From the holdings of the Department of Special Collections, Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes ("Fair use") as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyright law. We do not charge any permission or usage fees for the publication of images of material in our collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.

For more information regarding copyright, credit, and citation please review the information on the Copyright, Credit and Citation Guidelines page of our website. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form.

U.S. Copyright law:
The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law.

What Seems to be the Problem? Stigmatizing Language in Patient Medical Notes

Abinitha Gourabathina

Advisor: Professor Christiane Fellbaum

Submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science in Engineering

Department of Operations Research and Financial Engineering

Princeton University

April 2023

I hereby declare that I am the sole author of this thesis.

I authorize Princeton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Abinitha Gourabathina

I further authorize Princeton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Abinitha Gourabathina

Abstract

Stigmatizing language in medical notes can prevent a patient from acquiring proper treatment. Reading medical notes containing biased language can influence subsequent clinicians' perception of a patient, further compounding a patient's inability to receive adequate care. Thus, there is a clear need to correct patient notes to eliminate stigmatizing language. Prior work involving stigmatizing language in medical notes has largely remained qualitative where clinicians and researchers manually analyzed notes for stigmatizing keywords. Our work utilized a computational approach to obtain a more robust set of stigmatizing keywords. We created contextual word embeddings from BERT-based and BioBERT-based models that are trained on free-text patient-oriented clinical data. These state-of-the-art models allowed us to develop word vector representations, from which we identified 30 new stigmatizing keywords. We then complete a thorough analysis to build a grammar structure that categorizes stigmatizing keywords according to the ways they induce stigma and better understand the syntactical environments in which these keywords occur. Following our analysis, we developed a model called MedStiLE (Medical note Stigmatizing Language Editor) that utilizes the grammar structure and constituency parsing to edit notes containing the stigmatizing keywords to be non-stigmatizing. We conducted an evaluation to test the efficacy of MedStiLE using human raters and found that it significantly reduced stigma in notes. This research provides various novel insights in terms of methodology and results that can help shape future works involving the intersection of language and healthcare.

Acknowledgements

First, I would like thank my advisor, Professor Christiane Fellbaum. This project would not have been possible at all without her unparalleled guidance, support, and encouragement. For the past six years, her wisdom, kindness, and dedication have beyond inspired me. She has been a guiding light throughout my Princeton experience and the most incredible mentor. There truly are no words that would fully express my gratitude, despite the linguistics classes I have taken. Thank you for everything. I would also like to thank Professor Srinivas Bangalore who has been an immense support in creating this work. Thank you for teaching me so much about language and computation, and thank you for your consistent encouragement. I have learned invaluable lessons from your course, both as a student and course assistant.

Thank you to my friends who have brought so much joy. Gaya, thank you for the late-night walks and always understanding me. Annie, thank you for the sugary treats and uplifting words. Selinay, thank you for always making me laugh and giving the best hugs. Ivania, thank you for the fun stories and always making me feel at home. Thank you to all of the amazing people inside and outside of Princeton who have helped in the creation and completion of this work.

Finally, thank you to my parents for their unwavering love and support. Every step of the way, they believed in me and provided me with everything I could ever ask for. You are my motivation and my comfort.

To my parents

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

In the medical field, the adage "with great power comes great responsibility" holds true, as clinicians' judgments and decisions completely affect patient outcomes. Research shows that clinicians can be dismissive towards patients and cast doubt on patients' experiences and opinions [3, 4]. The consequences of the clinician-patient power dynamic can be exacerbated by additional factors. Medical professionals are not exempt from harboring potential biases against certain demographic groups. Numerous studies have shown how patients may receive poorer quality of care based on their gender, race/ethnicity, level of health literacy, and many other factors [5, 6, 7, 8].

As an individual's language can reflect their internal attitudes and implicit biases, medical notes can reveal how a clinician feels toward a patient. The study of stigmatizing language in medical notes is incredibly important in creating a more egalitarian healthcare system. Moreover, medical notes may cause more harm than just being a static receipt of bias. Stigmatizing language written in a patient's medical record can perpetuate negative attitudes and influence decision-making of clinicians subsequently caring for that patient, thereby hindering a patient's ability to obtain insurance, re-

ceive the medical care they need, and recover effectively. Thus, stigmatizing language in medical records must be corrected.

Most research regarding stigmatizing language in medical records has remained qualitative, with physicians and researchers alike closely studying medical notes to identify linguistic features most associated with stigma and categorizing types of stigmatizing language for easier study. To understand how and when stigmatizing language is used, rigorous quantitative research is required. Gourabathina used data-driven approaches to study language that casts doubt on patients [9]. The work analyzed how specific stigmatizing linguistic features that cast doubt on patients correspond to particular patient demographics, medical note types, and clinician categories. The statistically significant results showed that stigmatized language is present with greater likelihood in records of female patients, Black and Latinx patients, patients with federal health insurance plans, and in records for emergent rather than elective medical procedures. From these results, we see concrete evidence of how stigmatizing language is used more often with specific groups of patients, motivating us to develop a model that would prevent unequal treatment of patients. Gourabathina also looked at textual co-occurrences to identify other keywords related to bias beyond the linguistic features documented by previous literature [9]. Building upon these findings and with several methodological advancements, we have created a model named MedStiLE that edits notes to eliminate stigmatizing language.

1.2 Goal

Gourabathina's work has provided a preliminary computational model to identify additional stigmatizing keywords [9]. From there, to develop a more robust set of stigmatizing keywords, we created patient-oriented clinical word embeddings by training specialized BERT transformers on patient medical notes. Then, we created a gram-

mar structure to identify ways to edit or correct for stigmatizing language. By using the Stanford CoreNLP Parser to create constituency parses of the medical notes, we were then able to develop our MedStiLE (Medical note Stigmatizing Language Editor) model that automatically edits notes to be non-stigmatizing. The efficacy of MedStiLE in eliminating stigma from notes was evaluated. Overall, our work has contributed the following:

- We created contextual word embeddings from BERT models that are useful vector representations of patient-oriented clinical free-text data. Current models involving contextual word embeddings are not domain-specific for patient medical notes and hence not useful when encountering medical jargon and plain language in tandem. These word embeddings have specifically been trained on clinical notes that describe or mention patient behavior or experiences. While we have used these patient-oriented clinical word embeddings for the task of identifying stigmatizing keywords, they can be used for many other NLP tasks.

- We have identified 30 additional keywords of stigmatizing language in patient medical notes. Through meticulous study of 70 total stigmatizing keywords, we have categorized the keywords into four classes, providing additional insights in the ways that stigmatizing language is used in medical notes.

- We have done a thorough analysis of the syntactic environments in which 70 stigmatizing keywords occur and develop a grammar that restructures stigmatizing statements to become non-stigmatizing.

- We have effectively created a model that edits medical notes to remove stigmatizing language. To the best of our knowledge, our model is the first of its kind.

Our work serves as an example of how computational methods can be used to foster more equitable healthcare. We believe that our model shows how stigmatizing language in medical notes can be effectively addressed by the medical community, thereby opening avenues towards breaking down social barriers in healthcare.

Chapter 2

Background and Related Work

Medicine has long been a field that has emphasized professionalism and formality [10]. In doing so, the notion of medical authority and the power of the white coat have become deeply rooted in societal perceptions of physicians, patients, and the relative statuses of both parties, with physicians clearly being superior [11]. This power difference becomes even more pronounced when considering race, gender, ethnicity, sexuality, and other factors, especially given the history of the medical field. From the very beginnings of its conception, medicine has largely been developed by and for a very specific demographic: the wealthy, white, able-bodied, heterosexual, cisgender male [12, 13]. Individuals who do not fit these criteria are not only under-represented in health research and as medical professionals, but also tend to be ignored, dismissed, and devalued as patients [12, 13]. Such disparities in healthcare are not the consequence of individual clinicians but rather a larger systemic issue that is deeply ingrained in the medical field. This section briefly discusses the systemic disadvantages faced by certain patient demographics, the significance of stigmatizing language in medical notes, and the contributions of several relevant works.

2.1 Women and Hysteria

In recent years, more attention is being brought to medical gaslighting: "the interpersonal phenomenon of having one's experience of illness marginalized (including having one's self-reported or presenting symptoms downplayed, silenced, or psychologically manipulated) by a clinical provider or healthcare professional" [14, 15]. Women have been opening up about their individual experiences with medical gaslighting via social media, recounting how clinicians have dismissed their symptoms and complaints of pain, stemming from an implicit perception of women as overdramatic and needy [14, 16]. In fact, the misogynistic notion of women as dramatic and "hysterical" traces back to the very beginning of Western medicine.

In the 5th century BC, Hippocrates, a Greek Physician who is now considered one of the most influential figures in the history of medicine, coined the term "hysteria" from the Greek word for uterus, "hysteron" [15]. He classified "epilepsy" and "hysteria" as separate gendered diagnoses, with epilepsy as a condition pertaining to the brain and therefore corresponding to men while hysteria was deemed a condition pertaining to the uterus. He believed hysteria was an inherently female disease, caused by sexual frustration or lack of procreation. He described hysteria as any combination of anxiety, sense of suffocation, tremors, and even convulsions and paralysis. This notion of hysteria among women was widely accepted by medical scholars of the time and for centuries going forward; hysteria became a common medical diagnosis for women, applied whenever women displayed "inappropriate" emotions such as anxiety, anger or even sexual desire. In the 18th century, hysteria finally started becoming associated with the brain, as a mental disorder, rather than the uterus [17], but the concept of the hysterical woman has stood the test of time.

Women's health issues are likely to be misdiagnosed or dismissed by doctors as something less critical. A study published by the Academic Emergency Medicine Journal found that women with severe stomach pain waited for almost 33% longer

in the emergency room than men with the same symptoms [18]. Other studies have shown that compared with men, it takes longer for women to be properly diagnosed for cancer and heart disease [19]. In a study from the Journal of Women's Health, women were twice as likely as men to be diagnosed with a mental illness when their symptoms were, in actuality, consistent with heart disease [20]. Mikolic et al. has also shown that women are treated less aggressively for traumatic brain injury where they are offered pain medications at lower dosages [21].

2.2 Race in Medicine

Extensive research has been conducted about disparities in healthcare between white patients and patients of color. Beliefs that Black people and white people are fundamentally biologically different have been prevalent for centuries.

In the United States, Black men and women were subject to slavery and inhumane treatment during medical testing on the basis that they were a "different breed of human" [22]. During the 19th century, prominent physicians, such as Dr. Samuel Cartwright sought to identify the "physical peculiarities" of Black people compared to whites, as if to classify two subspecies of human [22]. Such "peculiarities" included thicker skulls, less sensitive nervous systems, and diseases inherent to dark skin [23]. Today, many continue to believe that the Black body is biologically and fundamentally different from the white body. Even medical students hold disturbingly ignorant beliefs regarding the physiological differences between black people and white people. In 2016, Hoffman et al. found that 40% of first- and second-year medical students from a large public university endorsed the belief that "Black people's skin is thicker than white people's" [24].

These perceived differences between white and Black bodies fosters an impression of Black people as superhuman and tough, thereby suggesting that Black patients do

not need the same medical attention or help that their white counterparts do. In a 2012 study, Sabin et al. found that pediatricians would be more likely to prescribe pain medication for white teenagers than Black teenagers [25]. A meta-analysis of 20 years of studies regarding pain management disparities found that Black/African American patients were 22% less likely than white patients to receive any pain medication [26]. Like Black patients, other ethnic groups face inequities in medicine, but the underlying stigmas have not yet been properly identified for other groups [27]. The aforementioned meta-analysis found similarly low pain medication prescription rates for Hispanic/Latino patients. Cintron et al.'s review paper regarding pain disparities reveals that 11 of 17 studies found that African American and Hispanic patients are less likely to receive pain medication and more likely to have their pain untreated compared to white patients [28]. Three studies revealed that minority patients are more likely to have under-treated pain compared to white people. All of these studies point to one conclusion: minorities' pain often goes ignored compared to their white counterparts. This dismissal by clinicians indicates a larger bias that still persists in the medical world where patients' symptoms and lived experiences are not taken seriously. While there is copious evidence to demonstrate how racism is still an issue in medicine, there have been no effective means to change racist beliefs within clinicians.

2.3 Importance of Studying Medical Notes

Explicitly stigmatizing language persists in everyday medical vernacular and may have consequences for patient care [29]. A recent study demonstrated that physicians who read a vignette with the term "substance abuser" as opposed to "having a substance use disorder" were more likely to have negative perceptions of the featured character and feel that they should face punishment for substance use [30]. Similarly, Goddu et al. used clinical vignettes to examine the effects of explicitly stigmatizing language

on providers' perceptions of patients with sickle cell disease (SCD) [31]. The authors found that when medical providers were shown a hypothetical chart note containing stigmatizing language, such as the derogatory term "sickler," they were more likely to have a negative perception of the patient's pain and to formulate a less aggressive pain management plan than when presented with a chart note with neutral language but the same objective patient information [31].

In medicine, clinicians typically look at patient medical history and read previous medical notes written by other clinicians. Not only are clinicians' attitudes reflected in the medical records they write, but previous medical records can influence clinicians' mindsets. As such, correcting for stigmatizing language in medical records would lessen the cycle of further generating stigma and hindering patient care. This paper seeks to deeply understand the use of stigmatizing language in patient medical notes to allow for a better grasp of its components and to enable the correction of stigmatizing language in medical notes.

2.4 Review of Studies on Stigmatizing Language in Medical Notes

Research involving stigmatizing language in medical notes has mainly been qualitative. Park et al. manually analyzed 600 randomly selected encounter notes from electronic medical records from an urban academic medical center [32]. The 600 encounter notes were written by 138 physicians in 2017. When categorizing the types of stigmatizing language used by physicians, the largest category identified was the questioning of the credibility of the patient, when the physician expresses disbelief of patient reports of their own experience or behaviors.

Beach et al. conducted a study where they analyzed 600 clinical notes to identify linguistic features that questioned a patient's credibility and cast doubt on the

patient's claims [33]. These linguistic features are: (1) quotes; (2) specific "judgment words" that suggest doubt; and (3) evidentials, words that allow for the rephrasing of a patients' symptoms or experience as hearsay.

Linguistic Feature	**Example**
Quotes	"The patient reports that she has 'pain' in her upper arm."
Judgement Words	"The patient claims she took her medication."
Evidentials	"The patient reports that the headache started yesterday."

Table 2.1: Three types of linguistic features that cast doubt on a patient

While quotes are often used to promote accuracy when citing a source, they can also have notes of sarcasm and dismissal. Quoting patients is often encouraged in medical training to humanize the patient and involve a patient in their own treatment plans. However, when clinicians make a conscious choice to write, "the patient reports she had 'pain' in her upper arm," they may be trying to indicate that they do not necessarily believe that the patient is experiencing pain or do not deem the pain sufficient for any treatment; instead the clinician is belittling a patient [32].

The authors compiled a list of specific judgment words that, when used to describe a patient's experience, convey a sense of suspicion or doubt by the clinician. The list of judgment words includes verbs like "claims", "insists", and "states." For example, by stating "the patient claims she took her medication," the clinician is directly casting doubt on whether the patient actually took her medication. Often times, these judgment words stem from the very roots of medical language that have now become standard. The use of the words "complain" and "deny" in the context of a medical case dates back to the 1800s [34, 35]. Medical training often refers to a patient's problem as a "chief complaint," which naturally leads to clinicians describing a patient's concerns as "complaints." Similarly, physicians have adopted the word "deny" to indicate a negative conclusion or lack of symptom. Even so, in considering

how the term 'denial' is used in daily vernacular, the term can correspond to a delusional refusal to accept an unpleasant or threatening truth. Despite the intention behind the term or how it was been normalized in the medical field, it does imply distrust towards a patient's experience [36].

While the judgment words serve as direct indications of casting doubt on patients, there are more subtle ways in which a medical professional can indicate their doubt. Evidentials are used to direct the truth of a statement to another source rather than the absolute truth. For example, a straight declarative statement ("The patient's headache started yesterday.") indicates certainty; the patient indeed has a headache, and it did start yesterday. As soon as the clinician adds an evidential ("The patient reports that the headache started yesterday."), it is no longer clear whether the patient actually has a headache or if the headache did indeed start yesterday. Instead, the clinician is describing the patient's statement as hearsay. These evidentials cast doubt in a more indirect sense by stating the patient's point of view rather than describing the statement as fact.

Sun et al. analyzed a sample of 40,113 history and physical notes (January 2019 – October 2020) from 18,459 patients for negative descriptors of the patient or the patient's behavior to understand the relationships between negative descriptors and race [37]. They used mixed effects logistic regression to determine the odds of finding at least one negative descriptor as a function of the patient's race or ethnicity. The authors manually curated a list of fifteen negative patient descriptors: "(non)adherent", "aggressive", "agitated", "angry", "challenging", "combative", "(non)compliant", "confront", "(non)cooperative", "defensive", "exaggerate", "hysterical", "(un)pleasant", "refuse", and "resist." These negative patient descriptors are clearly stigmatizing as they portray the patient in a negative light. Upon looking at these descriptors, we can see that the words generally suggest that the patient is non-compliant.

Gourabathina considered two objectives: (1) how the use of stigmatizing language that casts doubt on patients coincided with patient demographics, medical note types, and clinician categories and (2) common textual co-occurrences with known stigmatizing keywords to identify new keywords [9]. The former objective motivates this work in highlighting the importance of remediation of biases in medical notes. The statistically significant results showed that stigmatized language is present with greater likelihood in records of female patients, Black and Latinx patients, patients with federal health insurance plans, and in records for emergent rather than elective medical procedures. The latter objective provides us with keywords outside of those already documented in literature, such as "concerned", "confused", "delirious", and "dramatic". This work highlighted a new aspect of negatively-connoted language applied to patients: the emotional patient. When clinicians' notes highlight excessive emotions of patients, they indirectly question their credibility and represent their actions and words as consequences of emotion rather than reason. Other keywords referring to emotional states ("worried", "frantic", "upset", "distressed", "troubled") were identified.

Our model will incorporate the negative descriptors identified by Sun et al., the keywords that cast doubt on a patient identified by Beach et al., and the keywords identified by Gourabathina [37, 33, 9] (for full list of stigmatizing keywords, see Table 3.1). In general, we explore two categories of stimgatizing language in this work: language that (1) casts doubt on patients and (2) suggests non-compliance or excess emotion.

Chapter 3

Approach

We approached the task of creating this editing model by first considering our set of keywords and linguistic features from previous works (see Table 3.1). From there, we used patient medical notes to train a BERT transformer to create patient-oriented clinical word embeddings. These clinical word embeddings allowed us to find semantically similar words to the previously-identified keywords to obtain a quantitatively robust set of keywords. Following the development of our patient-oriented clinical word embeddings, we then created a thorough grammar structure that directs us from a statement containing stigmatizing language to one that is void of stigmatizing language. We used the Stanford CoreNLP parser to obtain constituency parses of the medical notes such that we could apply our grammar structure to medical notes to eliminate stigma.

3.1 Data

Our medical notes data is drawn from the MIMIC-IV-Note Clinical Database. MIMIC-IV-Note contains 331,794 deidentified discharge summaries from 145,915 patients admitted to the hospital and emergency department at the Beth Israel Deaconess

Beach et al. [33]	**Sun et al.** [37]	**Gourabathina** [9]
'alleges'	'(non-)adherent'	'concerned'
'asserts'	'aggressive'	'confused'
'attests'	'agitated'	'delirious'
'claims'	'angry'	'dramatic'
'complains'	'challenging'	'emotional'
'denies'	'combative'	'frantic'
'endorses'	'(non-)compliant'	'remarks'
'insists'	'confront'	'troubled'
'notes'	'(non-)cooperative'	'upset'
'proclaims'	'defensive'	'worried'
'protests'	'exaggerate'	
'reports'	'hysterical'	
'says'	'(un-)pleasant'	
'swears'	'refuses'	
'tells me'	'resists'	

(*) For all of the verbs, we look at the present tense plain form, past tense, and present participle. For instance, along with the keyword "claims", we also consider 'claim', 'claimed', and 'claiming.'

Table 3.1: Initial list of keywords and linguistic features used in our model.

Medical Center in Boston, MA, USA from 2008 to 2019 [38] [1]. The MIMIC-IV-Note database is an updated version of the MIMIC-III database [39]. Being newer, the MIMIC-IV database includes language that is more relevant and representative of clinician notes of today. Moreover, the notes in MIMIC-IV-Note have been compiled with the intention of use for NLP tasks with clinical free-text. Several notes in the MIMIC-III database often only included medical information, such as medications, prescriptions, and radiology scans or reports, with little information regarding patient behavior or any clinician commentary specific to a patient. The MIMIC-IV notes are intended for physician use and not available to patients. The notes are transferred to subsequent physicians as needed for the patient's care. The overall MIMIC-IV-Note data contains approximately 3.5 billion total words and a vocabulary of approximately 600 million unique words. Notes have a semi-structured header that includes the patient's name, gender, admission date, discharge date, and the category of the medical note, such as psychiatry, cardiothoracic, neurology, etc. The rest of the note consists of free-text that describes the patient's current conditions, past medical history, medications, and any other information deemed relevant by the clinician. Further information regarding pre-processing the clinical notes for development of the embedding models will be detailed in chapter 4.

3.2 Patient-Oriented Clinical Word Embeddings

Our goal with the development of clinical word embeddings is to create word vector representations specific to patient medical notes. Given our purposes of understanding stigmatizing language, our objective differs from that of previous contextual models in the clinical and biomedical domains. Other models have focused on understanding

[1] In the MIMIC-IV Clinical Database, identifying patient information has been removed, ensuring that subsequent identification of patients is not possible. A link is provided in the reference section to access the dataset. To view the data, users must be accredited and receive credentialed status by confirming that they will not maliciously use the data in any way and use the data only for the purpose of scientific research.

biomedical relations, such as drug-medical problem or medical problem-medical test, rather than actual semantic similarity. The language we are trying to decipher is not purely domain-specific or general purpose. To develop effective clinical word embeddings that capture stigma and identify semantically similar words to our keywords, we need our embeddings to both sift through and identify medical jargon but also understand non-medical language. Even language used in the medical domain that may not be particularly biomedical in nature can still have subtleties that are not applicable to English vernacular. [40, 36].

3.2.1 Contextual Word Embeddings

Word embeddings represent the semantic meaning of a word in vector form by considering its context or surrounding words. Two words that exist in similar contexts are considered more similar. Traditional (non-contextual) word-level vector representations, such as word2vec, GloVe, and fastText, express all possible meanings of a word as a single vector representation without taking the different contexts a word can appear into consideration [41, 42, 43]. For example, the word 'bark' would have a singular word embedding, despite the meaning being vastly different when referring to a 'dog bark' versus a 'tree's bark.' Contextual word representations such as ELMo and BERT have proven to be useful in tasks where context is essential and naturally lend themselves to the creation of more domain-specific word embeddings [44, 2]. After pre-training on a large text corpus that serves as a language model, ELMo creates context-sensitive embeddings for words in a given sentence, which will be fed into downstream tasks. BERT is similar to ELMo in the pre-training aspect, but is deeper and contains much more parameters, thus possessing greater representation power. More importantly, rather than simply providing word embeddings as features, BERT can be incorporated into a downstream task and gets fine-tuned as an integrated task-specific architecture, which allows for better evaluation for how the

model performs. BERT has generally been found to be superior to ELMo and even more specifically in the clinical domain [45]. For these reasons, our word embedding models are derived from BERT models.

There are two main components to BERT models: (1) pre-training and (2) fine-tuning. BERT uses a masked language modelling (MLM) objective, where some of the tokens of a input sequence are randomly masked. The MLM objective is to predict these randomly-masked positions by taking the corrupted sequence as input. BERT applies a Transformer encoder to attend to bi-directional contexts during pre-training. Additionally, BERT uses a next-sentence-prediction (NSP) objective. Given two input sentences, NSP predicts whether the second sentence is the actual next sentence of the first sentence. The NSP objective prioritizes reasoning over sentence pairs. For our purposes, this allows for the model to better understand the context of a discharge summary and the clinical narrative of a note. BERT uses special tokens to obtain a single contiguous sequence for each input sequence. Specifically, the first token is always a special classification token [CLS], and sentence pairs are separated using a special token [SEP]. The final hidden state of [CLS] is used for sentence-level tasks and the final hidden state of each token is used for token-level tasks. After the pre-training aspect, the model can be fine-tuned on a variety of tasks, such as entity recognition. Figure 3.1 details the general pre-training and fine-tuning procedures during the initial development of BERT [2].

Devlin et al. considered two model sizes when creating BERT: BERT-Base and BERT-Large. We turn our attention to BERT-Base, which has 12 layers, hidden size of 768, 12 self-attention heads, and 110M total parameters.

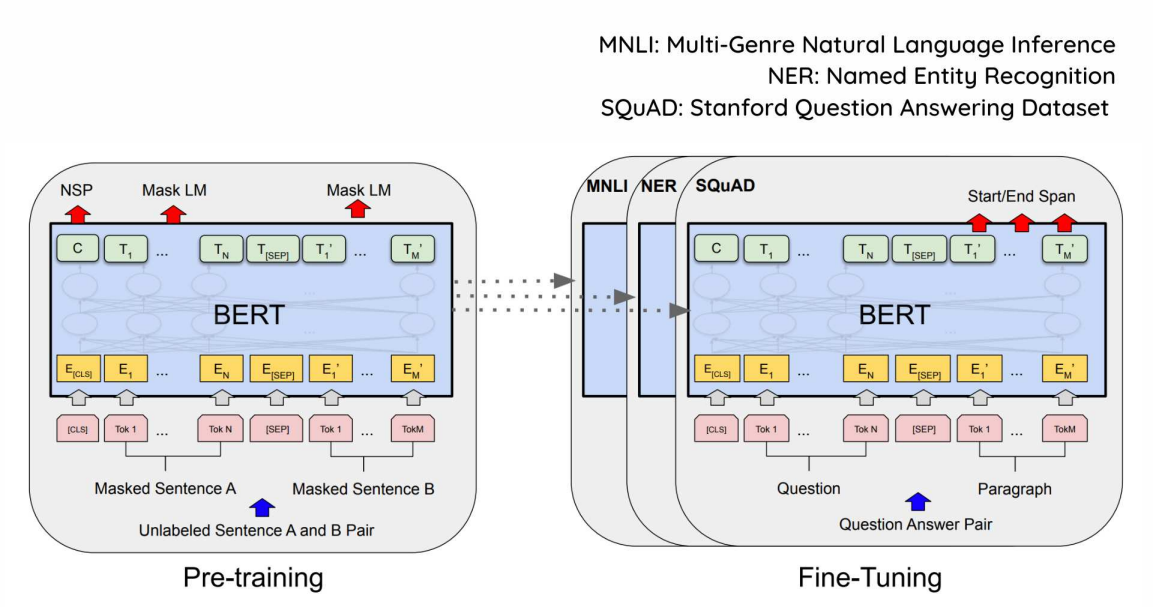

Figure 3.1: General pre-training and fine-tuning procedures for BERT [2]

3.2.2 Contextual Word Embeddings in Clinical and Biomedical Domains

Within the clinical and biomedical domains, there have been several approaches to create contextual embeddings, such as BioBERT and Clinical BERT models [45, 46]. BioBERT trains a BERT model over a corpus of biomedical research articles sourced from PubMed3 article abstracts and PubMed Central4 article full texts. They find the specificity offered by biomedical texts translated to improved performance on several biomedical NLP tasks.

Clinical BERT models are a collection of BERT-Base and BioBERT-initialized models trained on both clinical notes and only discharge summaries from the MIMIC-III database. Specifically, Discharge Summary BERT (initialized from BERT-Base) and Bio+Discharge Summary BERT (initialized from BioBERT) are models that are useful to us, given that they are trained on discharge summaries, the most comparable corpora to our MIMIC-IV-Note data. However, even when constricted to discharge summaries in MIMIC-III, the notes are not specifically free-text summaries

that describe patient behavior, experience, or health, as it is in MIMIC-IV. Essentially, MIMIC-III discharge summaries still consist of much more biomedical minutiae such as medications, reports, and medical tests than required for our purposes of better understanding stigmatizing language. As such, our approach will be to fine-tune the Discharge Summary BERT and Bio+Discharge Summary BERT models by training on MIMIC-IV-Note to create Patient-Oriented Discharge Summary BERT and Bio+Patient-Oriented Discharge Summary BERT. We will henceforth refer to these models as PODS BERT and Bio+PODS BERT for brevity. More detail regarding the architecture and training approaches will be described in chapter 4.

To check that PODS BERT and Bio+PODS are functioning properly, we apply the models to two clinical NLP tasks: (1) MedNLI natural language inference task and (2) 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [47, 48]. The performance of PODS BERT and Bio+PODS on these two tasks are simply sanity checks to ensure that the training process went correctly, and we do not expect these models to necessarily out-perform Discharge Summary BERT and Bio+Discharge Summary BERT, as that is not our goal. The MedNLI task provides a set of medical notes from the MIMIC-III database [2]. Patient information is then annotated by doctors into premise-hypothesis pairs. Each premise-hypothesis pair is classified as one of three classes: (1) entailment, (2) contradiction, or (3) neutral. For example, for the premise "Patient has type II diabetes.", we have the following three hypotheses:

1. Entailment: "Patient suffers from a chronic condition." (True)

2. Contradiction: "Patient's insulin levels are normal with no medication." (False)

3. Neutral: "Patient has hypertension." (Could be true or false)

[2]Like in the MIMIC-IV Clinical Database, identifying patient information has been removed, ensuring that subsequent identification of patients is not possible. To view the MedNLI data, users must be accredited and receive credentialed status by confirming that they will not maliciously use the data in any way and use the data only for the purpose of scientific research.

The goal of the task is to accurately classify each pair within the correct class.

The 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text consists of three sub-tasks: (1) a concept extraction task focused on the extraction of medical concepts from patient reports, (2) an assertion classification task focused on assigning assertion types for medical problem concepts, (3) and a relation classification task focused on assigning relation types that hold between medical problems, tests, and treatments. The data comes from manually-annotated patient reports from the VA Salt Lake City Health Care System in collaboration with i2b2 [3]. For the concept extraction task, the concepts are divided into (1) medical problems, (2) treatments, and (3) tests. The assertion categories are: (1) present, (2) absent, (3) possible, (4) conditional, (5) hypothetical, and (6) not associated with the patient. Relations may exist between: (1) medical problems and treatments, (2) medical problems and tests, and (3) medical problems and other medical problems. Given the uneven class distributions within the dataset for each sub-task, the Exact F1 metric has been used to calculate the accuracy of the 2010 i2b2/VA challenge.

Clinical NLP Task	**Metric**	**# of Sentences**		
		Train	**Dev**	**Test**
MedNLI	Accuracy	11232	1395	1422
2010 i2b2	Exact F1	14504	1809	27624

Table 3.2: An overview of the clinical NLP tasks used

These two tasks were chosen as they both serve as markers of whether the clinical text is semantically understood by PODS BERT and Bio+PODS BERT, but it must be noted that the tasks test whether the model understand biomedical concepts and relations rather than testing whether patient behavior, experience, and health is

[3]The i2b2 NLP data sets previously released on i2b2.org are now hosted by the Harvard University Department of Biomedical Informatics (DBMI) Data Portal under their new moniker, n2c2 (National NLP Clinical Challenges). To access the data set, one most confirm that they are using the data for only research purposes and with no malicious intent.

understood. To actually accomplish our goal, we then complete a nearest neighbors task on PODS BERT and Bio+PODS with respect to the 40 previously-identified keywords to identify additional keywords. To do so, we use the KNearestNeighbors algorithm from Scikit-learn with the metric of cosine similarity [49].

3.3 Grammar Structure and Constituency Parsing

After attaining our complete list of keywords, we then develop a grammar structure. Essentially, we develop a list of rules that takes into account every keyword and the syntactical environments in which the keyword occurs. The grammar rules translate a sentence from being stigmatizing to non-stigmatizing. Using the Stanford CoreNLP parser, we obtain constituency parses for each sentence in a given note. From there, we create our editing model that applies our grammar structure to the constituency parse to effectively edit the note. We name this editing model MedStiLE, which stands for Medical note Stigmatizing Language Editor.

The diagram below serves as a toy example of how MedStiLe uses the grammar structure and constituency parsing (see Figure 3.2). Chapter 6 will go into more detail about the extensive grammar structure and constituency parsing.

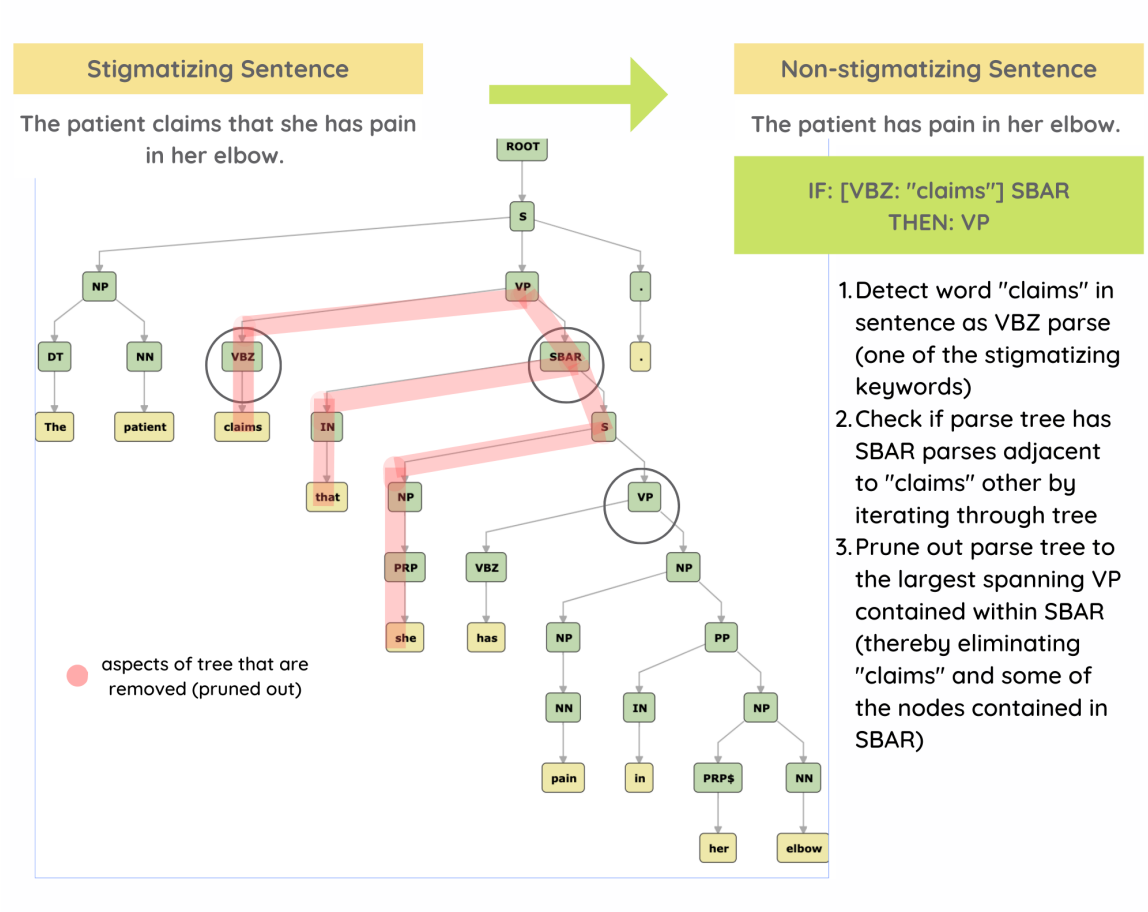

Figure 3.2: An example of the grammar structure and constituency parse method to edit notes to omit "claims" when occurring adjacent to a subordinating clause (SBAR tag). This is a common way that "claims" has occurred in the MIMIC-IV notes.

Chapter 4

Implementation

Our approach with creating the PODS BERT and Bio+PODS BERT models is summarized in Figure 4.1. This section will go through important implementation details, but the full code is available in Appendix A.

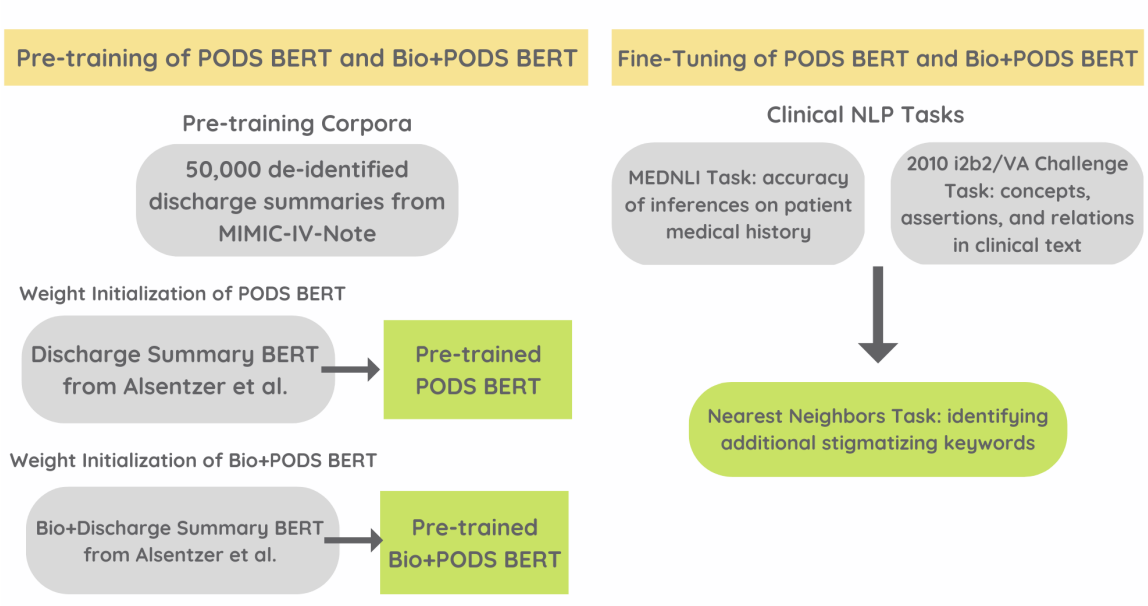

Figure 4.1: Overview of the pre-training and fine-tuning of PODS BERT and Bio+PODS BERT

4.1 Data Processing

We use clinical text from 50,000 deidentified discharge summaries from MIMIC-IV-Note. We first filtered the discharge summaries to only those that contained at least one of the 40 previously-identified keywords or relevant inflected forms of the keywords. Then, we selected the 50,000 longest discharge summaries of the filtered subset. We verified that all 40 of the previously-identified keywords were represented in the 50,000 selected notes.

Notes were cleaned to remove extra spaces and extraneous symbols such as asterisks. We then used the `en_core_sci_sm` tokenizer from Scispacy to perform sentence extraction on the notes [50]. The sentences are input into the Discharge Summary BERT and Bio-Discharge Summary BERT models for pre-training.

4.2 BERT Training and Evaluation

We trained two BERT models on clinical text: (1) PODS BERT, initialized from Discharge Summary BERT, and (2) Bio+PODS BERT, initialized from Bio+Discharge Summary BERT. For pre-training, the following hyperparameters were used:

- Batch size: 32
- Maximum sequence length: 128
- Learning rate: $5 * 10^{-5}$

Models were trained for 100,000 steps. For both clinical NLP tasks, the output BERT embedding was passed through a single linear layer for classification, based on the "begin sentence" token for MedNLI task and at a per-token level for the 2010 i2b2/VA task. For fine-tuning on the two tasks, the following hyperparameters were tested:

- Batch size: 16, 32

- Maximum sequence length: 128
- Learning rate: $3 * 10^{-5}$, $5 * 10^{-5}$
- Epochs: 2, 3

4.3 Nearest Neighbors Task

Following the training and fine-tuning process, we then retrieved the word vector representation of each stigmatizing keyword from the embedding matrix of the PODS BERT and Bio+PODS BERT models. We only consider the root form of the keyword (see Table 3.1) and not the inflected forms. BERT models create embeddings that are context-dependent, and thus, there are word vectors for each word in the entire text. Because there were several instances of each keyword and some instances were in a non-stigmatizing context, we manually selected one instance for each keyword for the associated word vector. For example, the word "reports" can occur in a non-stigmatizing context where a clinician is describing medical reports. The keyword instance was selected manually to ensure that the word vector corresponds to a stigmatizing occurrence. For each of these 40 word vectors, we calculated the five nearest neighbors through use of cosine similarity with the KNN algorithm of Scikit-learn [49].

Chapter 5

Results

5.1 Evaluation of Patient-Oriented Clinical Word Embeddings

We evaluate PODS BERT and Bio+PODS BERT on two clinical NLP tasks in comparison with Discharge Summary BERT and Bio+Discharge Summary BERT (see Table 5.1). We see that for MedNLI, the PODS BERT and Bio+PODS BERT slightly outperform Discharge Summary BERT and Bio+Discharge Summary BERT respectively. For 2010 i2b2, PODS BERT outperforms Discharge Summary BERT while Bio+PODS BERT slightly underperforms compared to Bio+Discharge Summary BERT. Again, our goal was not for PODS BERT and Bio+PODS BERT to serve as new performance benchmarks for MedNLI and 2010 i2b2 tasks. Instead, we wanted to verify that our models perform on par with the previously created Discharge Summary BERT and Bio+Discharge Summary BERT and are effective clinical embeddings.

Model	MedNLI	2010 i2b2
Discharge Summary BERT	80.2%	84.3%
Bio+Discharge Summary BERT	81.7%	86.5%
PODS BERT	80.4%	85.4%
Bio+PODS BERT	81.8%	86.1%

Table 5.1: Model performance on clinical NLP tasks in terms of Accuracy (MedNLI) and Exact F1 score (2010 i2b2)

5.2 Nearest Neighbors Task

For each of the 40 previously-identified keywords, we obtain the five nearest neighbors according to the word vectors from both PODS BERT and Bio+PODS BERT. We only consider the set of words from the 50,000 MIMIC-IV-Note discharge summaries used to train the models. We thus retrieve a total of 400 words. These 400 words fell into four general categories:

(1) Previously-identified keywords

(2) Non-stigmatizing words

(3) New keywords that cast doubt on the patient

(4) New keywords that suggest non-compliance or excess emotion

With words of type (1), we clearly do not add any new stigmatizing keywords to our list. These have been removed from the results table below. (2) refers to words that we deem not stigmatizing. Some of these words are clearly not stigmatizing, such as 'patient' and 'discharge.' Other words required closer inspection. To determine if a word fell into (3) or (4) in contrast to (2), we searched the discharge summaries for occurrences of the word. If there were no occurrences in which the word was casting doubt on the patient nor suggesting non-compliance or excess emotion, the word was categorized as (2). A word was categorized as (3) or (4) if there were any occurrences

in which the word could be deemed stigmatizing, even if not all the occurrences were in a stigmatizing context. This typology is described for ease in presenting the results, as there are certain words like 'argues' and 'demands' that can both cast doubt on the patient and imply excess or unreasonable emotion. Moreover, chapter 6 will consider a more detailed categorization of all of the keywords that is used for applying the grammar structure.

Overall, we have identified 30 new stigmatizing keywords. Upon taking a cursory glance at Table 5.2, it may seem that some of the words could be stigmatizing, especially in light of some of the previously-identified stigmatizing keywords such as 'hysterical' and 'emotional.' In taking a closer look at words like 'delusional', 'paranoid', 'hysteria', 'manic', and 'depressed,' we saw that all occurrences of these words in our data were describing a patient's psychiatric or mental state. Some of the previously-identified keywords and newly-identified type (4) keywords also occur in contexts of simply describing a patient's psychiatric or mental state and are not always used in a stigmatizing manner. Chapter 6 will discuss the keywords in much more detail including the contexts in which each of these words are stigmatizing and how they behave semantically similar to previously-identified keywords.

For the detailed results of the five nearest neighbors from the PODS BERT and Bio+PODS BERT embeddings for all 40 keywords, see Appendix B.

Type (2)	Type (3)	Type (4)
'ask'	'argues'	'annoyed'
'chief'	'believes'	'confrontational'
'delusional'	'comments'	'cranky'
'depressed'	'demands'	'difficult'
'discharge'	'explains'	'impulsive'
'(dis)likes'	'perceives'	'restless'
'distant'	'promises'	'annoyed'
'doctor'	'states'	'difficult'
'failed'	'suggests'	'frenzy'
'hallucination'		'horrible'
'hysteria'		'impulsive'
'manic'		'mad'
'paranoid'		'melodramatic'
'patient'		'negative'
'strange'		'panicked'
'time'		'reluctant'
'unsettled'		'resistant
'unsure'		'restless'
'unyielding'		'stubborn'
		'sad'
		'unhappy'

Table 5.2: Consolidated results from nearest neighbors task. Type (2) refers to words that are deemed non-stigmatizing. Type (3) is newly-identified keywords that cast doubt on a patient. Type (4) is newly-identified keywords that suggest non-compliance or excess emotion.

Chapter 6

Grammar

In order to create our grammar structure, we thoroughly consider the various contexts that the stigmatizing keywords we have identified appear and how we can correct for them.

6.1 Analyzing Stigmatizing Keywords

In total, we now have 70 stigmatizing keywords. As previously discussed, these keywords either cast doubt on a patient or portray the patient negatively in ways that suggest non-compliance or excess emotional sensitivity. Previous literature discussing language that casts doubt on patients details two types of words: (1) judgement words and (2) evidentials [33]. Judgment words are words that inherently suggest doubt, and evidentials are words that allow for the rephrasing of a patient's experience as hearsay. We also split the keywords that are negative descriptors of patients into types: (1) descriptors of noncompliance and (2) descriptors of emotion. As we previously discussed, there can be overlap between descriptors of noncompliance and emotion, where a certain keyword may function as both. Altogether, we have four classes of stigmatizing keywords (see Table 6.1). We use the word 'class' to avoid confusion with the 'types' mentioned in chapter 5.

(1) Judgement words

(2) Evidentials

(3) Descriptors of noncompliance

(4) Descriptors of emotion

This classification allows us to more easily develop a grammar structure.

6.1.1 Judgment Words

Judgment words more explicitly question a patient's credibility than evidentials do. An example of a judgment word from existing literature is the word 'claims' [33]. By stating that the patient is 'claiming' something when describing their symptoms, the doctor is directly casting doubt on the patient's words. We identified five additional judgment words from our patient-oriented clinical embeddings: 'believes', 'argues', 'demands', 'promises', and 'perceives' (see Table 6.1). We provide example sentences for each of the words below (see Table 6.2).

The word 'believes' and 'perceives' function quite similarly to 'claims' where the clinician creates a separation between reality and the patient's belief or perception. The words 'argues' and 'demands' suggest a sense of disagreement between the opinion of the patient and the clinician, where the clinician is clearly conveying an opposing point of view that they do not believe or agree with, thereby casting doubt on the patient. The use of these words emphasizes a sense of conflict between the patient and clinician. Finally, the word 'promises' casts doubt on the patient following through on their word and shifts blame if the patient were to break the promise. For example, in the sentence in Table 6.2, if the patient were to not take their medications properly, it would be the patient's fault and the clinician is removing themselves from that responsibility. On a more basic level, the clinician is certainly casting doubt on

Class 1	Class 2	Class 3	Class 4
'alleges'	'complains'	'(non)adherent'	'hysterical'
'asserts'	'denies'	'aggressive'	'exaggerate'
'attests'	'endorses'	'agitated'	'concerned'
'claims'	'notes'	'angry'	'confused'
'insists'	'remarks'	'challenging'	'delirious'
'proclaims'	'reports'	'combative'	'dramatic'
'protests'	'says'	'noncompliant'	'frantic'
'swears'	'tells'	'confront'	'troubled'
'argues'	'comments'	'(non)cooperative'	'emotional'
'believes'	'explains'	'defensive'	'worried'
'demands'	'states'	'unpleasant'	'upset'
'perceives'	'suggests'	'refuses'	'frenzy'
'promises'		'resists'	'impulsive'
		'annoyed'	'melodramatic'
		'confrontational'	'panicked'
		'cranky'	'restless'
		'difficult'	'sad'
		'frustrated'	'unhappy'
		'horrible'	
		'mad'	
		'negative'	
		'pessimistic'	
		'reluctant'	
		'resistant'	
		'rude'	
		'stubborn'	
		'uncooperative'	

Table 6.1: Complete list of keywords. Words in red are the newly-identified keywords. Class 1 is judgment words. Class 2 is evidentials. Class 3 is descriptors of noncompliance. Class 4 is descriptors of emotion.

Example Sentences of Judgment Words	
'argues'	She argues that her symptoms are consistent with something more serious.
'believes'	The patient believes his nausea has began since starting the new medication.
'demands'	He demands an increase in his dosage.
'perceives'	His GI symptoms are baseline or improved with the exception of epigastric discomfort that he perceives as new in the past week.
'promises'	She promises to take her medications consistently.

Table 6.2: Example sentences containing judgment words. These sentences are from the MIMIC-IV-Note database and have been modified to remove specific medication names.

the patient's ability to take the medications by noting down that the patient has 'promise[d].'

6.1.2 Evidentials

Evidentials are words that imply that information is coming from an external source (the patient) rather than the clinician themselves. For example, with the statement "The patient notes that she has pain in her lower back," there is a lack of certainty of whether she actually has pain in her lower back. By framing it as hearsay from the patient, the clinician casts doubt on the validity of the patient's statement. All four of the newly-identified evidentials are quite neutral in tone ('suggests', 'comments', 'states', and 'explains') and function similarly to 'notes.' Examples of each word in a stigmatizing context are provided below (see Table 6.3).

Example Sentences of Evidentials	
'comments'	He comments that the spot has been there for several months.
'explains'	She explains that her symptoms have all resolved since the last appointment.
'states'	He states his back pain has mildly improved.
'suggests'	She suggests she experienced palpitations prior to "passing out."

Table 6.3: Example sentences containing evidentials. These sentences are from the MIMIC-IV-Note database.

In general, evidentials may not be seen as explicitly stigmatizing compared to

judgment words, but if we compare a statement with an evidential to one without, the extent to which an evidential casts doubt becomes much more obvious:

- He comments that the rash has been there for several months.
- The rash has been there for several months.

6.1.3 Descriptors of Noncompliance

Descriptors of noncompliance generally suggest that the patient is difficult or not cooperating with the clinician. We can clearly see how all 14 of the newly-identified keywords of this class are semantically similar to the previously-identified descriptors of noncompliance ('annoyed', 'confrontational', 'cranky', 'difficult', 'frustrated', 'horrible', 'mad', 'negative', 'pessimistic', 'reluctant', 'resistant', 'rude', 'stubborn', and 'uncooperative'). These descriptors are very explicitly stigmatizing by characterizing the patient in a negative light (see Table 6.5). While sometimes these words can be used to indicate the mental state of a patient, especially in psychiatric notes, it is important for clinicians to be careful when using these words. They must recognize how their words portray patients and how a negative portrayal can harm a patient's ability to receive proper treatment and care. Moreover, these descriptors are often used to emphasize the difficult nature of a patient or emphasize a sense of conflict between the patient and clinician. Considering the power dynamic between a clinician and patient, such negative descriptors can be used to subjugate patients.

Example Sentences with Descriptors of Noncompliance	
'frustrated'	The patient became extremely frustrated and began yelling.
'cranky'	Patient was cranky and groggy following the operation.
'uncooperative'	He is largely uncooperative but vitals remain normal.
'stubborn'	He is too stubborn to accept input from his family.

Table 6.4: Example sentences containing some of the newly-identified descriptors of noncompliance. These sentences are from the MIMIC-IV-Note database.

6.1.4 Descriptors of Emotion

Descriptors of emotion may suggest that the patient is displaying emotion beyond what is considered reasonable. There are many cases in which words of this category are used in psychiatric settings where the clinician must take note of a patient's mental or emotional state. However, they can be used in stigmatizing ways that portray a patient as dramatic or overly emotional. Just like the descriptors of noncompliance, the clinician must be careful when using such words to describe a patient. In the following examples, some patients are described in less flattering ways to make them seem overly emotional and unreasonable. In reading the entirety of the notes from which these statements are derived, the tone of the sentences become even more clear, as these statements in isolation might seem like simple descriptors of patient behavior. Rather, by emphasizing the emotional state of the patient, clinicians portray the patients negatively, despite the patients having just undergone medical procedures and experiencing immense stress about their health. The words we have newly-identified are 'frenzy', 'impulsive', 'melodramatic', 'panicked', 'restless', 'sad', and 'unhappy.'

Example Sentences with Descriptors of Emotion	
'restless'	He is restless, inattentive, and continues to ignore my questions.
'panicked'	She panicked immediately, and then drove herself to the emergency department.
'melodramatic'	The patient is melodramatic about her family.
'unhappy'	She seems extremely unhappy with services from the hospice company.

Table 6.5: Example sentences containing some of the newly-identified descriptors of emotion. These sentences are from the MIMIC-IV-Note database.

6.2 Grammar Rules

To develop the grammar rules, we look at the sentence structures in which the stigmatizing keywords occur. We study each of the four classes of keywords separately, given how each class induces stigma differently.

6.2.1 Judgment Words

We consider the various ways judgment words are used in language. First, we have the occurrence of the judgment words occur adjacent to a noun phrase (NP). This would be a sentence like (6.1). In this sentence, 'claims' is a verb and 'pain in her elbow' is a noun phrase.

The patient claims pain in her elbow. (6.1)

Next, we consider a variation on the sentence where the judgment word occurs adjacent to a complete sentence (S). In (6.2), 'she has pain in her elbow' is a complete sentence (S):

The patient claims she has pain in her elbow. (6.2)

We additionally consider the following sentence with a subordinating conjunction 'that':

The patient claims that she has pain in her elbow. (6.3)

We also consider the following sentence that contains 'to' (TO tag):

The patient claims to have pain in her elbow. (6.4)

Another slight variation on (6.4) is using a gerund (VBG):

The patient claims to having pain in her elbow. (6.5)

Finally, we consider this last variation without the TO:

The patient claims having pain in her elbow. (6.6)

With each of these six variations of sentence structure, we develop grammar rules to correct from stigmatizing to non-stigmatizing by studying constituency parses. We tabulate the six variations below for easy reference (see 6.6).

Variations	
Variation 1	[keyword] NP
Variation 2	[keyword] S
Variation 3	[keyword] IN S
Variation 4	[keyword] TO VB
Variation 5	[keyword] TO VBG
Variation 6	[keyword] VBG

Table 6.6: Six variations of the sentence structure in which the judgment words occur. It must be noted that the 'S' in variations 2 and 3 are also equivalent to a subordinating clause (SBAR) due to the function of the judgment word as a verb.

Certain judgment words may not occur in these syntactical structures. For instance, "The patient believes pain in her elbow." does not occur, as the word 'believes' does not co-occur directly with a noun phrase in a stigmatizing context. A sentence like "The patient believes she has pain in her elbow." would occur. Nevertheless, we do not form individual rules according to each keyword. Because of the fast pace in which clinicians work and how notes consist of several non-grammatical elements, we apply the grammar rules developed for variations 1 to 6 for all of the judgment words in case there are discrepancies in clinicians' phrasing. This does not lead to any issues

because if a given judgment word does not occur in a certain syntactic context, the rule simply will not be applied.

The following figures illustrate the grammar we develop for judgment words (see Figure 6.1-6.3). When occurring adjacent to a noun phrase, as in variation 1, we simply change the judgment word to 'has' or 'have.' For variations 2 and 3, we consider the case of a judgment word occurring next to a subordinating clause (SBAR). In this case, we simply prune the tree to replace the complete sentence (S) containing the judgment word with the complete sentence (S) within the SBAR. For variations 4, 5, and 6, we modify the sentence structure such that the primary verb in the sentence (S) containing the judgment word changes to eliminate the judgment word and replace with the verb of the gerund or the verb adjacent to the 'TO' parse.

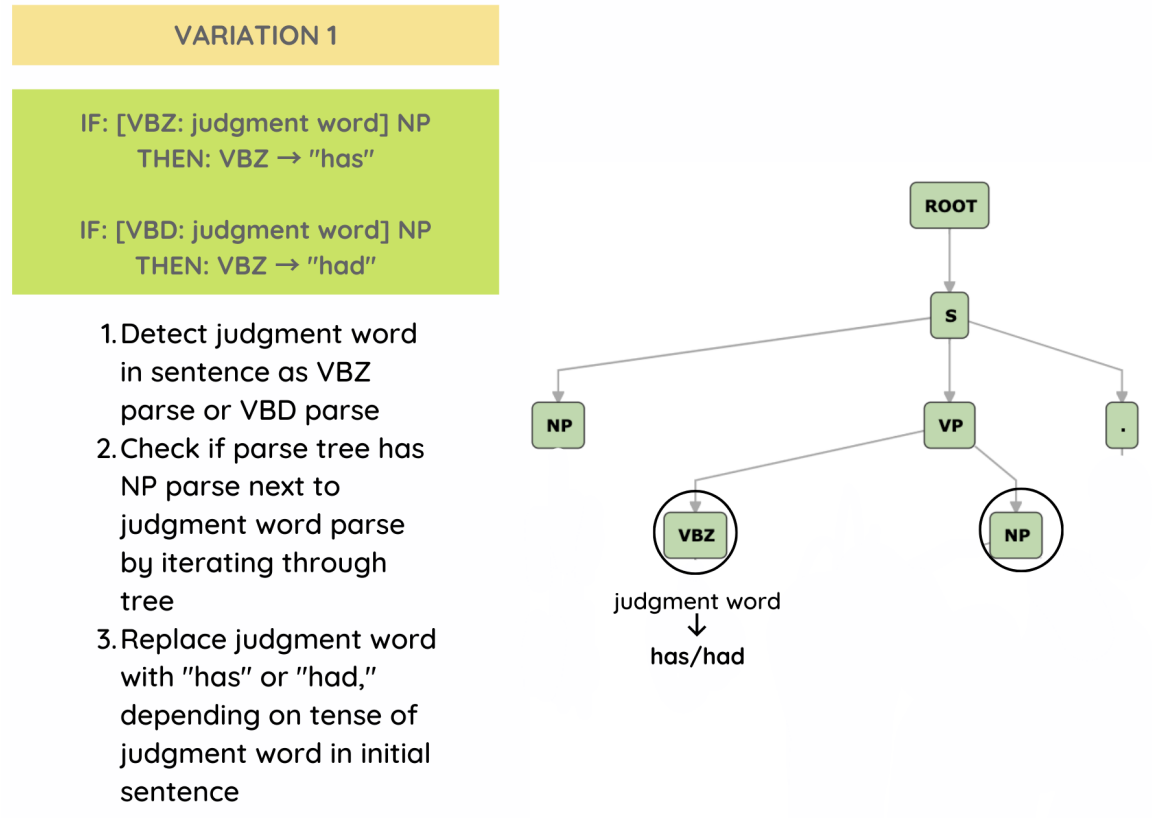

Figure 6.1: Grammar rule for variation 1, with judgment word adjacent to noun phrase

VARIATIONS 2 AND 3

IF: [VBZ: judgment word] SBAR
THEN: S

IF: [VBD: judgment word] SBAR
THEN: S

1. Detect judgment word in sentence as VBZ parse or VBD parse
2. Check if parse tree has SBAR parse next to judgment word parse by iterating through tree
3. Prune out parse tree to the largest spanning S contained within SBAR

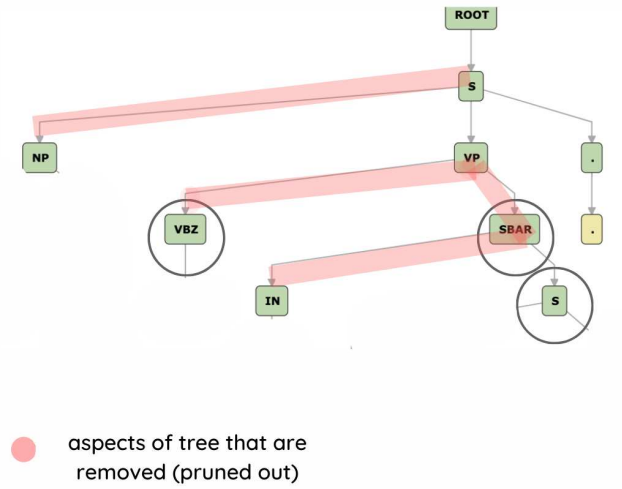

Figure 6.2: Grammar rule for variations 2 and 3, with judgment word adjacent to subordinating clause. It must be noted that the rule works regardless of whether there is a subordinating conjunction or not

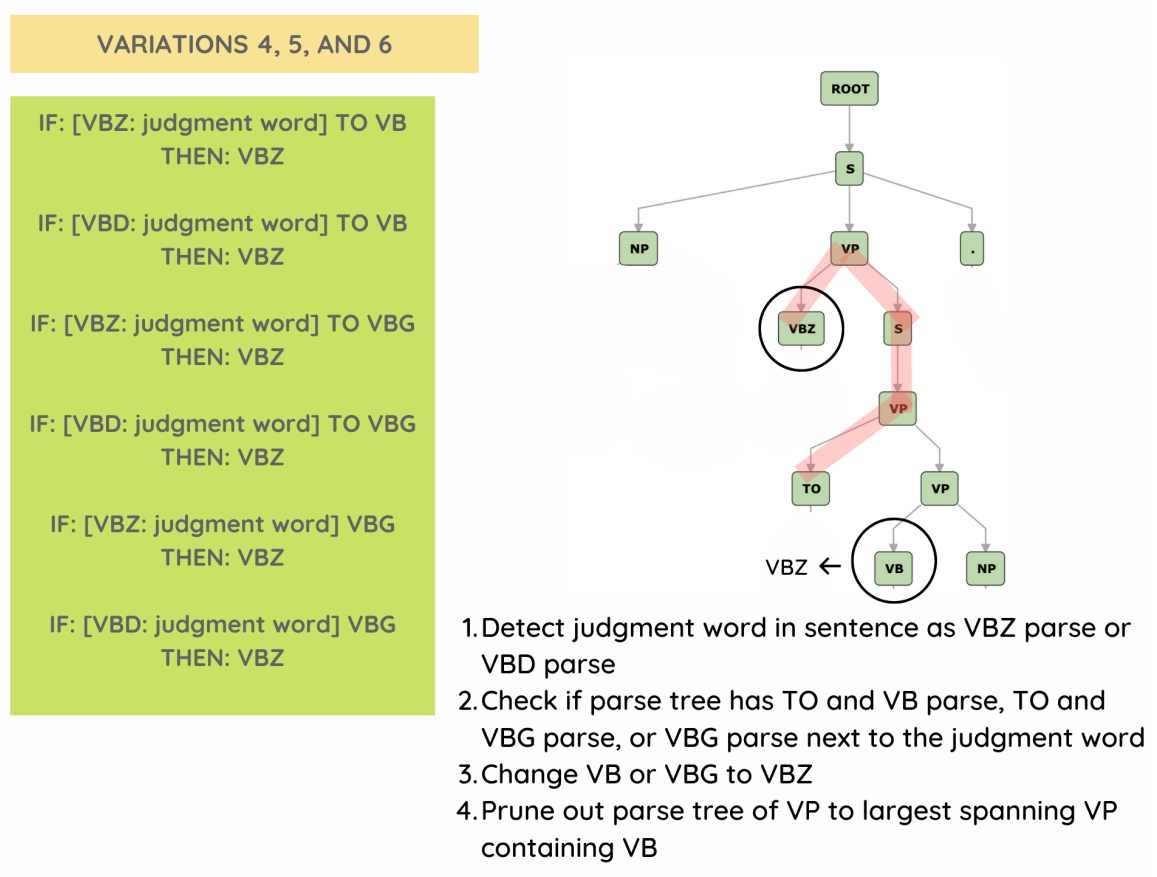

Figure 6.3: Grammar rule for variations 4, 5, and 6. The judgment word is either next to (1) TO and VB parse, (2) TO and VBG parse, or (3) VBG parse.

The figure below illustrates how sentences (6.1)-(6.6) would be corrected by the grammar structure (see Figure 6.4).

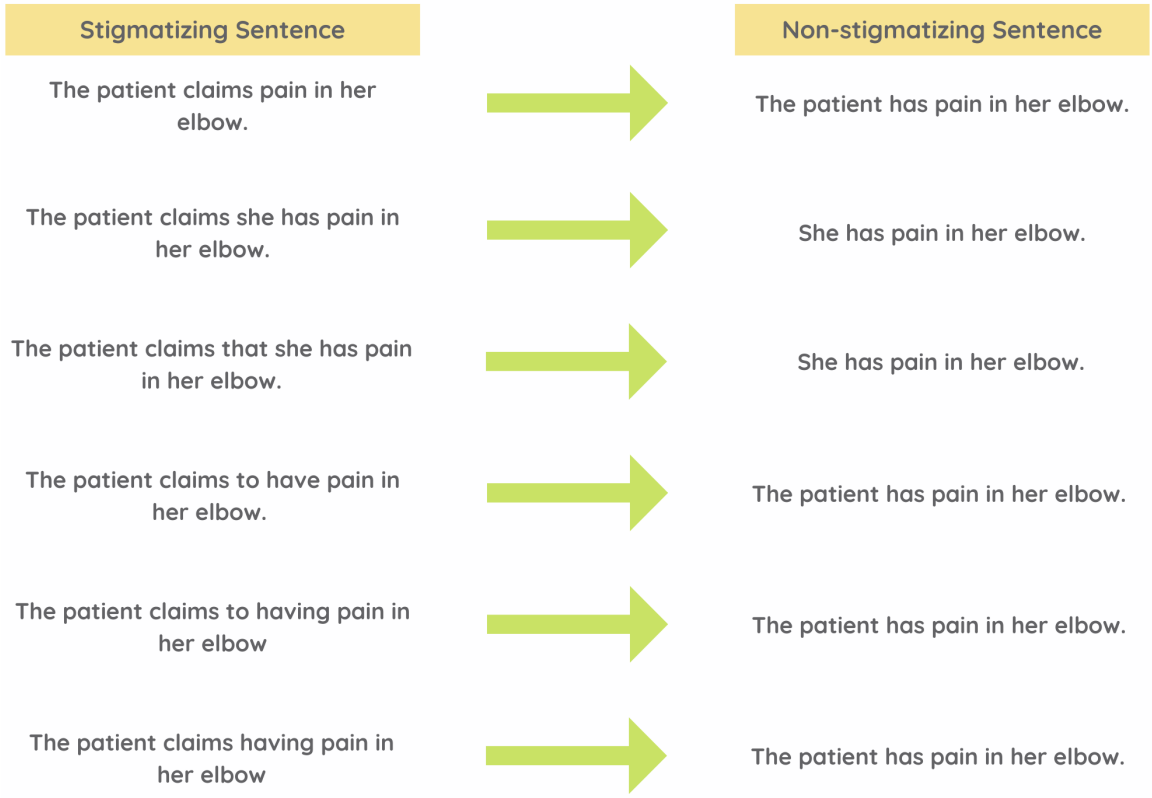

Figure 6.4: Correcting sentences (6.1)-(6.6) with grammar structure

6.2.2 Evidentials

Evidentials behave in similar ways syntactically to the judgment words. For example,

1. She notes pain in her elbow.
2. She notes she feels pain in her elbow.
3. She notes that she has pain in her elbow.
4. She notes having pain in her elbow.

As such, we apply the rules of variations 1, 2, 3, and 6 to the vocabulary of evidential stigmatizing keywords. Two of the evidential keywords, 'complains' and 'denies'

require additional attention. The word 'complains' often occurs in conjunction with the subordinating conjunction 'of', where it is followed by a noun phrase or gerund. For example, we have the statements "She complains of pain." and "She complains of feeling pain." To take this into account, we apply the grammar rules of variations 1 and 6 and view "complains of" and "complained of" as units. Unlike all the other evidentials, 'denies' negates the meaning of subsequent noun phrases, subordinating clauses, and gerunds. We thus modify the rules to insert 'does not' or 'did not' into the new sentence. For instance, when applying the rule of variation 6 to the sentence "She denies having pain in her elbow," we obtain "She has pain in her elbow." We now modify this statement to be "She does not have pain in her elbow."

6.2.3 Descriptors of Noncompliance and Emotion

The descriptors of noncompliance (class 3) and descriptors of emotion (class 4) are words that explicitly portray the patient in a negative light. We have two verbs in class 3 ('refuses' and 'resists'). All of the other words in class 3 and 4 are adjectives that plainly stigmatize a patient's behavior. Our grammar model simply removes sentences containing the adjectives of class 3 and class 4 from the note. However, for the two verbs in class 3, we believe that they can still retain important information on patient feedback, such as a patient declining a certain medical exam or not wanting to switch medical prescriptions. Thus, we simply develop a rule where we replace 'refuses' and 'resists' with 'does not want.' In cases where the the verbs are in past tense, 'refused' and 'resisted' are replaced with 'did not want.'

6.3 Applying Grammar Rules

To apply the grammar structure to a note, we first tokenize the sentences. Then, we do an initial pass of all the sentences to see if any of our 70 stigmatizing keywords

appear. For each sentence in which a keyword appears, we use the Stanford CoreNLP parser to obtain a constituency parse. From there, we apply our grammar rules. The figure below illustrates our overall grammar structure (see Figure 6.5). This entire framework of using the grammar structure and constituency parsing makes up our editing model: MedStiLE (Medical note Stigmatizing Language Editor). Given a note, MedStiLE produces a destigmatized output.

Judgment words:

IF: [VBZ: keyword] NP, THEN: [VBZ:keyword] → "has"

IF: [VBD: keyword] NP, THEN: [VBD:keyword] → "had"

IF: [VBZ: keyword] SBAR, THEN: S

IF: [VBD: keyword] SBAR, THEN: S

IF: [VBZ: keyword] TO VB, THEN: VBZ

IF: [VBD: keyword] TO VB, THEN: VBZ

IF: [VBZ: keyword] TO VBG, THEN: VBZ

IF: [VBD: keyword] TO VBG, THEN: VBZ

IF: [VBZ: keyword] VBG, THEN: VBZ

IF: [VBD: keyword] VBG, THEN: VBZ

Evidentials:

IF: [VBZ: keyword] NP, THEN: [VBZ:keyword] → "has"

IF: [VBD: keyword] NP, THEN: [VBD:keyword] → "had"

IF: [VBZ: keyword] SBAR, THEN: S

IF: [VBD: keyword] SBAR, THEN: S

IF: [VBZ: keyword] VBG, THEN: VBZ

IF: [VBD: keyword] VBG, THEN: VBZ

IF: ['complains of'] NP, THEN: ['complains of'] → "has"

IF: ['complained of'] NP, THEN: ['complained of'] → "had"

IF: ['complains of'] VBG, THEN: VBZ

IF: ['complained of'] VBG, THEN: VBZ

IF: ['denies'] NP, THEN: ['denies'] → "does not have"

IF: ['denied'] NP, THEN: ['denied'] → "did not have"

Descriptors of Noncompliance and Emotion:

IF: [JJ: keyword], THEN: ∅

IF: [VBZ: 'refuses', 'resists'], THEN: [VBZ: 'refuses'] → "does not want"

IF: [VBZ: 'refused', 'resisted'], THEN: [VBZ: 'refused', 'resisted'] → "did not want"

Figure 6.5: Complete list of rules of our grammar structure

Chapter 7

Evaluation

We have successfully developed our model MedStiLE that allows us to edit notes to de-stigmatize them. To fully analyze how effective MedStiLE is, we have conducted an evaluation plan in which human raters manually assessed a random sample of statements from notes. Half of the statements have been processed through our model and meant to be de-stigmatized. The other half are in their original form with no edits made.

7.1 Methodology

7.1.1 Sampling Approach

In selecting statements that contain stigmatizing language, we consider all four classes of stigmatizing language that our model targets: (1) judgment words, (2) evidentials, (3) descriptors of noncompliance, and (4) descriptors of emotion. We stratify our sample to contain all four classes to ensure that our model is effectively targeting all aspects of stigmatizing language and to also verify that our perceptions of what constitutes stigmatizing language are correct across the board. We select 10 statements corresponding to each category, with each statement corresponding to a different stig-

matizing keyword. These statements are directly from the MIMIC-IV-Note database, but we ensure that any patient information is removed. We then run all 40 statements through the model. After, we selected original statements and de-stigmatized statements from each category, such that none of the de-stigmatized statements were descended from the original statements that were selected. In other words, if the statement "The patient claims pain in her elbow." was selected, we did not also select "The patient has pain in her elbow.", as we did not want raters to view this evaluation as a matching task. Given that MedStiLE removes statements containing adjective keywords in class 3 and class 4, the statements processed by the model correspond only to the verbal keywords while the original statements include the adjectival keywords as well. We included the adjectival keywords in the original statements as we do want to see how they are evaluated by the raters. The following table shows our final distribution of statements in the sample:

	Class 1	**Class 2**	**Classes 3 and 4**
Original statements	5	5	10
Processed statements	5	5	10

Table 7.1: Class distribution for evaluation sample. Class 1 is judgment words. Class 2 is evidentials. Class 3 is descriptors of noncompliance. Class 4 is descriptors of emotion. For the original statements, there are five statements corresponding to class 3 and five statements corresponding to class 4.

7.1.2 Rating Approach

Each statement was evaluated by 20 independent raters [1]. The raters were given a guide to completing this evaluation task in the form of an assignment for TRA 301/COS 401. The guide includes details on the basis for stigmatizing language in patient medical notes and a brief explanation on what stigmatizing language may look

[1]The raters are the students of Princeton's Introduction to Machine Translation course (TRA 301/COS 401) in the Spring 2023 semester.

like. We intentionally do not provide example statements as raters may be biased in identifying certain keywords or phrases as stigmatizing. We then ask raters to rate each of the 40 statements on an integer scale from 1 to 5, where 1 represents "no stigmatizing language" and 5 represents "strongly stigmatizing language." See Appendix C for the full contents of this rating guide and the 40 statements used in the sample.

7.2 Results

We standardize the scores each rater assigns to each statement through z-score normalization. A rater's scores were subtracted by the mean of that rater's scores and divided by the standard deviation of that rater's score. We then looked at the distribution of standardized scores attributed to the original statements and the distribution of standardized scores attributed to the processed statements. A paired two-sampled t-test was conducted to compare the sample means of both distributions to determine if there is a statistically significant improvement in bias with our model. In chapter 8, we further discuss our rationale for using a paired two-sampled t-test and how we take the assumptions of this statistical test into account.

The original statement scores (standardized) had a mean of 2.44 with a standard deviation of 0.75 ($n = 20 * 20 = 400$), and the processed statement scores (standardized) from the raters had a mean of -2.17 with a standard deviation of 0.64 ($n = 400$). The paired two-sample t-test yielded a two-sided $p < .0001$, which is extremely statistically significant. We also looked at rater agreement. Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items [1]. There was moderate agreement between overall raters' judgments with $\kappa = 0.564$ (see Table 7.2).

κ	**Interpretation**
< 0	Poor agreement
$0.01 - 0.20$	Slight agreement
$0.21 - 0.40$	Fair agreement
$0.41 - 0.60$	Moderate agreement
$0.61 - 0.80$	Substantial agreement
$0.81 - 1.00$	Almost perfect agreement

Table 7.2: Interpretation of Fleiss' kappa scores for rater agreement [1]

Along with the overall score distribution, the score distributions for each class of stigmatizing keywords was also studied.

	Class 1	**Class 2**	**Classes 3 and 4**
Mean score for original statements	1.98	1.74	3.34
Mean score for processed statements	-2.45	-2.34	-1.17
Standard deviation for original statements	0.77	0.78	0.69
Standard deviation for processed statements	0.66	0.54	0.35
p-value from paired t-test	<0.0001	<0.0001	<0.0001

Table 7.3: Class-stratified evaluation results for paired two-sample t-test. Class 1 is judgment words. Class 2 is evidentials. Class 3 is descriptors of noncompliance. Class 4 is descriptors of emotion.

The rater agreement scores were also calculated for different classes and for both the original and processed notes (see Table 7.4).

	Overall	**Class 1**	**Class 2**	**Class 3**	**Class 4**
Original statements	0.507	0.498	0.423	0.722	0.552
Processed statements	0.646	0.654	0.778	0.607	
All statements	0.564	0.489	0.478	0.684	0.552

Table 7.4: Detailed Fleiss' kappa scores including results for each class of stigmatizing keywords and specified to original statements and processed statements. Class 1 is judgment words. Class 2 is evidentials. Class 3 is descriptors of noncompliance. Class 4 is descriptors of emotion. Because MedStiLE simply eliminates sentences containing keywords from class 4, no score is reported for class 4 and processed statements.

Chapter 8

Discussion

This section provides a discussion on the findings in chapters 5-7. Specifically, we consider the implications of our patient-oriented clinical word embeddings, grammar structure, and evaluation results.

8.1 Patient-Oriented Clinical Word Embeddings

To create our patient-oriented clinical word embeddings, we have developed two models: PODS BERT and Bio+PODS BERT. Compared to the baseline models of Discharge Summary BERT and Bio+Discharge Summary BERT, PODS BERT and Bio+PODS BERT perform similarly on two clinical NLP tasks. Because these clinical NLP tasks involve great semantic understandings with medical entities and relations, these embeddings can be used for a variety of tasks. Specifically, given that the primary purpose of these embeddings is on patient-oriented data, the embeddings can be used for relation-extraction and medical inference on free-text patient data.

PODS BERT and Bio+PODS BERT have been trained on 50,000 discharge summaries. With access to greater computational power, future works can expand on PODS BERT and Bio+PODS BERT by training the models on the entirety of the MIMIC-IV-Note dataset of 331,794 discharge summaries. Also, more time can be

spent in assessing model performance on the two clinical NLP tasks, with a more extensive hyperparameter tuning process.

For the nearest neighbors task, we find five nearest neighbors on all 40 previously-identified keywords with both the PODS BERT and Bio+PODS BERT model, yielding a total of 400 words. Looking at the complete results (see Appendix B), we can see that both models performed similarly in the nearest neighbors task. We anticipate that looking at more than five neighbors for each of the keywords would provide us with even more newly-identified keywords. We see several instances of keywords we have already identified or inflected forms of the keywords appear, which is not a surprising result. Given that our notes have been selected to include one or more of the previously-identified keywords, it was likely that many of the same keywords would appear in the results of the nearest neighbors task. Our rationale for filtering to only notes containing the previously-identified keywords was two-fold: (1) first, we had to downsample the data given our limitations in computational power and (2) second, we wanted to ensure that all of our discharge summaries focused on describing patient behavior rather than merely describing medical tests and procedures with primarily medical terminology. By selecting for notes that already contained previously-identified stigmatizing keywords, we ensure that the notes are patient-oriented.

There are certain instances of non-stigmatizing words that appear. This is likely because certain words like 'patient' and 'discharge' commonly occur in similar contexts to the stigmatizing keywords that describe patient behavior or experience. Some words were deemed non-stigmatizing since the MIMIC-IV-Note data did not contain any occurrences of the words in a stigmatizing context. It was decided a word like 'paranoid' or 'manic' was used in a non-stigmatizing context if the note was in the psychiatric category and used to plainly describe patient behavior. In contrast, words like 'melodramatic', 'sad', and 'impulsive' were used to critique patient response and

behavior in some notes. In total, we identified 30 new stigmatizing keywords (see Table 5.2).

8.2 Grammar Structure and Constituency Parsing

We divide the 70 total stigmatizing keywords into four classes: (1) judgment words, (2) evidentials, (3) descriptors of noncompliance, and (4) descriptors of emotion. Our typology of the keywords is not one that is unequivocal. Certain keywords may both cast doubt on a patient and also suggest noncompliance, such as 'argues.' Additionally, judgment words often function as evidentials where they rephrase a patient's statement as hearsay. For example, "She claims she has pain in her elbow." rephrases the patient's statement about experiencing pain in her elbow as hearsay. However, the word 'claims' also directly casts doubt on the validity of the patient's statement. In cases of ambiguity, our approach is as follows. If a word casts doubt on a patient's statement, it is considered either a judgment word or evidential, even if it may suggest noncompliance or emotion. If a word rephrases a patient's statement as hearsay but also carries meaning that explicitly casts doubt, it is considered a judgment word.

By analyzing the ways the keywords occur syntactically and leveraging constituency parsing, we create a grammar structure (see Figure 6.5). Our grammar effectively addresses several types of sentence structures. We not only account for several variations of sentences, but we also construct the grammar structure such that any clausal occurrences of the stigmatizing keywords are considered, and the rest of the sentence is retained. By harnessing the power of constituency parsing, we have allowed for edits that take larger syntactical aspects of sentences into account. For example, "The patient insists vehemently that she still experiences pain." corrects to "She still experiences pain." despite the presence of the adverb "vehemently" since the grammar

structure simply searches for the complete sentence (S) contained within the subordinating clause "that she still experiences pain." Our goal with all the rules is to retain as much information as possible while also eliminating stigma. Our corrections maintain a neutral tone, such as the rule that converts from 'refuses' to 'does not want.' The following table illustrates the complexity of sentences that can be corrected by the model (see Table 8.1).

Original Sentence	**Processed Sentence**
The medication prescribed to the patient has had positive effects on his blood pressure, but after three weeks, patient **complains** of chest pain.	The medication prescribed to the patient has had positive effects on his blood pressure, but after three weeks, patient has chest pain.
Patient is here for follow up after operation, where we see progress in vitals although she **argues** that she still feels a numbing sensation in her extremities.	Patient is here for follow up after operation, where we see progress in vitals although she still feels a numbing sensation in her extremities.
Patient also found to have a potassium of 2.7 and in clinic **reported** recent stomach issues, though on the floor she **says** her bowel movements have been normal, non-bloody, no diarrhea/constipation.	Patient also found to have a potassium of 2.7 and in clinic had recent stomach issues, though on the floor her bowel movements have been normal, non-bloody, no diarrhea/constipation.
Patient had previously been on CPAP, but due to his dementia, his agitation worsens with the CPAP mask and he **refuses** to wear the mask.	Patient had previously been on CPAP, but due to his dementia, his agitation worsens with the CPAP mask and he does not want to wear the mask.

Table 8.1: List of example sentences with complex structure and the output of our model.

While we do our best to account for all occurrences of stigmatizing language, there could be cases where our grammar structure does not perfectly edit notes. For instance, a sentence like "She claims God will cure her." corrects to "God will cure her." Clearly, this is not an appropriate edit of the note. As such, clinicians using our model should use their judgement to determine whether edits are appropriate or not. Future work on this project can focus on identifying incorrect edits and expanding on our grammar structure.

8.3 Evaluation Discussion

In this section, we present and discuss the results of the evaluation for the model introduced in Chapter 7. We decided to standardize raters' scores through z-score normalization due to potential differences in scoring perception amongst raters [51]. Some raters may be more conservative in their scoring approach than others. Given that our goal is to compare the scores between original statements and processed statements, normalizing scores for each rater does not affect this premise, as we are simply controlling for rater scoring perception. This standardization allows for us to assume independence between scores and rater identity. We then calculated the mean score and standard deviation for both the original statements and processed statements. To determine whether the difference in mean scores is statistically significant, we employed a paired two-sample t-test. A paired two-sample t-test assesses whether two samples of related units have the same mean [52]. Here, the two samples of related units are the original statements and processed statements. The paired t-test, as opposed to the unpaired t-test, is used when samples are related. In our case, the original statements and processed statements are related, as they both are groups of statements about patients in a medical setting. The processed statements originate from notes that have sentences that are semantically similar to the original statements presented to raters. Moreover, the paired t-test allows for the two samples to have differing variances, which applies to the distributions of scores for original statements and processed statements.

To ensure the validity of the paired t-test, we carefully considered the following criteria of the test:

1. The dependent variable is normally distributed
2. The observations are sampled independently
3. The dependent variable is continuous

4. The dependent variable should not contain any outliers.

For criteria 1, the Shapiro-Wilk normality test was used on the score distributions of original statements and processed statements [53]. The null-hypothesis of this test is that the data is normally distributed. Thus, if the p-value is less than the chosen alpha level, then the null hypothesis is rejected and there is evidence that the data is not normally distributed. On the other hand, if the p-value is greater than the chosen alpha level, then we fail to reject the null, and the data could be normally distributed. We obtained a two-sided p-value of $p = 0.257$ for the original statement scores and a two-sided p-value of $p = 0.213$ for the processed statement scores. Given an alpha value of 0.05, we fail to reject normality for both distributions. Although the Shapiro-Wilk normality test does not prove normality, the test does not suggest that the score distributions are non-normal, so we proceed with the assumption that the dependent variable is normally distributed.

We assume that the statements can be thought of as independent observations. While raters may compare between statements when assigning scores, we do not think these potential comparisons would cause high levels of dependence between statements. Additionally, raters were not allowed to collaborate in assigning scores, so we do not assume any dependence between rater scores.

The original scores were on an integer scale from 1 to 5, making the scores discrete variables, but with the z-score normalization of scores for each rater, the scores can now be considered a continuous dependent variable.

Finally, to assess whether there were any outliers present, the Grubbs test was used [54]. The Grubbs test is used to detect a single outlier in a univariate data set that follows an approximately normal distribution. The null-hypothesis of this test is that there are no outliers. We obtained a two-sided p-value of $p = 0.462$ for the original statement scores and a two-sided p-value of $p = 0.613$ for the processed statement scores. Given an alpha value of 0.05, we fail to reject that there are no

outliers. Additionally, we inspect the score distributions and note that there are no scores that lie outside of the three standard deviation range, an empirical rule applied to normal distributions where outliers are defined as points that lie more than three standard deviations from the mean.

Since all the criteria are fulfilled, we conducted a paired t-test on the original notes and processed. For even deeper analysis, we also stratified the notes according to the class of keywords they correspond to and performed three additional paired t-tests. According to the results of the t-tests, our model produced statistically significantly less biased notes than the original notes overall and also for all specific classes of keywords we have identified.

Rater agreement is an important statistic in assessing the validity of our evaluation. Amongst the 20 independent raters, there is moderate agreement overall, with a Fleiss' Kappa score of $\kappa = 0.564$. According to Fleiss et al., scores between 0.400 and 0.750 represent fair to good agreement beyond chance [55].

Upon taking a closer look at the stratified results, amongst the original notes, the ratings for the descriptors of noncompliance and emotion have the least variance and the highest mean scores. This finding is not surprising, as they are the least subtle indicators of stigmatizing language, especially the adjectival keywords. Below are some of the statements the raters viewed from the original notes that contained keywords from classes 3 and 4 (see Appendix C for complete list):

1. The patient is cranky and uncooperative with staff.

2. He remains agitated despite gentle care.

3. She is uncooperative and continues yelling loudly at me.

4. The patient is dramatic when it comes to needles.

Across all classes, the ratings of processed notes have less variation compared to that of the original notes. This is to be expected, as the stigmatizing keywords that

are removed from the original notes behave as trigger words that induce a response in some raters and not have an effect on others, thereby increasing the variance in scores given. The processed notes, however, have much more plain language, including statements, such as

1. The patient has pain in her lower back.

2. She does not want additional scans.

3. Patient had an allergic reaction to previously prescribed analgesic.

Similarly, amongst the agreement scores, we see greater levels of agreement for processed notes across all classes compared to the original notes. There is substantial rater agreement for processed notes scores overall and for all classes, based on the interpretation of agreement strength by Landis et al. [1]. For the processed notes, the agreement scores are pretty similar across all classes, with evidentials having the highest agreement score. For the original notes, the agreement score is highest for descriptors of noncompliance.

We must note that for class 4, the original notes and processed notes do not correspond to the same set of keywords. We consider original notes containing stigmatizing keywords that are both adjectives and verbs, whereas we only consider processed notes corresponding to verbs ('refuses' and 'resists'). For class 4, we do not have any processed notes since the grammar structure we have developed simply removes the descriptors of non-compliance. Nonetheless, raters still found class 3 and class 4 statements to contain stigmatizing language, and it is reasonable to assume that they would find these statements more stigmatizing than the lack of any statement. Ultimately, the statements processed by our model MedStiLE are found to be significantly less stigmatizing than the original statements.

Chapter 9

Conclusion

This paper approached the problem of developing a model that edits patient medical notes to no longer contain stigmatizing language. Previous research regarding stigmatizing language in medical notes has mostly been qualitative, as work has centered on compiling keywords manually from analyzing medical notes. To obtain a more robust set of stigmatizing keywords, we created contextual word embeddings from BERT-based and BioBERT-based models that are trained on free-text patient-oriented clinical data. These state-of-the-art models allowed us to create word vector representations based on patient-oriented clinical text. From there, we identified 30 new stigmatizing keywords by calculating similar word vectors to those of the 40 previously-identified keywords. After compiling a total list of 70 keywords, we complete a thorough analysis to categorize the keywords according to the ways they induce stigma and better understand the syntactical environments in which these keywords occur. We then develop our model MedStiLE to edit notes containing the stigmatizing keywords to be non-stigmatizing by utilizing a grammar structure and constituency parsing. MedStiLE now automatically edits notes from stigmatizing to non-stigmatizing, and we conducted an evaluation to test our model's efficacy using human raters.

Many aspects of this project are promising. First, the development of our BERT-based and BioBERT-based models on free-text patient-oriented clinical data can be used not just for identifying semantic similarity between words in a patient-based medical context but also for a variety of tasks such as medical annotation and inference. Our models, which we have named PODS BERT and Bio+PODS BERT, are the first, that we know of, to be trained specifically on free-text designed to describe patient behavior, experience, or health in addition to medical terminology. Second, we have identified 30 new stigmatizing keywords through a completely computational procedure. Previous compilations of stigmatizing keywords have been entirely qualitative, and it is clear that a qualitative searching methodology is inadequate in obtaining a comprehensive list of stigmatizing keywords. We believe that further use of our methodology would lead to the discovery of even more stigmatizing keywords. Third, our thorough examination of the syntactical contexts where the stigmatizing keywords occur provide several new insights on the nature of stigmatizing language in medical notes. Fourth, our model MedStiLE effectively edits notes to become non-stigmatizing. Overall, we believe that the goal of our work is novel and can act as a precursor to future models that seek to edit stigmatizing language in medical notes. Ultimately, we feel that correcting stigmatizing language in medical notes is a necessary step towards equitable healthcare.

9.1 Future Work

Our model can be expanded in various ways. Most feasibly, additional keywords can be found through the word embeddings of PODS BERT and Bio+PODS BERT and incorporated into the grammar structure. A more thorough analysis can be done on outlying syntactical contexts that are not properly considered by our grammar structure. Our model is entirely built on data from the Beth Israel Deaconess Medical

Center (BIDMC). Because our data comes from critical care units and primarily emergency procedures, our findings cannot be applied to smaller private medical practices. More generally, differences in care practices across institutions can be significant, making using notes from multiple institutions important. While there is a shortage of publicly available clinical data, we hope that there will be more access to shared clinical text in the future that will enable greater levels of generalizability and applicability to various medical settings. There are several directions to pursue in developing models to edit stigmatizing language. We hope that our methods and results will influence and inform models and related projects of future researchers.

Appendix A

Code

The code used for this project can be found at the following Google Drive link: https://drive.google.com/drive/folders/1xRXJgUbg5J9X9mc-k-zgfEb0lbK1Rdml?usp=share_link.

Please note that access to the MIMIC-IV-Note, MedNLI, and 2010 i2b2/VA challenge data is granted to only credentialed users. To apply for credentialed status for the MIMIC-IV-Note database and MedNLI data, see this: https://physionet.org/. To obtain access to the 2010 i2b2/VA challenge data, see this: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/.

Appendix B

Complete Results from Nearest Neighbors Task

For completion, we include the complete results from the nearest neighbors task in Table B.1. These results show how both PODS BERT and Bio+PODS BERT have performed. The table also shows the relationship between the new keywords and previously-identified keywords, as in which previously-identified keywords the new keywords are closest to. We can also see the various previously-identified keywords that appeared again through this task. We imagine that the repetition of previously-identified keywords in this task may be elevated by the way that we filtered our notes to only those that included previously-identified words.

	PODS BERT	Bio+PODS BERT
'claims'	'claimed'	'claimed'
	'claiming'	'claiming'
	'believes'	'tells'
	'argues'	'says'
	'says'	'believes'
'insists'	'convinced'	'insisted'
	'claims'	'convince'
	'demands'	'claims'
	'says'	'says'
	'insisted'	'claimed'
'proclaims'	'claims'	'claims'
	'swears'	'swears'
	'promises'	'says'
	'complains'	'states'
	'says'	'insists'
'asserts'	'argue'	'assert'
	'assert'	'declares'
	'denies'	'denies'
	'deny'	'insists'
	'patient'	'claims'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'protests'	'protest'	'claims'
	'claims'	'protest'
	'claimed'	'declares'
	'states'	'denies'
	'angry'	'deny'
'swears'	'swear'	'swear'
	'promises'	'promises'
	'promise'	'resists'
	'resists'	'denies'
	'denies'	'deny'
'attests'	'suggests'	'claims'
	'reports'	'states'
	'patient'	'suggests'
	'time'	'reports'
	'discharge'	'says'
'alleges'	'alleged'	'alleged'
	'swears'	'swears'
	'promises'	'promises'
	'resists'	'denies'
	'denies'	'resists'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'complains'	'complaint'	'complaint'
	'chief'	'concerned'
	'concerns'	'concerns'
	'frustrated'	'claims'
	'dislike'	'frustrated'
'denies'	'denied'	'denied'
	'insists'	'deny'
	'claims'	'claims'
	'refuses'	'refuses'
	'refused'	'failed'
'endorses'	'rejects'	'endorsed'
	'urge'	'urges'
	'endorsed'	'says'
	'doctor'	'doctor'
	'likes'	'likes'
'notes'	'comments'	'comments'
	'reports'	'says'
	'says'	'reports'
	'remarks'	'claims'
	'tells'	'tells'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'reports'	'reported'	'reported'
	'tells'	'says'
	'says'	'tells'
	'alleges'	'claims'
	'claims'	'alleges'
'says'	'said'	'said'
	'tells'	'tells'
	'told'	'say'
	'states'	'states'
	'explains'	'explains'
'tells'	'told'	'told'
	'says'	'says'
	'said'	'said'
	'states'	'states'
	'explains'	'say'
'nonadherent'	'negative'	'negative'
	'pessimistic'	'rude'
	'agitated'	'agitated'
	'cranky'	'refuses'
	'resistant'	'resists'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'aggressive'	'angry'	'negative'
	'forceful'	'fast'
	'impulsive'	'angry'
	'confrontational'	'rude'
	'negative'	'frustrated'
'agitated'	'restless'	'restless'
	'angry'	'angry'
	'negative'	'negative'
	'aggressive'	'aggressive'
	'cranky'	'cranky'
'angry'	'frustrated'	'frustrated'
	'worried'	'worried'
	'negative'	'annoyed'
	'annoyed'	'resistant'
	'resistant'	'mad'
'challenging'	'difficult'	'difficult'
	'frustrating'	'hard'
	'noncompliant'	'restless'
	'restless'	'agitated'
	'angry'	'angry'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'combative'	'resistant'	'resistant'
	'noncompliant'	'noncompliant'
	'confrontational'	'confrontational'
	'unpleasant'	'unpleasant'
	'angry'	'angry'
'noncompliant'	'resistant'	'difficult'
	'confrontational'	'stubborn'
	'unpleasant'	'restless'
	'angry'	'angry'
	'combative'	'combative'
'confront'	'ask'	'question'
	'question'	'deny'
	'resist'	'refuse'
	'refuse'	'confrontational'
	'confrontational'	'said'
'noncooperative'	'uncooperative'	'difficult'
	'resistant'	'resistant'
	'noncompliant'	'noncompliant'
	'different'	'uncooperative'
	'unyielding'	'refuse'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'defensive'	'negative'	'combative'
	'refuse'	'negative'
	'distant'	'angry'
	'angry'	'resistant'
	'resistant'	'resists'
'exaggerate'	'perceive'	'dramatic'
	'dramatic'	'unsure'
	'confused'	'confused'
	'claims'	'claims'
	'claim'	'claimed'
'hysterical'	'delusional'	'hysteria'
	'paranoid'	'delusions'
	'delirious'	'delusional'
	'strange'	'delirious'
	'hysteria'	'worries'
'unpleasant'	'painful'	'bad'
	'horrible'	'rude'
	'severe'	'severe'
	'bad'	'refuses'
	'intense'	'combative'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'refuses'	'refusal'	'refusal'
	'denies'	'denies'
	'denied'	'denied'
	'reluctant'	'reluctant'
	'resists'	'resists'
'resists'	'resistant'	'resistant'
	'counters'	'combative'
	'denies'	'denied'
	'denied'	'complains'
	'complains'	'denies'
'concerned'	'concern'	'worried'
	'concerns'	'concern'
	'worried'	'worries'
	'worries'	'concerns'
	'upset'	'sad'
'confused'	'confuse'	'confuse'
	'concerned'	'worry'
	'disturb'	'concern'
	'unsettled'	'panicked'
	'delirious'	'strange'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'delirious'	'hysterical'	'hysterical'
	'restless'	'unsettled'
	'unsettled'	'hallucination'
	'hallucination'	'manic'
	'manic'	'restless'
'dramatic'	'melodramatic'	'melodramatic'
	'frantic'	'frantic'
	'stressed'	'stressed'
	'troubled'	'troubled'
	'worried'	'worried'
'frantic'	'frenzy'	'frenzy'
	'worried'	'worried'
	'stressed'	'stressed'
	'upset'	'upset'
	'hysterical'	'hysterical'
'troubled'	'worried'	'trouble'
	'concerned'	'worried'
	'confused'	'confused'
	'emotional'	'emotional'
	'worries'	'concerned'

Continued on Next Page

	PODS BERT	Bio+PODS BERT
'emotional'	'upset'	'mental'
	'sad'	'crying'
	'anxious'	'hysterical'
	'depressed'	'manic'
	'bipolar'	'confused'
'worried'	'concerned'	'worries'
	'worry'	'worry'
	'concern'	'concern'
	'panicked'	'concerned'
	'odd'	'upset'
'upset'	'unhappy'	'unhappy'
	'worried'	'sad'
	'worry'	'worried'
	'sad'	'depressed'
	'depressed'	'distressed'
'remarks'	'remark'	'remark'
	'says'	'says'
	'tells'	'said'
	'states'	'states'
	'explains'	'notes'

Table B.1: Complete results form nearest neighbors task from PODS BERT and Bio+PODS BERT where the five identified words are listed from highest cosine similarity (most similar) to lowest cosine similarity (least similar)

Appendix C

Evaluation Details

This section details the evaluation guide provided to the raters for judging the extent of stigma in original statements from medical notes and statements processed by MedSTiLE. This guide served as a short assignment for Princeton's Introduction to Machine Translation course (TRA 301/COS 401) in the Spring 2023 semester. We specifically do not provide raters with examples of stigmatizing language, as to not skew them into selecting statements according to keywords and instead exercise their own judgment in determining if a statement carries stigma.

In section C.2, we list the statements as provided to raters. Sections C.1 and C.2 consist of the exact guide given to raters. More specifically, section C.2 includes the order of the statements given to the raters. Section C.3 details the actual categorization of notes into four classes where class 1 is judgment words, class 2 is evidentials, class 3 is descriptors of noncompliance, and class 4 is descriptors of emotion, to provide more context to our findings in chapters 7 and 8. We also discuss the original statements from which the processed notes were edited. The contents of section C.3 were not shown to raters.

C.1 Introduction and Instructions

Medical professionals can have biases against certain groups of individuals. Numerous studies have shown how patients may receive poorer quality of care based on their gender, race/ethnicity, level of health literacy, and many other factors. Stigmatizing language written in a patient's medical record can perpetuate negative attitudes and influence decision-making of clinicians subsequently caring for that patient, thereby hindering a patient's ability to obtain insurance, receive the medical care they need, and recover effectively. Thus, biased language in medical records must be corrected.

Stigmatizing language can be any of the following:

- Portraying a patient negatively in a way that is irrelevant to medical care or diagnosis
- Emphasizing conflict or tensions between patient and clinician(s)
- Casting doubt on a patient's experiences
- Placing blame on a patient
- Stereotyping a patient (racism, sexism, xenophobia, homophobia, etc.)

We have developed a model that edits notes to remove stigmatizing language. The following survey consists of synthetically-generated statements based on patient medical notes from the MIMIC-IV-Note clinical database. All sensitive patient information has either been removed or modified for purposes of this survey. Some of the statements have been processed by the model while others have not. This survey is meant to evaluate how effective the model is in correcting stigmatizing language.

Please evaluate each of the following statements from level 1 (no stigmatizing language) to level 5 (strongly stigmatizing language). For clarity, level 3 would represent some stigmatizing language.

C.2 List of Statements

Table C.1 is the list of statements in the same order they were presented to the raters.

Statements for Model Evaluation
He claims to be experiencing pain three weeks post operation.
The patient is cranky and uncooperative with staff.
The patient does not want to continue treatment.
She became hysterical at the idea of sharing her room.
She complains of blurry vision in both eyes.
He is restless, inattentive, and continues to ignore my questions.
She does not want additional scans.
Patient had an allergic reaction to previously prescribed analgesic.
Patient insists that the wound is still open.
He does not want to take additional medications, specifically blood thinners.
The patient is dramatic when it comes to needles.
The patient does not want to sleep for longer periods of time due to stress.
Over the past month, her symptoms, including upper extremity weakness and numbness, have worsened.
He remains agitated despite gentle care.
She has been coughing for the past few days, occasionally producing phlegm.
He states that he does not have any refills at home.
She does not want to undergo major surgery at this time.
Patient refuses to tell me about the cause of the accident.
She does not want to participate in most social activities.
He is frantic and lacks stability.
She swears to having taken the prescribed dosage.

Continued on Next Page

Statements for Model Evaluation
She has been taking in good fluid hydration.
The patient does not want to share more details about the accident.
She is uncooperative and continues yelling loudly at me.
Patient attests that chest pain is persistent.
She has been experiencing worsening depression with increasing thoughts of suicide over the last couple of weeks.
In addition, when enquired patient remarks that he does not use drugs or alcohol.
She panicked immediately, and then drove herself to the emergency department.
The patient does not want a pelvic exam presently.
She resists disclosing information about her sexual past.
Patient checks his heart rate daily.
He argues that he wants additional scans of his lungs.
Patient has been depressed since the age of 18.
He does not want mental health services due to financial barriers.
Medication was giving her vivid dreams and leading to poor quality sleep.
She notes that she had some photophobia so felt that her symptoms were most consistent with her chronic migraines.
Patient does not want a colonoscopy
Patient is regularly taking medication as prescribed by her previous doctor.
He denies having multiple sexual partners.
She has never before had a small bowel obstruction.

Table C.1: List of statements provided to raters

C.3 Classification of Statements

The following tables show how the statements of the evaluation guide correspond to the four classes of stigmatizing keywords where class 1 is judgment words, class 2 is evidentials, class 3 is descriptors of noncompliance, and class 4 is descriptors of emotion. We also show the original statements from which the processed statements were edited.

Original Statements (Class 1)
He claims to be experiencing pain three weeks post operation.
Patient insists that the wound is still open.
She swears to having taken the prescribed dosage.
Patient attests that chest pain is persistent.
He argues that he wants additional scans of his lungs.

Table C.2: List of original statements provided to raters that contain judgment words (class 1 of stigmatizing keywords)

Processed Statements (Class 1)	
The patient claims to have pain in her lower back.	The patient has pain in her lower back.
Patient asserts that he had an allergic reaction to previously prescribed analgesic.	Patient had an allergic reaction to previously prescribed analgesic.
Over the past month, she argues her symptoms, including upper extremity weakness and numbness, have worsened.	Over the past month, her symptoms, including upper extremity weakness and numbness, have worsened.
She claims she has been taking in good fluid hydration.	She takes in good fluid hydration.
Patient claims he has been depressed since the age of 18.	Patient has been depressed since the age of 18.

Table C.3: List of processed statements provided to raters that contain judgment words (class 1 of stigmatizing keywords). We also include the original statements from which the processed statements are edited in the left column for reference.

Original Statements (Class 2)
She complains of blurry vision in both eyes.
He states that he does not have any refills at home.
In addition, when enquired patient remarks that he does not use drugs or alcohol.
She notes that she had some photophobia so felt that her symptoms were most consistent with her chronic migraines.
He denies having multiple sexual partners.

Table C.4: List of original statements provided to raters that contain evidentials (class 2 of stigmatizing keywords)

Processed Statements (Class 2)	
Patient notes checking his heart rate daily.	Patient checks his heart rate daily.
Patient states that medication was giving her vivid dreams and leading to poor quality sleep.	Medication was giving her vivid dreams and leading to poor quality sleep.
She explains she has been coughing for the past few days, occasionally producing phlegm.	She has been coughing for the past few days, occasionally producing phlegm.
Patient reports she is regularly taking medication as prescribed by her previous doctor.	Patient is regularly taking medication as prescribed by her previous doctor.
She tell me she has never before had a small bowel obstruction	She has never before had a small bowel obstruction

Table C.5: List of processed statements provided to raters that contain evidentials (class 2 of stigmatizing keywords). We also include the original statements from which the processed statements are edited in the left column for reference.

Original Statements (Class 3)
The patient is cranky and uncooperative with staff.
He remains agitated despite gentle care.
Patient refuses to tell me about the cause of the accident.
She is uncooperative and continues yelling loudly at me.
She resists disclosing information about her sexual past.

Table C.6: List of original statements provided to raters that contain descriptors of noncompliance (class 3 of stigmatizing keywords)

Original Statements (Class 4)
He is restless, inattentive, and continues to ignore my questions.
The patient is dramatic when it comes to needles.
He is frantic and lacks stability.
She panicked immediately, and then drove herself to the emergency department.
She became hysterical at the idea of sharing her room.

Table C.7: List of original statements provided to raters that contain descriptors of emotion (class 4 of stigmatizing keywords)

Processed Statements (Class 3)	
Patient notes checking his heart rate daily.	Patient checks his heart rate daily.
Patient states that medication was giving her vivid dreams and leading to poor quality sleep.	Medication was giving her vivid dreams and leading to poor quality sleep.
She explains she has been coughing for the past few days, occasionally producing phlegm.	She has been coughing for the past few days, occasionally producing phlegm.
Patient reports she is regularly taking medication as prescribed by her previous doctor.	Patient is regularly taking medication as prescribed by her previous doctor.
She tell me she has never before had a small bowel obstruction	She has never before had a small bowel obstruction

Table C.8: List of processed statements provided to raters that contain evidentials (class 2 of stigmatizing keywords). We also include the original statements from which the processed statements are edited in the left column for reference.

Bibliography

[1] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, March 1977.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[3] Elise Paradis and Cynthia R Whitehead. Louder than words: power and conflict in interprofessional education articles, 1954–2013. *Medical Education*, 49(4):399–407, April 2015.

[4] Laura Nimmon and Terese Stenfors-Hayes. The "Handling" of power in the physician-patient encounter: perceptions from experienced physicians. *BMC Medical Education*, 16(1):114, December 2016.

[5] Brian B. Drwecki, Colleen F. Moore, Sandra E. Ward, and Kenneth M. Prkachin. Reducing racial disparities in pain treatment: The role of empathy and perspective-taking. *Pain*, 152(5):1001–1006, 2011.

[6] Carmen R. Green, Karen O. Anderson, Tamara A. Baker, Lisa C. Campbell, Sheila Decker, Roger B. Fillingim, Donna A. Kaloukalani, Kathyrn E. Lasch, Cynthia Myers, Raymond C. Tait, Knox H. Todd, and April H. Vallerand. The Unequal Burden of Pain: Confronting Racial and Ethnic Disparities in Pain. *Pain Medicine*, 4(3):277–294, September 2003.

[7] P. Adam Kelly and Paul Haidet. Physician overestimation of patient literacy: A potential source of health care disparities. *Patient Education and Counseling*, 66(1):119–122, April 2007.

[8] Elizabeth N. Chapman, Anna Kaatz, and Molly Carnes. Physicians and Implicit Bias: How Doctors May Unwittingly Perpetuate Health Care Disparities. *Journal of General Internal Medicine*, 28(11):1504–1510, November 2013.

[9] Abinitha Gourabathina. Detecting stigmatizing language in patient medical records through a data-driven approach. Princeton University, Junior Independent Work, May 2022.

[10] Lynne M. Kirk. Professionalism in Medicine: Definitions and Considerations for Teaching. *Baylor University Medical Center Proceedings*, 20(1):13–16, January 2007.

[11] Havi Carel and Ian James Kidd. Epistemic injustice in healthcare: a philosophial analysis. *Medicine, Health Care and Philosophy*, 17(4):529–540, November 2014.

[12] Paige L. Sweet. The Sociology of Gaslighting. *American Sociological Review*, 84(5):851–875, October 2019.

[13] Elena Ruíz. Cultural Gaslighting. *Hypatia*, 35(4):687–713, 2020.

[14] Jennifer C. H. Sebring. Towards a sociological understanding of medical gaslighting in western health care. *Sociology of Health & Illness*, 43(9):1951–1964, November 2021.

[15] Cecilia Tasca, Mariangela Rapetti, Mauro Giovanni Carta, and Bianca Fadda. Women And Hysteria In The History Of Mental Health. *Clinical Practice & Epidemiology in Mental Health*, 8(1):110–119, October 2012.

[16] Margrit Shildrick. *Leaky bodies and boundaries: feminism, postmodernism and (bio)ethics*. Routledge, London ; New York, 1997.

[17] Jon Stone, Roger Smyth, Alan Carson, Steff Lewis, Robin Prescott, Charles Warlow, and Michael Sharpe. Systematic review of misdiagnosis of conversion symptoms and "hysteria". *BMJ*, 331(7523):989, October 2005.

[18] Esther H. Chen, Frances S. Shofer, Anthony J. Dean, Judd E. Hollander, William G. Baxt, Jennifer L. Robey, Keara L. Sease, and Angela M. Mills. Gender Disparity in Analgesic Treatment of Emergency Department Patients with Acute Abdominal Pain. *Academic Emergency Medicine*, 15(5):414–418, May 2008.

[19] Nafees U. Din, Obioha C. Ukoumunne, Greg Rubin, William Hamilton, Ben Carter, Sal Stapley, and Richard D. Neal. Age and Gender Variations in Cancer Diagnostic Intervals in 15 Cancers: Analysis of Data from the UK Clinical Practice Research Datalink. *PLOS ONE*, 10(5):e0127717, May 2015.

[20] A.H.E.M. Maas and Y.E.A. Appelman. Gender differences in coronary heart disease. *Netherlands Heart Journal*, 18(12):598–603, November 2010.

[21] Ana Mikolić, David van Klaveren, Joost Oude Groeniger, and Wiegers. Differences between Men and Women in Treatment and Outcome after Traumatic Brain Injury. *Journal of Neurotrauma*, page neu.2020.7228, October 2020.

[22] J. D. Guillory. The pro-slavery arguments of Dr. Samuel A. Cartwright. *Louisiana History*, 9:209–227, 1968.

[23] Rana Asali Hogarth. The Myth of Innate Racial Differences Between White and Black People's Bodies: Lessons From the 1793 Yellow Fever Epidemic in Philadelphia, Pennsylvania. *American Journal of Public Health*, 109(10):1339–1341, October 2019.

[24] Kelly M. Hoffman, Sophie Trawalter, Jordan R. Axt, and M. Norman Oliver. 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, April 2016.

[25] Janice A. Sabin and Anthony G. Greenwald. The Influence of Implicit Bias on Treatment Recommendations for 4 Common Pediatric Conditions: Pain, Urinary Tract Infection, Attention Deficit Hyperactivity Disorder, and Asthma. *American Journal of Public Health*, 102(5):988–995, May 2012.

[26] Salimah H. Meghani, Eeeseung Byun, and Rollin M. Gallagher. Time to Take Stock: A Meta-Analysis and Systematic Review of Analgesic Treatment Disparities for Pain in the United States. *Pain Medicine*, 13(2):150–174, February 2012.

[27] Nicole A. Hollingshead, Leslie Ashburn-Nardo, Jesse C. Stewart, and Adam T. Hirsh. The Pain Experience of Hispanic Americans: A Critical Literature Review and Conceptual Model. *The Journal of Pain*, 17(5):513–528, May 2016.

[28] Alexie Cintron and R. Sean Morrison. Pain and Ethnicity in the United States: A Systematic Review. *Journal of Palliative Medicine*, 9(6):1454–1473, December 2006.

[29] Jeffrey Glassberg, Paula Tanabe, Lynne Richardson, and Michael DeBaun. Among emergency physicians, use of the term "Sickler" is associated with negative attitudes toward people with sickle cell disease. *American Journal of Hematology*, 88(6):532–533, June 2013.

[30] John F. Kelly and Cassandra M. Westerhoff. Does it matter how we refer to individuals with substance-related conditions? A randomized study of two com-

monly used terms. *International Journal of Drug Policy*, 21(3):202–207, May 2010.

[31] Anna P. Goddu, Katie J. O'Conor, Sophie Lanzkron, Mustapha O. Saheed, Somnath Saha, Monica E. Peek, Carlton Haywood, and Mary Catherine Beach. Do Words Matter? Stigmatizing Language and the Transmission of Bias in the Medical Record. *Journal of General Internal Medicine*, 33(5):685–691, May 2018.

[32] Jenny Park, Somnath Saha, Brant Chee, Janiece Taylor, and Mary Catherine Beach. Physician Use of Stigmatizing Language in Patient Medical Records. *JAMA Network Open*, 4(7):e2117052, July 2021.

[33] Mary Catherine Beach, Somnath Saha, Jenny Park, Janiece Taylor, Paul Drew, Eve Plank, Lisa A. Cooper, and Brant Chee. Testimonial Injustice: Linguistic Bias in the Medical Records of Black Patients and Women. *Journal of General Internal Medicine*, 36(6):1708–1714, June 2021.

[34] W. F. Hertzog. A CASE OF PUERPERAL ECLAMPSIA. *Journal of the American Medical Association*, I(7):220–221, August 1883. _eprint: https://jamanetwork.com/journals/jama/articlepdf/421202/jama_i_7_009.pdf.

[35] R. D. BARKER. A CASE OF TYPHLITIS, WITH AUTOPSY. *Journal of the American Medical Association*, I(9):273–274, September 1883. _eprint: https://jamanetwork.com/journals/jama/articlepdf/421255/jama_i_9_001g.pdf.

[36] David B. Sykes and Darren N. Nichols. There Is No Denying It, Our Medical Language Needs an Update. *Journal of Graduate Medical Education*, 7(1):137–138, March 2015.

[37] Michael Sun, Tomasz Oliwa, Monica E. Peek, and Elizabeth L. Tung. 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, February 2022.

[38] Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. MIMIC-IV-Note: Deidentified free-text clinical notes. Version Number: 2.2 Type: dataset.

[39] Alistair Johnson, Tom Pollard, and Roger Mark. MIMIC-III Clinical Database Demo, 2019. Type: dataset.

[40] Berkeley Franz and John W. Murphy. Reconsidering the role of language in medicine. *Philosophy, Ethics, and Humanities in Medicine*, 13(1):5, December 2018.

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.

[42] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[43] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[44] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations, 2018.

[45] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, sep 2019.

[46] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings, 2019.

[47] Chaitanya Shivade. MedNLI — A Natural Language Inference Dataset For The Clinical Domain, 2017. Type: dataset.

[48] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, June 2011. _eprint: https://academic.oup.com/jamia/article-pdf/18/5/552/33015280/supplemental_materials_amiajnl-2011-000203.pdf.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[51] Rahul Wadbude, Vivek Gupta, Dheeraj Mekala, and Harish Karnick. User bias removal in review score prediction. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, CODS-COMAD

'18, page 175–179, New York, NY, USA, 2018. Association for Computing Machinery.

[52] George Waddel Snedecor and William Gemmell Cochran. *Statistical methods*. Iowa State University Press, Ames, 8th ed edition, 1989.

[53] Prabhaker Mishra, Chandra M. Pandey, Uttam Singh, Anshul Gupta, Chinmoy Sahu, and Amit Keshri. Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, 22(1):67–72, 2019.

[54] Frank E. Grubbs. Procedures for detecting outlying observations in samples. *Technometrics*, 11(1):1–21, 1969.

[55] *The Measurement of Interrater Agreement*, chapter 18, pages 598–626. John Wiley Sons, Ltd, 2003.