From Stigma to Support: A Parallel Monolingual Corpus and NLP Approach for Neutralizing Mental Illness Bias

Mason Choey
The Nueva School
San Mateo, California, USA
mason@choey.com

Abstract

Negative attitudes and perceptions towards mental illness continue to be pervasive in our society. One of the factors contributing to and reinforcing this stigma is the usage of language that is biased against mental illness. Identifying biased language and replacing it with person-first, neutralized language is a first step towards eliminating harmful stereotypes and creating a supportive and inclusive environment for those living with mental illness. This paper presents a novel Natural Language Processing (NLP) system that aims to automatically identify biased text related to mental illness and suggest neutral language replacements without altering the original text's meaning. Building on previous work in the field, this paper presents the Mental Illness Neutrality Corpus (MINC) comprising over 5500 mental illness-biased text and neutralized sentence pairs (in English), which is used to fine-tune a CONCURRENT model system developed by Pryzant et al. (2020). After evaluation, the model demonstrates high proficiency in neutralizing mental illness bias with an accuracy of 98.7%. This work contributes a valuable resource for reducing mental illness bias in text and has the potential for further research in tackling more complex nuances and multilingual biases.

1 Introduction

Globally, 970 million people are currently living with mental illness. Each year, 14.3% of deaths, or approximately 8 million people, are caused by mental disorders. Depression affects over 300 million people worldwide, affecting people of all demographic and socioeconomic backgrounds. Anxiety disorders are almost as common, affecting 284 million people worldwide (Children's Hopechest, 2022). During the first year of the COVID-19 pandemic, incidences of depression and anxiety increased by an astounding 25% (WHO, 2022). In addition to depression and anxiety disorders, suicidal ideation, bipolar disorder, autism spectrum disorder (ASD), substance use disorder, and eating disorders such as bulimia and anorexia nervosa fall under the umbrella of mental illness (Zhang et al., 2022).

Mental illnesses have traditionally been one of the most stigmatized health conditions. Stigma and derogatory language about mental illness can be found everywhere, from casual conversations to media, even among medical professionals. Language can be used to reinforce stigma or fight it. Guidelines for responsible reporting by the media emphasize how language used in media can either encourage help-seeking behavior or inadvertently lead to suicide contagion (Reporting on Suicide, 2023). As an example, the phrase "committed suicide" is commonly used but has been replaced by the preferred "died by suicide" to avoid an association with "committing (a murder)" and demonstrate compassion through word choice (Mental Health Coalition, 2021).

In casual conversation, mental health diagnostic terms are frequently used to describe a non-medical event, such as the weather as "bipolar" or a situation as "insane". Although seemingly harmless, misusing diagnostic terms can lead to misunderstandings about mental illness by undermining the severity of mental illness and reinforcing negative stereotypes (Volkow, 2021).

Biased language affects how those with mental illnesses view themselves, how others treat them, and whether they seek help (Rose et al., 2007). People with mental illnesses, particularly

Proceedings of Recent Advances in Natural Language Processing, pages 249–254
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_028

substance use disorder, who perceive a high degree of public stigma about their condition, were half as likely to seek help as those who perceive a low degree of stigma (Canadian Centre for Substance Abuse and Addiction, 2019). Medical professionals and mental health service providers with bias against mental illness are less likely to offer appropriate treatment or refer those with mental illness to the specialty care they need (Volkow et al., 2021).

Natural Language Processing presents an opportunity to apply a language model to identify biased language and neutralize it by suggesting more respectful and compassionate language to replace it.

This project aims to develop an NLP system to identify text (in English) biased against mental illness and automatically replace it with proposed edits of neutral language without changing its original meaning.

2 Related Work

2.1 NLP Studies in Mental Illness

To date, NLP studies in mental illness have focused on (1) sentiment analysis (Nadkarni et al., 2011), (2) symptom detection (Jackson et al., 2017), (3) mental health surveillance (Mukherjee et al., 2020) (4) mental health portrayal in print media (Chen et al., 2017), and (5) text classification (Ive et al., 2020).

Most studies applying NLP to mental illness have focused on early indicators to support detection, prevention, and treatment (Zhang et al., 2022).

Existing studies have also focused on specific mental illnesses, such as PTSD or post-traumatic stress disorder (Sawalha et al., 2022), suicide, depression, or data sources such as social media and non-clinical texts (Zhang et al., 2022).

No studies have been conducted on neutralizing biased language related to mental illness.

2.2 Automatically Neutralizing Subjective Bias in Text

Pryzant et al. (2020) pioneered the development of the first generative model designed to mitigate biased text. They also introduced three valuable tools and frameworks into the discourse: the Wiki Neutrality Corpus (WNC), a corpus of 180,000 sentence pairs of subjective and neutralized text from Wikipedia, and two generative models that were trained on the WNC to (1) identify subjective bias in text, and (2) propose edits to neutralize it.

Notably, the groundbreaking use of a joint embedding architecture to integrate bias identification and text generation sets their work apart. Their paper is considered the first to successfully combine both tasks and utilize the identification algorithm to directly fine-tune a generative algorithm. Furthermore, the construction methodology of the Wiki Neutrality Corpus serves as a valuable framework for constructing other types of bias-related corpora. It is worth mentioning that their work focuses exclusively on subjective bias, but their methodology provides a promising foundation for exploring the mitigation of other forms of bias.

This project builds upon the model proposed by Pryzant et al. (2020) and extends its application to specifically address mental illness bias by creating a parallel corpus, fine-tuning Pryzant et al.'s (2020) model, and then evaluating the results.

3 Mental Illness Neutrality Corpus (MINC)

This paper introduces the Mental Illness Neutrality Corpus (MINC)[1], a novel parallel monolingual (specifically, English) corpus of mental illness-biased text. This dataset is comprised of 5500+ mental illness-biased text, neutralized sentence-pairs, and metadata. To construct the MINC, several language guides[2] were referenced to compile a list of biased expressions and suggested text replacements of appropriate and respectful word choices. In addition to general terms describing mental illness, the corpus contains biased text describing substance use and eating disorders, which fall under the umbrella of mental illness as defined by several of the language guides mentioned above.

To create sentences with biased text, ChatGPT (OpenAI, 2023) was prompted to pull real-world

[1] https://github.com/masonchoey/from-stigma-to-support

[2] National Recreation and Park Association's Mental Health Substance Use Disorder Language Guide, Well Beings' Mental Health Language Guide, The Mental Health Coalition's Language Guide, Hogg Foundation of Mental Health's Language Matters in Mental Health, DBSA's 10 Ways to Combat Discrimination with Compassionate Language, Canadian Centre on Substance Use and Addiction's Overcoming Stigma Through Language: A Primer, and "280 Labels Used to Stigmatize People with Mental Illness" (Rose et al. 2007).

examples of biased text (i.e., source sentence), which was paired with suggested neutralized replacements (i.e., target sentence). ChatGPT is an AI language model trained on data obtained from books, web texts, Wikipedia, news articles, scientific journals, and social media; in total, 570 GB of data and pieces of writing were collected from the internet. In a few instances, ChatGPT's content restrictions would prohibit prejudicial language from being included in our prompts. In these cases, we performed the reverse task of prompting ChatGPT to generate the target sentences, which were then paired with corresponding biased text.

After referencing literature on a list of commonly used biased text about mental illness, four categories of mental illness language bias were identified:

1. **Derogatory depiction of mental illness**: words intended to degrade those living with mental illness.
2. **Outdated language for mental illness**: words without harmful intentions but have been replaced with more respectful and compassionate language.
3. **Person-first language**: words that focus on a person's abilities instead of their limitations; putting the person first before the mental illness.
4. **Using mental illness as a metaphor**: words to describe something other than a person experiencing the disorder, using a metaphor to describe something unrelated to mental illness.

Each biased sentence is annotated in MINC into one of these four categories (refer to Table 1 for examples). Each biased sentence was paired with suggested edits (i.e., target sentence) of neutralized text to form sentence-pairs.

Source	Target	Category
The **crackhead** was unable to hold down a job due to their addiction.	The **person with cocaine use disorder** was unable to hold down a job due to their addiction.	Derogatory depiction of mental illness.
Struggling with depression for many years ultimately led him to **kill himself.**	Struggling with depression for many years ultimately led him to **die by suicide**.	Outdated language for mental illness.
Hospitalized for malnutrition, the **anorexic's** weight had dropped too low.	Hospitalized for malnutrition, the **person living with anorexia nervosa's** weight had dropped too low.	Person-first language.
The weather was **bipolar** today, with sunshine and rain alternating throughout the day.	The weather was **oscillating** today, with sunshine and rain alternating throughout the day.	Using mental illness as a metaphor.

Table 1: Samples from MINC. Biased text and neutralized text are in bold. Each sentence-pair is annotated with category.

Subcategory	% of corpus
Derogatory depiction of mental illness	33.3
Outdated language for mental illness	21.33
Person-first language	16.00
Using mental illness as a metaphor	29.33

Table 2: Percentage of mental illness biased text by category in MINC.

4 Approach

This project employs the CONCURRENT system proposed by Pryzant et al. (2020) and fine-tunes it using MINC. This CONCURRENT model architecture consists of two different modules, a detection module, and an editing module. The detection module aims to identify which word in the sequence is likely to be biased. It is a neural-sequence tagger that estimates p_i, or the chance that a word is biased using the equation:

$$p_i = \sigma(b_i, W_b + e_i, W_e + b) \quad (1)$$

where b_i represents the semantic meaning of the contextualized word vector as produced by BERT. W_b, W_e, and b are learnable parameters. The editing module takes a subjectively biased sentence s and edits it to a more neutral replacement sentence t. First, a bi-LSTM encoder takes the problematic sentence and converts it to a sequence of hidden states H_1 , H_2 , H_3 … then the LSTM decoder generates text one token at a time, according to which tokens are more likely to be biased.

First, when taking an input, the detection module labels the sentences according to which words are more likely to be biased. Once the potentially biased sentence has been identified with the words that are most likely to be biased, the detection and editing modules are connected using a *join embedding* mechanism, which, using the probabilities of each word being biased from the detection module $p = (p_1, \ldots, p_n)$, is added to the hidden state in the editing module using the following equation:

$$h_i' = h_i + p_i + v \quad (2)$$

where v is the *join embedding* vector that is multiplied by the probabilities, then added to each hidden state. In doing so, the words that are more likely to be biased are weighted higher in the encoder-decoder architecture. Finally, a token-weighted loss function is used to evaluate the model.

5 Training

Using the new MINC, the CONCURRENT system (Pryzant et al., 2020) is fine-tuned. The MINC data was split 85% training data and 15% testing data, with roughly 4120 sentences used for training data and roughly 730 sentences used for testing data. The fine-tuning process was implemented using PyTorch and the Adam optimizer with a learning rate 5e-5. Batch size of 16 and all vectors of length h=512. Gradient clipping with a maximum gradient norm of 3 was used and a dropout probability of 0.2 for the inputs of each LSTM cell. The BERT model was initialized using the bert-based-uncased pre-trained parameters (Devlin et al., 2019). The other parameters were randomly initialized on the range [-0.1, 0.1]. After pre-training using the neutral text, the CONCURRENT model was fine-tuned using the training data in addition to 710 sentences of neutral data for 20 epochs. The training time was approximately 3 hours, using the Apple M1 chip.

6 Results and Analysis

6.1 Evaluation

After employing human evaluation by three validators, the results were divided into three categories: Perfect (P), Good (G), and Incorrect (X). A "P" rating denotes results in which the model corrected the biased sentence to the sentence proposed in the corpus. This includes neutralized sentence data points in which (1) the model did not replace the answer and (2) sentences that the model corrected by removing biased language and inserting the language in the target sentence. A "G" rating is given when the model correctly identifies and neutralizes harmful language but inserts a synonym or slightly different word(s) instead of the suggested replacement word(s) in the target sentence. However, these instances were included in the accuracy score since the source sentence's bias was correctly identified and neutralized. Finally, an "X" rating is given when the model either (1) does not correct the biased language, (2) tries to correct a neutral example, or (3) results in a grammatically incorrect sentence. Human intervention for evaluation and annotation was necessary to detect grammatical errors.

Category	% of total	Counted towards accuracy
P	40.0	Yes
G	58.7	Yes
X	1.3	No
Accuracy	98.7	

Table 3: Summary of human annotations of results.

6.2 Analysis

The results indicate that the model performed very well at neutralizing biased language and performed exceptionally well at identifying biased language. "G" ratings came up frequently since the dataset included multiple suggested replacements to neutralize biased text (e.g., living with, experiencing, etc.) As such, the model cannot accurately predict which will occur in the target sequence, and they are subsequently given a "G" rating. Combining "P" and "G" ratings provides a more accurate view of how successfully the model neutralizes text.

7 Conclusion

Bias against mental illness is pervasive in our culture, frequently appearing as biased language in the media or casual conversation. Although bias can be nuanced, implied, or unconscious, language biased against mental illness has a negative impact on those living with mental illness. Identifying and reducing bias is crucial to reducing prejudice and helping those with mental illness seek and obtain the needed treatment.

The proposed models in this study were highly proficient in providing appropriate neutralized suggestions for reducing subjective bias for the biased sentences generated by ChatGPT.

This paper presents the annotated corpus of mental illness biased text (MINC). The MINC is a novel monolingual parallel corpus generated by ChatGPT from real-world text and trained on data from a wide variety of sources such as news media, social media, Wikipedia, books, personal websites, etc. Human intervention was necessary for annotation and review of grammatical errors. Several language guides for journalists and writers were consulted to obtain a list of commonly used biased terms and phrases and replacements that were respectful and compassionate towards those with mental illness.

This paper is a first step towards reducing bias in language describing mental illness, but further study should tackle more complex text such as multi-word, multilingual, and cross-sentence bias, as well as nuances and implicit language, taking into consideration that language, slang in particular, is ever-evolving. Also worth noting is the MINC is entirely in English. Additional work to study language bias and applying our model to non-English languages would be a logical next step.

Language is a complex, ever-evolving field of study. While NLP is increasingly sophisticated, it has yet to replace human language cognition completely. However, using NLP models to reduce bias in real-world text is a significant step toward addressing and lessening mental illness bias in our society. Given the substantial negative impact of biased language used to describe those with mental illness, creating more sophisticated detection models should be a high priority.

References

Thushari Atapattu, Mahen Herath, et al. 2022. EmoMent: an Emotion Annotated Mental Health Corpus from Two South Asian Countries. In *Proceedings of the 29th International Conference on Computational Linguistics.* Pages 6991–7001, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. https://doi.org/10.48550/arXiv.2208.08486.

Marian Chen, Stephen Lawrie. 2017. Newspaper depictions of mental and physical health. *BJPsychBull*. 2017 Dec;41(6):308-313. https://doi.org/10.1192/pb.bp.116.054775.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ACL Anthology.* 2019. https://doi.org/10.48550/arXiv.1810.04805.

Swapna Gottipati, Mark Chong, et al. 2021. Exploring media portrayals of people with mental disorders using NLP. In *Proceedings of the 14th International Conference on Health Informatics HEALTHINF 2021: Part of BIOSTEC 2021*, Virtual, February 11-13. 5, 708-715. Research Collection School of Computing and Information Systems. https://doi.org/10.5220/0010380007080715.

Julia Ive, et al. 2020. Generation and evaluation of artificial mental health records for Natural Language Processing. *npj Digit. Med.* 3, 69. https://doi.org/10.1038/s41746-020-0267-x.

Richard G. Jackson, Rashmi Patel, et al. 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open*. 7(1):e012012. https://doi.org/10.1038/s41746-020-0267-x.

Min Hyung Lee and Richard Kyung. 2022. Mental Health Stigma and Natural Language Processing: Two Enigmas Through the Lens of a Limited Corpus, *2022 IEEE World AI IoT Congress (AIIoT)*, Seattle, WA, USA, 2022, pages 688-691, https://doi.org/10.1109/AIIoT54504.2022.9817362.

Mukherjee, Sankha S. et al. 2020. Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry* 4, 76–106. https://doi.org/10.1162/cpsy_a_00030.

Nadkarni Prakash M., Lucia Ohno-Machado, Wendy W. Chapman. 2011. Natural language processing: an introduction. *J Am Med Inform Assoc*., 18(5):544-51. https://doi.org/10.1136/amiajnl-2011-000464.

Reid Pryzant, Richard Diehl Martinez, R. et al. 2020. Automatically Neutralizing Subjective Bias in Text. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(01), 480-489. https://doi.org/10.1609/aaai.v34i01.5385.

Diana Rose, Graham Thornicroft, et al. 2007. 250 labels used to stigmatise people with mental illness. *BMC Health Services Research* 7, 97 https://doi.org/10.1186/1472-6963-7-97.

Jeff Sawalha, Muhammed Yousefnezhad, et al. 2022. Detecting Presence of PTSD Using Sentiment Analysis From Text Data. *Front Psychiatry*. 12:811392. https://doi.org/10.3389/fpsyt.2021.811392.

Nora Volkow, Joshua Gordon et al. 2021. Choosing appropriate language to reduce the stigma around mental illness and substance use disorders. *American College of Neuropsychopharmacology*, Dec;46(13):2230-2232. https://doi.org/10.1038/s41386-021-01069-4.

Tianlin Zhang, Annika Schoene, et al. 2022. Natural language processing applied to mental illness detection: a narrative review. *npj Digit. Med*. 5, 46 https://doi.org/10.1038/s41746-022-00589-7.

References: *Website Links*

Born This Way Foundation. 2023. Mental Health First Aid: Teen Mental Health First Aid. https://www.mentalhealthfirstaid.org/wp-content/uploads/2023/02/MHFA_Teen_Flyer.pdf. Accessed: 2023-05-15.

Canadian Centre on Substance Use and Addiction. 2019. Overcoming Stigma Through Language: A Primer. https://https://www.ccsa.ca/overcoming-stigma-through-language-primer. Accessed: 2023-05-15.

Children's Hopechest. 2022. Global Mental Health Statistics https://www.hopechest.org/global-mental-health-statistics. Accessed: 2023-05-15.

Depression and Bipolar Support Alliance. 2023. 10 Ways to Combat Discrimination with Compassionate Language. https://www.dbsalliance.org/wp-content/uploads/2019/02/DBSA_language.pdf. Accessed: 2023-05-15.

Hogg Foundation for Mental Health, University of Texas-Austin. Language Matters in Mental Health. https://hogg.utexas.edu/news-resources/language-matters-in-mental-health. Accessed: 2023-05-15.

OpenAI. 2023. *ChatGPT* (January 9 Version). https://chat.openai.com.

The Mental Health Coalition. 2021. Language Guide. https://www.thementalhealthcoalition.org/wp-content/uploads/2020/05/The-Mental-Health-Coalitions-Language-Guide.pdf. Accessed: 2023-05-15.

National Recreation and Park Association. 2021. Mental Health and Substance Use Disorder Language Guide. https://www.nrpa.org/globalassets/research/mental-health-and-substance-use-disorder-language-guide-december-2021.pdf. Accessed: 2023-05-15.

Reporting on Suicide. 2023. Best Practices and Recommendations for Reporting on Suicide. https://reportingonsuicide.org/recommendations/. Accessed 2023-05-15.

WETA. 2021. Well Beings Mental Health Language Guide. https://wellbeings.org/wp-content/uploads/2021/11/Well-Beings_Language-Guide_FINAL_111821.pdf. Accessed: 2023-05-15.

World Health Organization. 2022. COVID-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide. http://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-worldwide. Accessed: 2023-05-15.