and struggled with domain-specific nuance (e.g., distinguishing clinically appropriate from subtly biased phrasing). Prompt sensitivity and inconsistencies across sessions also limited replicability. Human

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The introduction and abstract clearly state the goals (bias-aware NLP for clinical communication), methods (benchmarking and agentic feedback pipeline), and societal motivation (health equity), aligning well with the actual contributions.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The manuscript acknowledges the sociolinguistic challenges in defining stigma, cross-cultural variance, and limitations in generalizing across specialties or domains.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include formal theoretical results or proofs; it is primarily empirical and system-driven.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental procedures, model types, datasets, and evaluation metrics are described in sufficient detail for replication. Key dataset access instructions are also included.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: At least one dataset (MHSI) is available via author request, and code is released on the anonymous website.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The manuscript includes description of model architectures, dataset splits, and training parameters. Additional hyperparameter details are included in the appendix.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper includes accuracy and F1-macro scores with standard deviations across multiple folds in a k-fold CV, allowing readers to assess statistical robustness. Further, when evaluated on the test set, bootstrap resampling was used to estimate the confidence intervals of the accuracy and F1-macro scores.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the specifications in the Appendix that summarizes the compute environment of a reproducible notebook originally executed on Kaggle on 2025-09-15 to 2025-09-16. Section D

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [Yes]

   Justification: The work promotes equitable healthcare, minimizes harm, avoids sensitive data misuse, and transparently reports limitations and design decisions.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: The manuscript discusses both positive impacts (supporting equitable care, improving clinical communication) and risks (reinforcing subtle biases, misuse of AI-generated feedback) in a dedicated section.