UNIVERSITY OF PRINCE EDWARD ISLAND

Identifying Disability Insensitive Language in Scholarly Works using Machine Learning

by

Roshna Roby

A thesis submitted in partial fulfillment for the
degree of Honours in Computer Science

in the

Faculty of Science
School of Mathematical and Computational Sciences

April 2025

Declaration of Authorship

I, Roshna Roby, declare that this thesis titled, 'Identifying Disability Insensitive Language in Scholarly Works using Machine Learning' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UNIVERSITY OF PRINCE EDWARD ISLAND

Abstract

Faculty of Science
School of Mathematical and Computational Sciences

Honours in Computer Science

by Roshna Roby

Language plays an important role in shaping societal views of disability. Over time, the terminology used to describe disability has evolved to be more inclusive; however, instances of insensitive or outdated language persist in written communication. This thesis studies the presence of insensitive disability-related language in the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS) conference paper abstracts from 1994 to 2024 as a starting point to build a system that can automatically detect such language within scholarly writing.

Guidelines around disability language from the UN, ADA, and the Accessible Canada Act were used to build a set of terms/phrases marked as insensitive. This set of lexicons was used to build a corpus of sentences within ASSETS abstracts. A BERT-base model was fine-tuned on the resulting corpus, achieving 92.66% test accuracy. However, it was identified that the corpus had a disproportionate distribution in the number of examples available within each category, i.e., “insensitive” or “notInsensitive” for each term from the list of lexicons. Therefore, GPT-4o-generated data was incorporated to address this disproportion, resulting in an augmented corpus. The BERT model fine-tuned on the augmented corpus achieved the best overall performance with a test accuracy of 93.58% on source ASSETS abstract sentences and 96.28% accuracy on the augmented test set. This augmented model was able to generalize well to real-world source sentences better than all other models trained on the same corpora.

This study acts as a step towards understanding the use of inclusive disability terminology within academic writing and the feasibility of building a model solely to identify insensitivity around disability language.

Acknowledgements

I am especially thankful to my supervisors, Dr. Christopher Power and Dr. Paul Sheridan, for their unwavering support throughout the thesis and to Dr. Andrew Godbout for taking on the role of reader. Thanks to Yvonne Terry and Rodney Macleod for their mentorship on web accessibility and inclusive language, which greatly inspired this work. I would also like to thank my parents for being by my side throughout my degree.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language, being the foundation of communication, has the ability to both reinforce or dismantle stereotypes within human society. The composition of words and the human ability to understand specific linguistic rules differentiates humans from other species [1]. Therefore, rules around how a set of words should be stringed together to convey emotions and needs are what underline human language [1]. This leads to the evolution of language according to different guidelines in various contexts, including the ways in which society refers to disabilities. Over the years, terms or phrases used to refer to individuals with disabilities have evolved from being derogatory and dehumanizing to being more inclusive [2], fostering a shift in the social outlook towards people with disabilities. Inclusive language is also a clear indicator of the commitment to learning and changing marginalizing perspectives, cultural beliefs, and stereotypes [3] by organizations or even individuals alone; hence, it plays a very important role in propagating respect across communities.

This study proposes developing an automated system to detect insensitive language in academic writing. The goal is to potentially create a tool that supports scholars in using more inclusive language. To ensure relevance to terms referring to disability mainly in academic contexts, we analyze the terminology used in the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS) conference papers, specifically their abstracts, and compare them to more "inclusive" language specified by Disability Rights Organizations such as the Americans with Disabilities (ADA) National Network [4] and United Nations (UN) specified guidelines [5].

1.1 Research Goals

We study language around disability within scholarly content by using the sentences extracted from ASSETS conference papers based on a non-exhaustive and non-definitive set of terms

and phrases that are recognized to be outdated and insensitive to fine-tune Machine Learning models (ML) like BERT (Bidirectional Encoder Representations from Transformers) and logistic regression. Please note that insensitive and stereotypical terminology is being used within the study; however, the inclusion of such terms is for research and analysis purposes only. Insensitive language, as discussed throughout this paper, is studied and referenced only within the context of the English language.

The goals of this study are:

1. Compile a list of guidelines and lexicons around insensitivity in disability-related language, specifically within written communication.
2. Create an annotated corpus of sentences that use the terminology referred to by the consulted guidelines within the ASSETS conference papers over the years from the very first conference in 1994 to 2024, the latest one as per the date of this study.
3. Study the potential of using OpenAI's GPT-4o supported data augmentation where the corpora lacks sufficient examples.
4. Predict insensitivity in disability language used within academia using a fine-tuned BERT model.

We compile a non-exhaustive list of terms that may be insensitive based on the context as described by ten different guideline articles, i.e., the ground truth documents for this thesis (see Chapter 3). These terms are then used to generate a corpus that contains sentences extracted from the abstracts of all ASSETS conferences over the years. By manually compiling and annotating the corpus, we aim to study the examples of disability-related terminology specifically within the academic context in a transparent and traceable manner.

Given the binary classification nature of the study, i.e., treating each sentence within the corpus as either insensitive or not, the corpus was first used to train a logistic regression model, which acts as the baseline model for this study, and compare it with the metrics generated by fine-tuning the same records on a pre-trained BERT-base model. The compiled corpus was found to underrepresent certain terms and phrases; therefore, this thesis also describes some methods used to augment the data in order to balance out the number of sentences corresponding to each term. It compares the results from the initial models with those from the models trained on augmented data. The comparison includes a total of five models:

- Three BERT models - one trained on unaugmented data, another trained on synthetic data generated by GPT-4o through one-shot prompting, and finally, a model trained on augmented data.

- Two logistic regression models - one trained on unaugmentated data based on sentences extracted from ASSETS abstracts and another trained on augmented data to act as baseline models (see Section 5.1).

Future works could consider placing a focus on extending the initial lexicon list of 160 terms and phrases used to extract sentences from the abstracts. Expanding this list would mean an increased variety of terms can be captured during the extraction stage; such a dynamically extendable list could ensure that the model would later have more diverse data to learn from. The current ability of the augmented BERT model to generalize to different contexts can be improved significantly, which could ultimately be used as a foundation to build a more resilient and robust system that can identify insensitivity in written content across domains.

The rest of this study is set up as follows. Chapter 2 explains inclusive language, the different models of disability language, and some related works around toxic language detection using Natural Language Processing (NLP) techniques. Chapter 3 describes how the corpus was compiled and annotated according to guidelines on inclusive disability language. Chapter 4 explores the potential of fine-tuning/training on the compiled unaugmented corpus from Chapter 3 using machine learning models like BERT and logistic regression. Chapter 5 focuses on supplementing the unaugmented corpus using GPT-4o and evaluating the models trained on this new augmented corpus, while Chapter 6 compares these performance metrics and analyzes key patterns within the results produced by each model. Chapter 7 outlines all the key limitations of the study along with suggestions for future research within the domain. Chapter 8 summarizes the entire study and its impact on identifying inclusive disability language.

Chapter 2

Background Information

This chapter discusses the different models of disability language currently in use and the challenges around writing more inclusively. The chapter also explains related works done within the domain of toxic language detection and how that impacts the detection of insensitive language solely around disability.

2.1 Different Models of Disability Language

A study on the evolution of disability language [6] explains how some of the words that were historically used to dehumanize individuals with any form of a disability have originated from a moral model that considered disability a result of "moral failing", with some words like "cripple" being in use since the early 9th century until the 17th century before even being recognized as insensitive. Multiple models of disability have evolved since then. The two models discussed below are the most common across literature [6].

1. The medical model: This is a long-established model that often categorizes individuals in general by their clinical diagnosis viewing disability as something that deviates from the "normal" functioning of the human body and hence needs to be "prevented" or "corrected" i.e as an "impairment", for example, the term, "the spastic quadriplegic" [6, 7]. The medical model often manifests itself when individuals are segregated into separate "special" programs or facilities sold as a form of "deficit-based" alternatives [8]. This impairment-focused model fails to recognize that an individual may experience different forms of "impairments" in their lifetime while others may be part of one's identity [7].

2. The social model: This model was a result of disability activism and advocacy for an alternate model to the medical model by people with disabilities in the 1970s. The model defines disability as originating from barriers and how the removal of barriers would help

individuals to participate within society beyond their "differences" [7, 9]. The social model involves the use of more "positive" language like "disabled person," "wheelchair user," and "person with learning disabilities" [10]

It is necessary to recognize that many institutions may identify more or fewer models. For example, Employment and Social Development Canada defines just the medical and social models [9], while a scoping study of models and theories of disability across literature [7] identified two other emerging models: a human rights model and CDS (Critical Disability Studies) model. The human rights model suggests that disability should not be viewed based on just medical terms or removing societal barriers but rather emphasizes that legal protections against discrimination are part of the basic rights of the individual. The CDS model is even more recent and closely related to the social model; it focuses on reframing the linguistic construction of "disabilities" and how society labels people with disabilities by providing "critical" theories on what a "functional body" is and open discussions around "capacity, potential, and possibilities" for people with disabilities. However, this model has been criticized, as it is unclear how such a "specialized" language may help improve "accessibility" in general [7].

2.2 Inclusive Language

Despite continuous efforts over the years, ensuring that language is fully inclusive remains a difficult task. Several factors contribute to the issue, including:

1. Language is constantly evolving, as hinted at previously; words may have multiple meanings within different contexts, and new meanings may arise within short time periods [2]. General language evolves fast enough, but language that refers to minority communities may evolve even more quickly as people feel safe enough within the society to express concerns [2] about the negative undertones and marginalizing nature of existing terms.

2. Individual preferences about how one wishes to be referred to may vary even within the same community. The Employment and Social Development Canada Guide [9] suggests an individualistic approach where the person's self-identification preference needs to be cross-checked, and this takes precedence over any other standardized term or phrase. Socio-cultural preferences like that of people belonging to the Deaf community, indicated by the capital-D Deaf term, where an individual identifies with the attitudes within their local community and the level of sign language skill over the amount of hearing loss [2], emphasizes the importance of respecting preferences.

3. Lack of clear guidelines and definitions around terms and phrases while referring to disabilities. For instance, identity-first language, i.e., that emphasizes the individual's

"disability identity," such as "disabled person," is often the encouraged standard in certain countries like New Zealand [9, 11]. However, person-first language that focuses on the person and not the disability, e.g., "people with disabilities," seems to be preferred in the United States [11] and Canada [9]. There is no definitive universal standard that can be followed across the different contexts that surround disability language [9].

Inclusive language, as defined by the Government of British Columbia [12], is a commitment to "using language that's free from prejudice, stereotypes, or discriminatory views of specific people or groups.". This intentional commitment ensures that no individual or community is intentionally or unconsciously excluded, disrespected, or marginalized [9, 11].

2.3 Toxic Language Detection

With the improvement in NLP algorithms in recent years, an increasing number of research has been carried out to build models that are capable of automatically detecting racism, sexism, hate speech, or other forms of toxicity within language; most of these are focused on the language used in social networks [13–15]. The Waseem and Hovy [13] dataset is common across studies that compare model performances in detecting toxicity [14]; this dataset focuses on racism and sexism expressed through tweets on X (previously Twitter), i.e., no disability language was considered. Similarly, the Hatebase database [16] with lexicons that generated the popular Davidson dataset [14, 15] contained only 25 "offensive" English terms referring to disability.

Although multiple traditional classifiers such as Support Vector Machine (SVM), Naive Bayes, and logistic regression models have been compared by related studies on the toxic language detection problem [15, 17, 18], deep learning approaches mainly models such as BERT and LSTMs have shown state-of-the-art performance since they can be fine-tuned to recognize misspelled or coded toxic language (i.e., use of symbols to replace letters) [19]. However, this study does not attempt to detect such variations within the disability-related terms used. Instead, we adopt a rule-based approach that relies on exact string matching for identifying standalone terms and fuzzy matching for phrases appearing in the corpus as described in Subsection 3.2.2.

There is less work done with an explicit focus on toxicity in disability language; this may have stemmed from the concerns raised in [15, 19] about the lack of clear distinctions as to what can be categorized as toxic, offensive or hate speech. As mentioned in [14, 15], classifying a sentence as hateful or not, or for the scope of this study, insensitive or not, is greatly affected by the individual biases of the researcher and annotators as well, making it difficult to build a model that is resilient and fully bias-free.

With the lack of enough datasets specifically focused on disability-related language, the study aims to build a reliable corpus as the first step. We focus on the ASSETS conference papers by

manually annotating sentences extracted from the abstract that contain terminology specified by the guideline documents. The following chapter explains the sentence extraction and annotation process in detail.

Chapter 3

Corpora Compilation and Annotation

This chapter details the process for compiling and annotating the sentences extracted from the abstracts of ASSETS conference papers.
A short description of the different corpora referenced throughout the study is provided below.

- Unaugmented corpus: A set of sentences extracted from ASSETS conference proceeding abstracts. See Section 3.2
- Synthetic corpus: A set of sentences generated via GPT-4o to mimic the use of insensitive and not insensitive source sentences from the unaugmented corpus. See Subsection 5.1.2
- Augmented corpus: A combination of sentences from the unaugmented corpus and synthetic sentences. See Subsection 5.1.3

3.1 Lexicon List Generation

As discussed in the introduction, language is inherently dynamic and prone to change, and this study recognizes that individual preferences may vary; therefore, no term or phrase is categorized as explicitly sensitive. Instead, the label "notInsensitive" is being used to refer to all terms that can be used within contexts that are neutral, acceptable for medical use, or reflective of people-first and identity-first approaches.

The first step of this study was to accumulate a list of terms that have been defined as "insensitive" across ten ground truth documents. These documents were a combination of multiple official recommendations, policies, and guidelines from the UN, Employment Canada [5, 9, 12] and non-official written language recommendations [20]. Quite a few words, like "handicapped," "crippled," etc., were referred to multiple times across different documents showing consistency

in the insensitivity surrounding these words. The words or phrases were sampled into a spreadsheet, forming the initial list of lexicons. The sheet was organized into columns labeled "terms," "Alternative," "Possible classification," and "Source." Although this study focuses only on two categories, i.e., "'Insensitive" and "notInsensitive," the lexicon list makes a note of when another possible subcategory within insensitive terms is emphasized by the documents. For example, the subcategories are labeled as:

- Insensitive-Noun: When the person's disability is used as a noun equating the individual with their condition, e.g., the deaf, epileptics.
- Insensitive-Slur: The term used is regarded as derogatory within the English language, e.g., cripple.
- Insensitive-Patronizing: These refer to terms that are used as mainstream euphemisms that may be unnecessarily patronizing, like "handicapable".
- Insensitive-Negative: These are terms that have an overtone that makes the disability seem negative or as a burden, e.g., suffering from autism [11, 21, 22].

The guidelines investigated also included some alternate language at times to be used for certain terms or phrases they defined as insensitive; the lexicon list makes a note of these alternatives, labeling them as "notInsensitive".

Each document was synthesized to extract core principles for writing about disabilities. Although phrasing varied across the documents, the underlying recommendations and guidelines were largely similar. Therefore, general principles have been rephrased for consistency. The synthesis of the guideline documents identified the following general principles for inclusive writing:

1. Use the terms "non-disabled" or "persons without disabilities" rather than "normal" or "healthy" as the word "disabled" is an adjective and should not be used as a noun [3, 9, 11, 21–23].
2. Avoid using "living with a disability" because that can sound like the disability is a negative burden [9].
3. Avoid language that equates persons with their condition (e.g., epileptics, the deaf) [3, 11, 21–23].
4. Avoid language that has unnecessary negative overtones or projected feelings (e.g., stroke victim or sufferer) [11, 21, 22].
5. Avoid language that is regarded as a slur (e.g., cripple) [11, 21].

6. Avoid trendy euphemisms like physically challenged, special, differently-abled, and handicapable, as they can be patronizing [11, 21, 22].

7. Use proper definitions of terms like "impairment," "disability," and "handicap" based on contextual differences [11, 21, 22].

8. In a cultural context, it is appropriate for the word Deaf to be spelled with a capital D to refer to members of the Deaf Community [11, 21].

9. "Hearing impaired" and similar impairment language should only be reserved for medical writing [11, 21].

10. The term "Hard of hearing" should be reserved for those not identifying with the deaf community [11, 21].

11. When referring to one's ability to use a keyboard or standard mouse, use specific terms like "motor or dexterity" rather than a generalized "mobility impairment"[11, 21].

12. Be concise and specific when referring to cognitive disabilities. For example, specific cognitive disability such as language processing difficulties [11, 21].

13. Use "person with a mental disability" to refer to different forms of clinical illnesses such as schizophrenia, depression, and emotional disorders [11, 21].

14. Use terms such as "neurotic" and "psychotic" only for clinical writing [11, 21].

15. The phrase "visually impaired" is NOT recommended in scientific writing: specify within the context of the study, i.e., screen reader user or degree of vision loss [11, 21, 22].

16. When the emphasis is on a specific topic, then the individuals' preferences are to be referenced instead of their disability, which may have less relevance to the discussion. For instance, when writing about communication styles, mentioning that an individual uses sign language may be enough without the need to specify their disability. Similarly, when the focus is hearing aids, the degree of hearing loss can be used to reference the situation [4, 9, 22].

17. When referencing accommodations, place focus on the term "accessible" rather than the disability, i.e., "accessible parking" vs. "handicapped parking" [4, 9].

18. Instead of "Person who has trouble/difficulties/challenges," use "Person who needs..." to avoid perpetuating condescending stereotypes [4, 9].

19. Define the context of specific terms. For example, does "blind people" mean people who having varying visual acuity or people who primarily use screen readers to access the computer [11, 21]?

20. Ensure consistency, i.e., if an impairment model has been chosen at the beginning of the document, it should be followed throughout the writing [11, 21].

It is important to note that this list of guidelines is not exhaustive but forms the foundation for the codebook used to annotate sentences later in the study (see Section 3.3). Refer to Appendix C for the lexicon list.

3.2 Automated Extraction and Initial Annotation

In order to construct the corpus, we focused on the ASSETS conference papers as these were expected to have a good amount of disability terminology within an academic context, given the conference's emphasis on accessibility research. The following sections detail the step-by-step process followed to build the corpus from the ground up.

3.2.1 Abstract Compilation

For each conference year between 1994 and 2024, the abstracts of all papers were attained by exporting the total citations of each conference from the ACM Digital Library's ASSETS Proceedings page in a BibTeX format. A simple Python script was then run to extract the title, DOI, and abstract from the BibTeX for each proceeding. It is important to note that keynote citations for all of the conferences were excluded as these do not have abstracts. The year 2024 had the highest number of proceedings, with a total of 138, while 1994 had the lowest, with only 22 proceedings.

3.2.2 Keyword Extraction Algorithm

An initial sample of ten papers was used to test for a suitable rule-based keyword-matching approach to potentially carry out automated first-pass annotation on the extracted abstracts. A few algorithms were tried at this stage, and these are minimally outlined below:

1. Exact Word Match: This algorithm did a simple search looking for "exact" matches of terms in the lexicon list within sentences of abstracts and extracted these sentences. As expected, this method was very inflexible and did not look for variations in phrasing; however, it had fewer misinterpreted subwords, i.e. "crip" was not considered as part of "description".

2. Fuzzy Matching + Adjusted Thresholds: This approach involved using a fuzzy matching algorithm [24] with partial matching that led to a lot of incorrectly tagged words that existed within an entirely different context, for example, “special” within “especially.”

3. Combined approach: The final approach was a combination of exact matching for terms and fuzzy matching for phrases. This method achieved the best results for the tested sample.

The feasibility of using Keyword spaCy [25] was very briefly explored to extract the sentences from abstracts; however, it captured significantly fewer sentences than the combined approach described above. With access to more computational resources, a detailed comparison between the rule-based approach used in this study and other machine learning-based keyword extraction methods could be explored to find an algorithm with the highest success in capturing variations of terms/phrases from the lexicon list.

The combined approach was applied to the file containing all the compiled abstracts to extract sentences based on the lexicon list. The machine-annotated sentences were stored in spreadsheet in the format: Title, Sentence, Matched_Terms, Automated_Annotation, Manual_Annotation, Comments, and Source_File. This file contains 1090 records and is referred to as the **unaugmented corpus** throughout the rest of the study.

3.3 Manual Annotation and Dataset Refinement

A subset of 200 sentences, about 18.3%, was carefully selected from the unaugmented corpus to ensure that edge cases, i.e., sentences containing less frequent terms like “homebound,” are represented; this could not be ensured if the sentences were selected at random. Two annotators used a data annotation guidebook that outlined the basic criteria similar to those found in Section 3.1, along with specific examples to label a given sentence as “insensitive” or “notInsensitive.” Detailed examples provided to the annotators can be found in Appendix B. The annotators separately labeled this subset within their respective spreadsheets, providing the reasoning behind certain choices in a “Comment” column. Inconsistencies in labels were discussed in detail during a review meeting, and a final classification was assigned only after both annotators reached a consensus; the remaining sentences were annotated based on the insights gained from the review meeting. This structure for evaluating manual annotation is consistent with existing literature [15].

When terms that are explicitly marked as “insensitive” are mentioned rather than used within a sentence, it was tagged as “notInsenstive”; for example, “In these cases, we draw similarities between sighted and visually impaired users, in that sighted users cannot see the target of a Web

Anchor and are therefore "handicapped" by the technology" is labeled as "notInsensitive" in comparison to the sentence "Making information more accessible to the disabled has obvious synergistic benefits for "non-handicapped" people alike as reflected by the importance of the concept of workforce diversification for overcoming unexpected future challenges and potential stumbling blocks" that is marked as "insensitive".

3.3.1 Inter-annotator Reliability

The final Cohen's κ was 0.8223 with 16 mismatched records. A Cohen value κ between 0.81-1.00 is considered as "almost perfect agreement" [26, 27]. During the review meeting, the following discrepancies were discovered between annotator A's and annotator B's labeling style, which may have risen due to differences in interpretation and lack of specific examples in the annotation guide:

1. Certain sentences could be annotated with either label depending on the broader context and surrounding sentences. For example, consider the following sentence that both annotators interpreted differently "We have designed a virtual coach system, in which an animated character engages users in simulated face-to-face conversation to provide health education and motivate healthy behavior." What healthy behavior refers to is unclear in this example without more context. For the scope of this thesis, it was decided that the sentences would be labeled as a standalone component following a sentence-level granularity for the analysis done throughout the study.

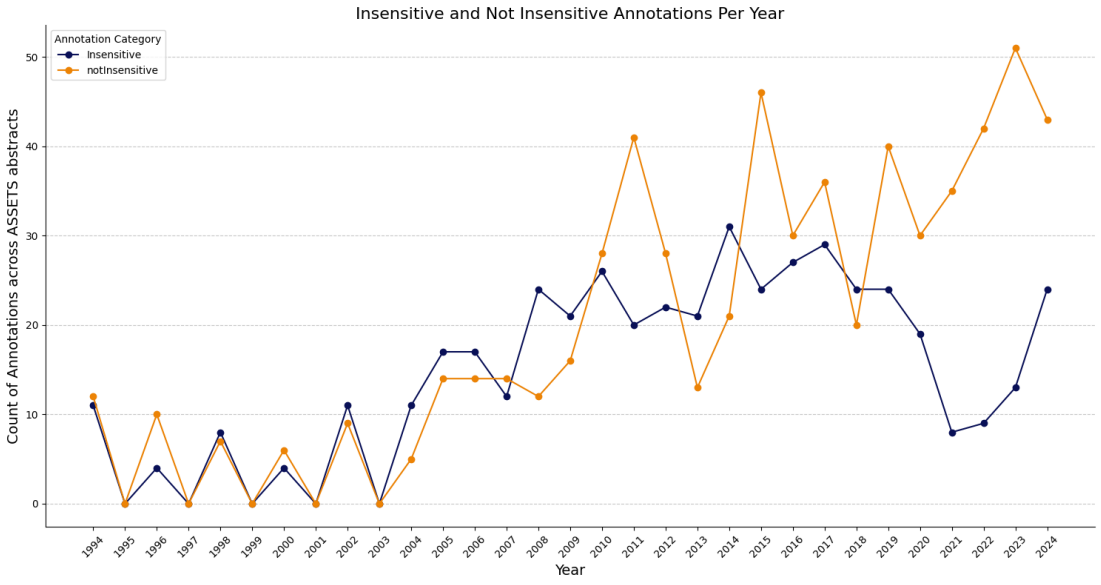
2. The phrase hard of hearing: Annotator A marked multiple occurrences of the term "hard of hearing" as "notInsensitive" whereas Annotator B marked the same sentences as "insensitive." In the discussion, it was discovered that this inconsistency occurred as a result of Annotator A followed [11] that still recognizes the term while Annotator B used [9] to justify their choice. It was decided that sentences with the term "hard of hearing" would be inherently considered "notInsensitive" unless another term within the sentence leads to insensitivity. This decision was made under the impression that the resource Annotator B used is fairly recent (2024); multiple other sources still accept "hard of hearing," such as [28] and the Canadian Association of the Deaf calling it a "medical and sociological term" [29]. Once this misunderstanding was resolved across Annotator B's label, the final Cohen's κ reached 0.8223 with 16 mismatched records from an initial value of 0.6138.

The mismatch due to the "hard of hearing" is a clear example of how the change in language over time often makes it challenging to devise proper standards and guidelines around disability.

3.4 Corpus Analysis

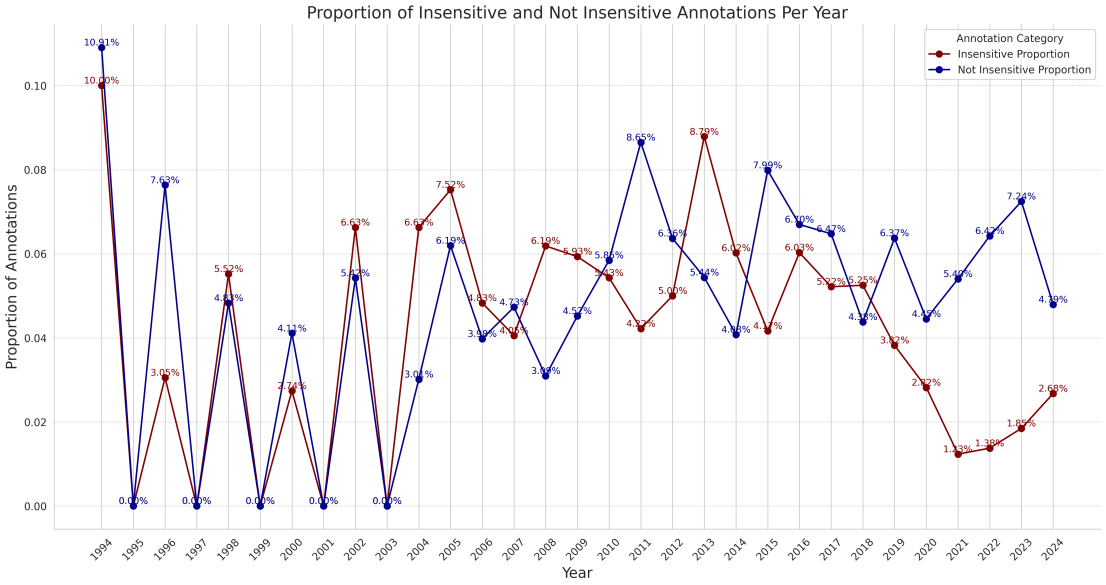
The unaugmented corpus was analyzed to prepare for data augmentation and model training. We removed all duplicate records before running the analysis below. Figure 3.1 visualizes the

Figure 3.1: Yearly count of sentences annotated as "Insensitive" and "Not Insensitive" in ASSETS conference paper abstracts. The chart highlights the frequency of these annotations for each year.



count of sentences that were manually annotated as "Insensitive" and "notInsensitive" over all the ASSETS conference paper abstracts. Note that the years 1995, 1997, 1999, 2001, and 2003 have a count of zero since these years did not hold an ASSETS conference, hence, did not have any papers to consider. In the later years, i.e., 2019 onward, an increase in the number of sentences that were annotated as "notInsensitive" can be identified. While the number of "insensitive" annotations initially increased over the earlier years, there was a gradual decline after 2017. However, an upward trend is seen once again from 2021 onward. Figure 3.1 shows that the year 2023 had the highest count of sentences that were being tagged as "notInsensitive" based on the phrases/terms in the Lexicon List, whereas 2014 shows the highest count of "insensitive" sentences. This information may be misleading if interpreted without Figure 3.2, which normalizes the count of annotations based on the total number of sentences within each abstract of every conference paper in a year; the years 1994 and 1996 are now being penalized more as 10% of the sentences in the 1994 abstracts were insensitive while 2024 had only 2.68% of the sentences being tagged as "insensitive." Figure 3.3 shows the distribution of terms or phrases from the Lexicon list that were found within the analyzed abstracts. The keyword extraction algorithm identified 43 terms in total; note that this is just 17.41% of the total number of terms, both "insensitive" and "notInsensitive" that were present in the Lexicon List. Figure 3.3 shows the logarithmic representation of the number of "insensitive" and "notInsensitive" sentences for each term, with the actual counts displayed on each bar. There are 623 "notInsensitive" sentences

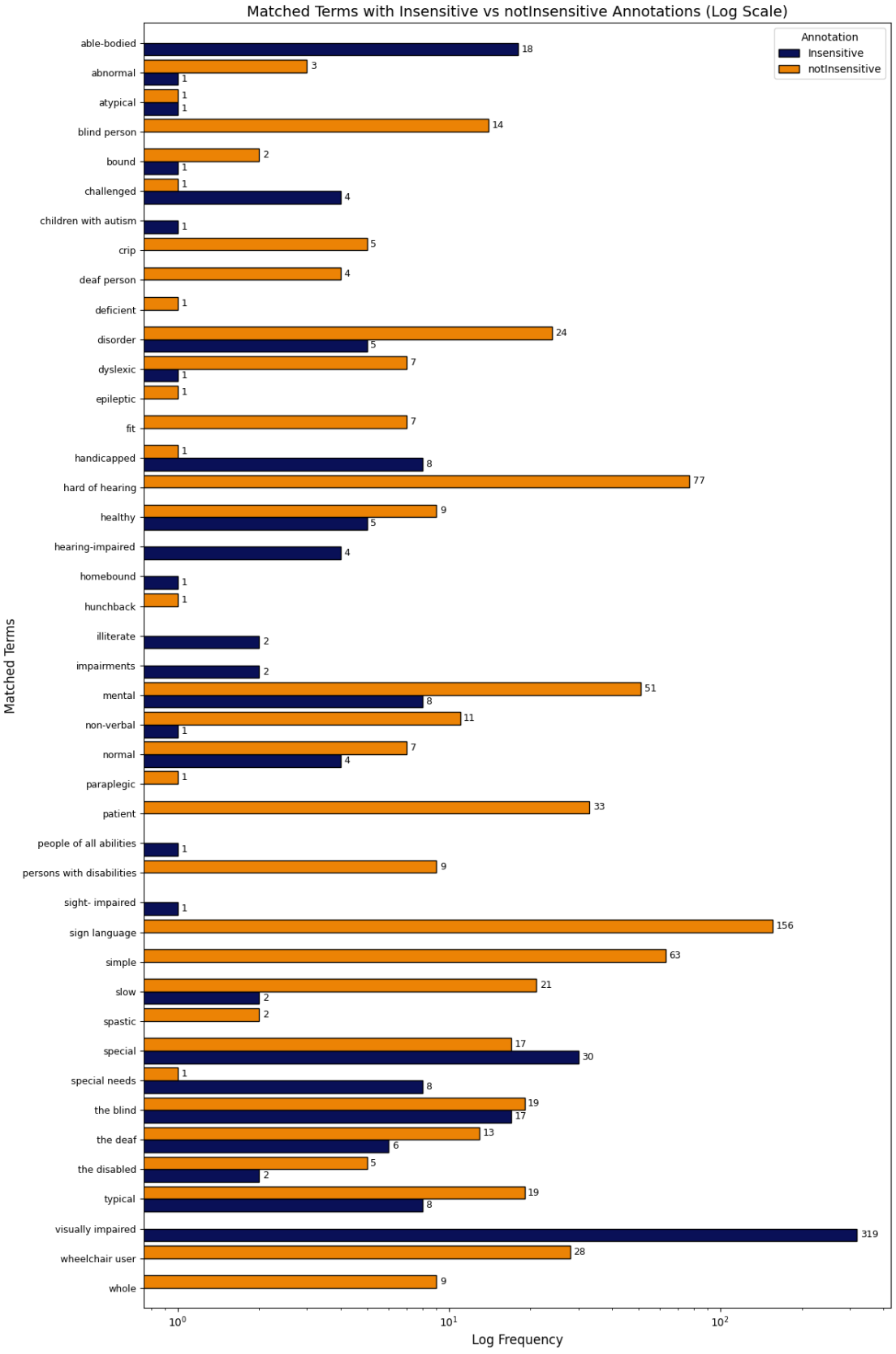
Figure 3.2: Yearly proportion of sentences annotated as “Insensitive” and “Not Insensitive” in ASSETS conference paper abstracts. The chart normalizes annotation counts by the total number of sentences within abstracts per year.



and 461 “insensitive” sentences. Note how the phrase “visually impaired” makes up 29.43% of the total sentences. The imbalance in the corpus is not solely because of the lower number of ”insensitive” examples but also due to the disproportionate occurrence of certain terms and phrases.

The next chapter investigates whether this manually compiled and annotated corpus could be used to build a model to detect instances of insensitive language within sentences of the ASSETS conference abstracts.

Figure 3.3: The bar chart illustrates the logarithmic frequency of sentences for each term found within the unaugmented corpus with the “insensitive” and “notInsensitive” categories side by side.



Chapter 4

Empirical Studies

This chapter builds on the unaugmented corpus created in Chapter 3. It describes the methodology used to investigate the feasibility of using machine learning models to identify insensitive terms and phrases within academic text. As discussed in Chapter 2, machine learning has emerged as a powerful tool for automating the detection of toxic or insensitive language over the years. Given the evolving nature of language, manually identifying insensitive phrases is resource-intensive, making automation a lucrative alternative, and the use of machine learning models to learn from annotated examples and generalize to unseen text helps detect potentially insensitive language more efficiently than manual processes. All model training and testing were performed in **Google Colab** using **Python 3.11.11**. The exact configurations of the models, along with any custom classes used, are detailed in Appendix A. The complete source code and experiment notebooks are available in the following GitHub repository: https://github.com/RobyRoshna/Insensitive-Lang-Detection.

4.1 Models

In order to build a model that is able to recognize "insensitive" terms based on the compiled unaugmented corpus, we fine-tune BERT. Developed in 2018 by researchers at Google AI Language, BERT is the state-of-the-art model for Natural Language Processing tasks [30, 31]. Moreover, BERT models have been excessively used for toxic language detection with great success, as they are capable of capturing nuances in language caused by contextual differences and, hence, can improve classification accuracy [32]. The original BERT paper [30] introduces BERT-base with 110 million parameters and BERT-large with 340 million parameters. The performance offered by BERT-large is higher than that of BERT-base in the tasks studied by [30], but the base model still showed better performance than prior state-of-the-art models,

even on datasets with very little training data. In this study, we fine-tune the pre-trained BERT-base uncased (BERT-base version that does not differentiate between lowercase and uppercase characters) to keep computational resource usage minimal while training multiple models.

In order to have a baseline for performance comparison with the BERT model, we use logistic regression. Traditional classifiers are compared by related works [15, 17, 18], and logistic regression was one of the traditional models that performed best on the language detection problem, though in most cases, performance is heavily dependent on the corpus [14, 15]. Unlike BERT, which captures contextual relationships between words using deep transformer-based representations for words, logistic regression is a linear model that does not account for word order or semantic differences caused by certain word combinations [17]. It estimates the probability that a given sentence belongs to a particular class, our case, "insensitive" or "not insensitive," using a logistic function applied to weighted representations of the input features (words) [17]. In this study, we use a standard text vectorization technique, TF–IDF (Term Frequency–Inverse Document Frequency), to convert sentences into numerical feature representations suitable for input into the logistic regression model.

4.2 Measures of Model Performance

Standard classification report metrics from Scikit-learn, mainly accuracy, precision, F1 Score, and recall, are being used to evaluate each model. These metrics are widely used to evaluate the performance of a classifier model [18]. They are detailed below:

- Accuracy: The number of correctly classified instances out of the total number of samples (true positives and true negatives).

$$\text{Accuracy} = \frac{TP \text{ (True Positives)} + TN \text{ (True Negatives)}}{TP + TN + FP \text{ (False Positives)} + FN \text{ (False Negatives)}}$$

- Precision: This is the proportion of correctly predicted positives out of all samples predicted as positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall: This is the proportion of actual positive samples that were also predicted as positive.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1 Score: The harmonic mean of precision and recall that balances both metrics.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Note that the default cross-entropy loss function for binary classification is used to calculate the test and validation losses for the Bert base models.

4.3 Exploratory Study: Evaluating the Unaugmented Corpus

4.3.1 Method

To prepare for fine-tuning on BERT-base, we did some preprocessing on the unaugmented corpus by dropping all other columns apart from "Sentence," " Manual_Annotation," and "Matched_Terms." The labels within the "Manual_Annotation" columns were encoded with 1 for "insensitive" and 0 for "notInsensitive." The corpus was then divided into the training, validation, and test sets using 80:10:10 splits where 80% of the dataset is used for training the BERT model, and the remaining 20% is divided equally between the validation and test sets [33]. Although hyperparameter tuning is not performed in this study, the presence of the validation set keeps this open for future work to use to monitor for improvements in the accuracy or generalizability of the model upon adjusting the hyperparameters, as demonstrated by Xu et al. [34], relying solely on a single split between training and test data can lead to misleading performance estimates.

The Hugging Face Transformers library was used for model training and evaluation with unmodified default hyperparameters and model configurations from the BERT-base implementation [35] (See Appendix A for exact configurations). The study uses the Scikit-learn package [36] for both preprocessing and training algorithms. For ease of reproducibility, each stage of data processing, model training, and testing uses 42, a common random seed [37]. It is important to note that 42 was also explicitly specified as the random_state value for all train-test splits. All the BERT-base models in the study were fine-tuned using `BertForSequenceClassification` for binary labels i.e, "insensitive" and "notInsensitive". The model was trained for three epochs. An epoch refers to one complete pass through the entire training set during model training, where the model updates its internal weights based on the data to possibly minimize errors as it learns over time [38]. Future works could monitor the improvement in model performance upon adjusting the number of epochs.

The baseline logistic regression model was also trained using the same unaugmented corpus splits as the BERT model with unmodified default parameters from Scikit-learn consistent with related works [15, 18] that use traditional classifiers as baseline models.

Table 4.1: Performance Metrics of the BERT Model Trained and Tested on Unaugmented Corpus

Dataset	Accuracy	F1 Score	Precision	Recall
Training	0.9942	0.9932	0.9919	0.9946
Validation	0.9722	0.9670	0.9778	0.9565
Test	0.9266	0.9069	0.9750	0.8478

4.3.2 Results: Models Trained on the Unaugmented Corpus

The BERT-base model fine-tuned on the unaugmented corpus achieved a test accuracy of 92.66% (see Table 4.1). The accuracy of the training set was comparatively higher at 99.42%, bringing the difference in accuracies between the test and training set to almost 7%. Therefore, the model seems to be learning very well from the training set but is slightly struggling to generalize to unseen data from the test set - this is an indicator of potential overfitting to training data.

Compared to the BERT model, the logistic regression model, trained in the same unaugmented training, validation, and test sets as the BERT model, has a slightly lower test accuracy of 88. 99% (see Table 4.2) with no false positives and 12 false negatives; that is, it misclassifies 12 "insensitive" sentences as "notInsensitive".

Table 4.2: Performance Metrics of the Logistic Regressor Trained and Tested on Unaugmented Corpus

Dataset	Accuracy	F1 Score	Precision	Recall
Training	0.9619	0.9535	0.9941	0.9160
Validation	0.9259	0.9048	1.0000	0.8261
Test	0.8899	0.8500	1.0000	0.7391

4.3.3 Analysis

Table 4.3: Loss and Performance Trends Across Training Epochs

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1 Score
1	0.2412	0.1693	0.9259	0.9524	0.8696	0.9091
2	0.1697	0.1223	0.9722	0.9778	0.9565	0.9670
3	0.0405	0.1191	0.9722	0.9778	0.9565	0.9670

Table 4.3 displays the loss trends across the three epochs that the BERT model was trained for. In the first epoch, accuracy was 0.9259, and it ultimately reached 0.9722 in the third epoch. The training loss started at 0.2412 and decreased significantly to 0.0405 by the third epoch, whereas the validation loss seems to have a more gradual decline, as it starts at 0.1693 and reaches just 0.1191 by the third epoch. This widening gap between training and validation losses may indicate potential overfitting. As shown in Figure 3.3, certain terms have a disproportionate number of training examples within each category (see Section 3.4). With this context about the disproportionate representation of terms within the corpus combined with the loss trends observed during training, it raises concerns that the model may be memorizing patterns associated

with specific terms. Therefore, the model is at risk of just overfitting to patterns and characteristics of the current corpus rather than learning to generalize to unseen corpora within the same domain.

From Table 4.1, the test precision is 0.9750, so the model rarely has false positives where

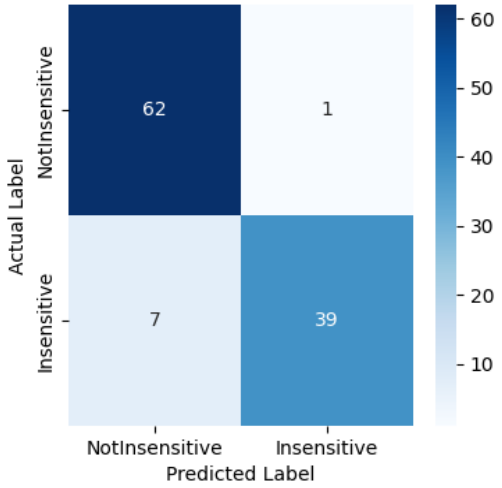


Figure 4.1: Unaugmented BERT Confusion Matrix

"notInsensitive" sentences are misclassified as "insensitive"; in fact, it only predicted just one such sentence. Recall for the test set is comparatively lower at 0.8478. This means that the model could not identify some truly "insensitive" sentences with confidence. Figure 4.1 shows the model's confusion matrix with seven false negatives and one false positive. The disparity between precision and recall explains the test F1-score of 0.9069 (see Table 4.1). An ideal classifier would be expected to have a F1-score of 1 [39]; therefore, though this unaugmented BERT model indicates positive metrics, it may still fail to identify a considerable amount of "insensitive" sentences.

The BERT model outperforms the logistic regression model, but the disproportional distribution of "insensitive" and "notInsensitive" sentences across each term in both the training and testing datasets raises concerns about the reliability and generalizability of the model. The next chapter explores data augmentation as a possible strategy to get a more representative and balanced corpus and analyses whether a model trained on the resulting augmented corpus could achieve comparable performance metrics.

Chapter 5

Expanding the Corpus: Augmentation and Fine-tuning of a New Model

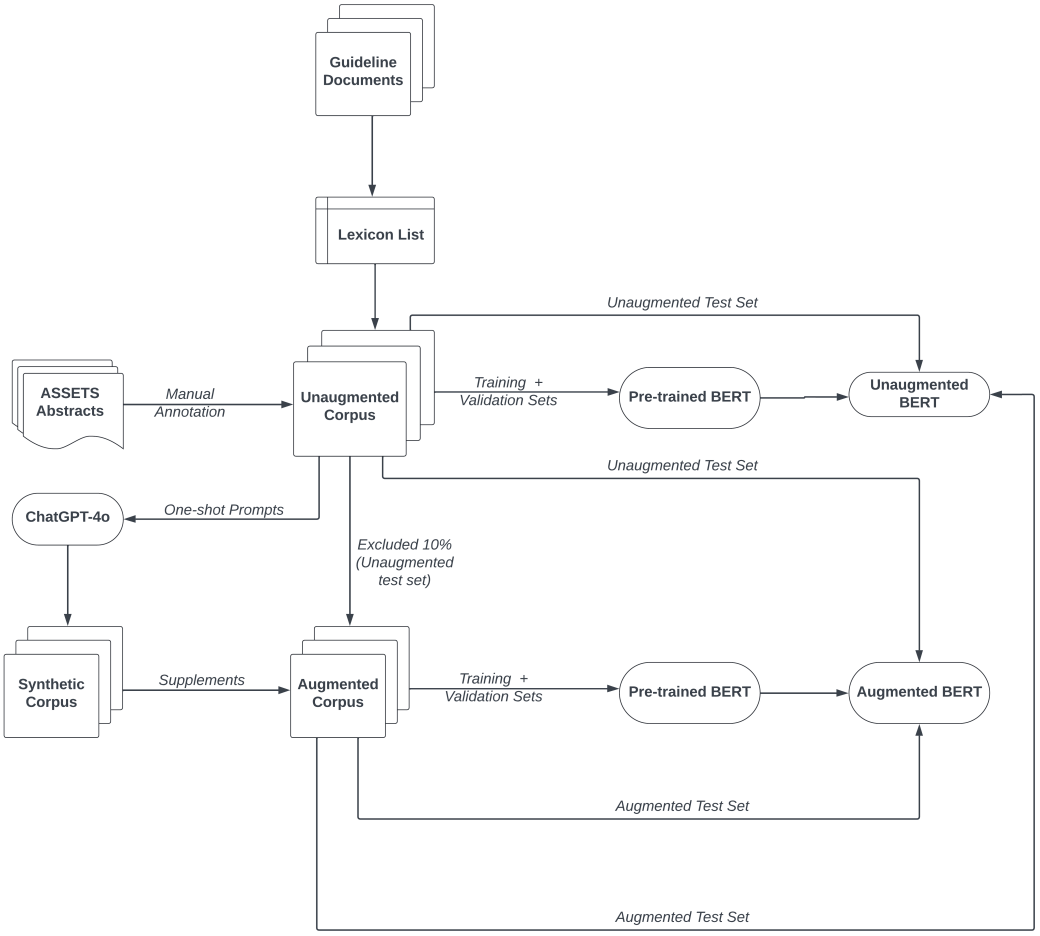
This chapter outlines the augmentation process followed to supplement the unaugmented corpus, the evaluation of the models built upon the resulting augmented corpus, and how they compare to the model discussed in Chapter 4. Figure 5.1 illustrates the workflow followed by the study indicating the flow of different datasets across training and evaluation of each model studied.

5.1 Data Augmentation

Given the obvious imbalance in the unaugmented corpus as analyzed in Section 3.4, a data augmentation option was considered to ensure that no particular term or phrase like "visually impaired" had a disproportional advantage during the model training process. We use a similar strategy as [40] by using GPT-4o to generate more sentences that are similar to the ones in our dataset. However, before the augmentation process, we set aside 10% of the unaugmented corpus, specifically the test set of the unaugmented BERT model (see Subsection 4.3.1) to evaluate the performance of the model trained on augmented data on completely unseen sentences from the abstracts.

For the purposes of this study and to ensure minimal use of resources and computational power, the corpus was augmented to 50 sentences for each of the 43 terms or phrases that existed in the unaugmented corpus. The choice between 25, 50, or 100 examples was based on a possible trade-off: the existence of certain terms like "paraplegic" with no insensitive occurrences within the corpus would mean that all the "insensitive" examples in the augmented corpus would be purely synthetic for this term. To minimize the model's reliance on synthetic data while learning patterns, 50 examples per word was chosen as an optimal balance. However, future

Figure 5.1: The workflow of the transition from ASSETS and guideline documents to an annotated corpus that can be used to fine-tune BERT



work could experiment by varying this number and comparing model performance for each. The final augmented corpus has a total of 2150 data records where each term has 25 "insensitive" examples and 25 "notInsensitive" examples. The steps followed for augmentation are detailed below:

5.1.1 Step I: Random Selection of an Existing Example

A one-shot prompting technique was used where the model is given an example along with a general description of the task. To construct the prompt, one sentence from the "insensitive" category and another from the "notInsensitive" category were randomly selected from the unaugmented corpus. This method, described in [41, 42], is based on the findings that language models can better understand and generalize the task when given some examples. In the case of this study, examples were necessary for the generated sentences to remain within the same domain.

5.1.2 Step II: Customizing Prompts and Sentence Generation

Figure 5.2 is the general prompt structure followed for every term recorded in the unaugmented corpus. Note that there exists certain terms that lacked examples for a given category, in such

> Below are examples labeled as "Insensitive" and "Not Insensitive" that use the term ⟨Term⟩:
>
> **Special Note (Optional):**
> ⟨Provide any other information specific to the given term⟩
>
> **Insensitive example:**
> ⟨Insensitive example selected at random from the unaugmented corpus⟩
>
> **Not Insensitive example:**
> ⟨Not Insensitive example selected at random from the unaugmented corpus⟩
>
> Generate 25 new sentences using the term ⟨Term⟩ in different contexts where it could be perceived as "Insensitive" such that each sentence should follow the domain, tone, and style of the example sentence.
>
> Additionally, generate 25 sentences using the term ⟨Term⟩ in contexts that are "Not Insensitive" such that each sentence should follow the domain, tone, and style of the example sentence.

Figure 5.2: One-shot prompt structure used to generate synthetic sentences

cases the "Special Note" option was leveraged to provide other instructions about the term based on the guidelines in Section 3.1. Every generated example was manually verified and re-annotated using the same annotation codebook used for the unaugmented corpus. The 2150 annotated sentences were then recorded in a spreadsheet with the columns "Term", "Prompt", "Sentence", and "Label"; this is referred to as the **synthetic corpus** throughout the study.

5.1.3 Step III: Augmented Corpus Creation

The augmentation process gave precedence to the unaugmented corpus made up of source data before selecting sentences at random from synthetic data to account for the remaining examples within each category. Note that only 90% of the unaugmented corpus was available to select from at this stage in the study, as 10% was set aside for testing later during model evaluation. Figure 5.3 shows the contributions from the unaugmented source data and synthetic data within both the "insensitive' and "notInsensitive' categories for every word. For some words like "able-bodied", all its "notInsensitive" examples are GPT-4o generated, while others like "special" have

sufficient "insensitive" examples from the source data alone. The resulting corpus is referred to as the **augmented corpus** throughout the study.

5.2 Exploratory Fine-tuning on Purely Synthetic Sentences

The synthetic corpus generated in Subsection 5.1.2 was also split into the train, validation, and test sets in the 80:10:10 ratio consistent with the methodology in Section 4.3 in order to explore how BERT learns from solely synthetic data. Another BERT-base model was fine-tuned on this purely synthetic corpus to visualize the impact of using just GPT-4o generated data. The model

Table 5.1: Performance Metrics of the BERT Model Trained on Pure GPT-4o Generated Data

Dataset	Accuracy	F1 Score	Precision	Recall
Training	1.000	1.0000	1.0000	1.0000
Validation	0.9767	0.9772	0.9640	0.9907
Test	0.9861	0.9862	0.9727	1.0000

achieved a test accuracy of 98.61% (see Table 5.1). The accuracy of the validation set followed closely behind at 97.67%. The test precision is 0.9727, so it also rarely has false positives; three sentences in total were wrongly tagged as "insensitive". Table 5.1 shows perfect metrics for the training set, suggesting that the model is learning the training examples too well, possibly due to the fully synthetic nature of the data.

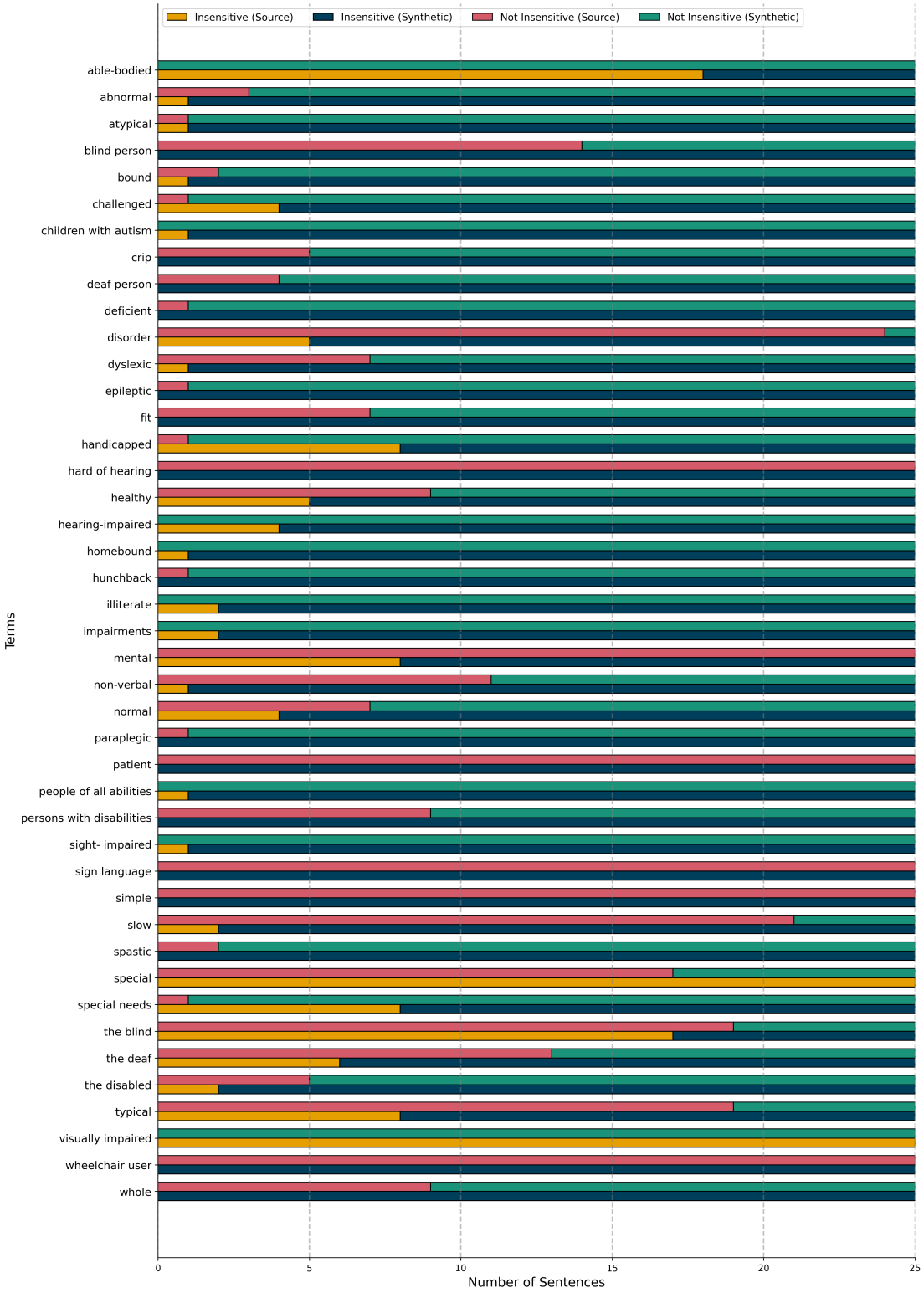
Recall for the test set is a perfect one; this means that the model could identify all the "insensitive" sentences. A perfect recall could also indicate that the model may be just memorizing the training set, and the training and testing sets could have been very similar to each other. Table 5.2 shows

Table 5.2: Loss and Performance Trends Across Training Epochs (Pure GPT-4o Data)

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1 Score
1	0.2127	0.0641	0.9674	0.9810	0.9537	0.9671
2	0.0508	0.0750	0.9814	0.9727	0.9907	0.9817
3	0.0075	0.1061	0.9767	0.9640	0.9907	0.9772

clear signs of overfitting as the validation loss increases while the training loss decreases across epochs. More training data alone could not improve the model in this case, as the model may be learning other underlying patterns that may have seeped into the generated data, and this is somewhat expected behavior when training on pure synthetic data [43].

Figure 5.3: Bar chart illustrating the distribution of sentences with the “Insensitive” and “notInsensitive” categories side by side. Each category shows contributions from the unaugmented corpus and GPT-4o. The vertical axis lists the terms, while the horizontal axis indicates the number of sentences.



5.3 Fine-tuning on the Augmented Corpus

5.3.1 Method

The augmented corpus from Subsection 5.1.3 was processed by keeping only columns "Sentence", "Label," and "Term." The rest of the model training and testing followed the same methodology as that of the unaugmented corpus.

5.3.2 Results: Models Trained on the Augmented Corpus

From Table 5.3, the BERT model trained on the augmented corpus achieved a 96.28% accuracy on its test set in comparison to a training accuracy of 99.59%. The difference in accuracy is quite low, hinting that the augmented model may be learning and generalizing well. The test F1 score is at 0.9626, with a perfect balance between precision and recall. Therefore, the model predicted exactly 4 false positives and false negatives out of the 215 sentences in the test set of the augmented corpus (see Figure 5.4). Table 5.4 shows that the validation loss starts at 0.2423 in the first epoch and gradually decreases along with the decrease in the training loss as it reaches the third epoch. Future works could investigate whether training for more epochs would further increase accuracy.

Table 5.3: Performance Metrics of the BERT Model Trained and Tested on Augmented Corpus

Dataset	Accuracy	F1 Score	Precision	Recall
Training	0.9959	0.9959	0.9954	0.9965
Validation	0.9674	0.9671	0.9810	0.9537
Test	0.9628	0.9626	0.9626	0.9626

Table 5.4: Loss and Performance Trends Across Training Epochs (Augmented Corpus)

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1 Score
1	0.1985	0.2423	0.9209	0.9027	0.9444	0.9231
2	0.1325	0.1993	0.9488	0.9619	0.9352	0.9484
3	0.0131	0.1419	0.9674	0.9809	0.9537	0.9671

The logistic regression model on the same augmented training, validation, and test datasets achieved a test accuracy of 89.30%, as shown in Table 5.7. This classifier had nine false positives and 14 false negatives, and the misclassifications mostly consisted of sentences from the synthetic corpus.

5.3.3 Analysis

Since the augmented BERT model showed strong metrics on the augmented test set, we decided to evaluate its performance on the unaugmented test set that does not have any synthetic data.

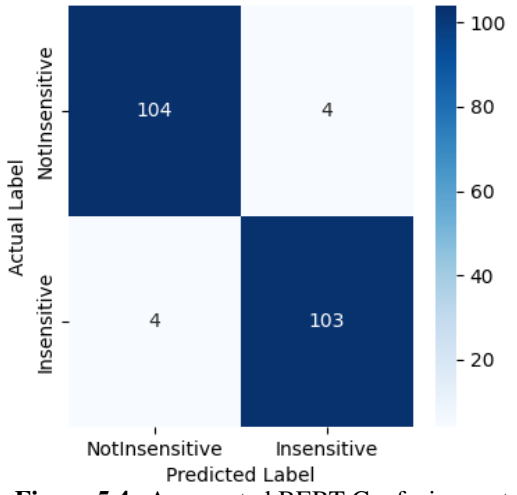


Figure 5.4: Augmented BERT Confusion matrix

This allows us to assess how well the model generalizes to real-world sentences, particularly since the augmented corpus used during training does contain GPT-4o generated examples. The presence of these synthetic sentences within the augmented corpus may introduce other patterns or structures not typically found in original ASSETS abstracts that could affect the model's performance as seen in Section 5.2. We also evaluate the performance of the unaugmented BERT from Section 4.3 on the augmented test set to measure how our initial BERT model would react to an entirely different test set and compare the resilience of the two BERT models.

5.3.3.1 Evaluating the BERT Model Trained on Augmented Corpus Using the Unaugmented Test Set

Table 5.6: Augmented BERT Model's Performance on its Augmented Test Set vs. Unaugmented Abstracts Test Set

Metric	Augmented Test Set	Unaugmented Test Set
Accuracy	0.9628	0.9358
Precision	0.9626	0.9354
Recall	0.9626	0.9327
F1 Score	0.9626	0.9340

The accuracy of this augmented BERT model on the unaugmented test set from Subsection 4.3.1 was 93.58%. The difference in performance in terms of accuracy for both test sets is just 2.7%; this is significantly better than the unaugmented BERT model in Subsection 5.3.3.2 . There were three false positives and four false negatives on this test set, see Table 5.5. This suggests that the model is able to generalize quite well to entirely new data, even when the test set solely consists of source sentences from the abstracts. However, it is important to note that the unaugmented test set is considerably smaller than that from the augmented corpus in Table 5.6. For comparison, the logistic regression model trained on the augmented corpus was also evaluated using the same

Table 5.5: False Positives and False Negatives from evaluating the unaugmented test set on the augmented BERT model.

Sentence	Annotated Label	Matched Term	Confidence Score
False Positives (Predicted: Insensitive, Annotated: NotInsensitive)			
We started working with the patient when he was two and a half years old.	0	patient	0.9557
However, language immersion can be particularly challenging for hearing parents of deaf children to provide as they may have to overcome many difficulties while learning sign language.	0	sign language	0.9946
Gaze typing for people with extreme motor disabilities, like full body paralysis, can be extremely slow and discouraging for daily communication.	0	slow	0.9891
False Negatives (Predicted: NotInsensitive, Annotated: Insensitive)			
Our research contributes empirical evidence demonstrating that that autistic sense-making on Twitter is constituted by (1) engaging in dynamic discussions of life experiences, (2) countering stigma with actions of advocacy, and (3) enacting neuro-atypical social norms.	1	atypical	0.9987
In this paper, we describe why designers need to look beyond the twin aims of designing for the 'typical' user and designing "prostheses".	1	typical	0.9979
In our system, besides the user's speech, the speech of individuals with the same type of disorder and even different types of disorders is also incorporated.	1	disorder	0.9967
Children with autism spectrum disorder and other developmental disorders tend to have difficulty in language and communication, especially in abstract language concepts like prepositions.	1	disorder	0.8182

Table 5.7: Logistic Regression Model's Performance on the Augmented Test Set vs. Unaugmented Abstracts Test Set

Metric	Augmented Test Set	Unaugmented Test Set
Accuracy	0.8930	0.7064
Precision	0.9118	0.9375
Recall	0.8692	0.3261
F1 Score	0.8900	0.4839

unaugmented test set. Table 5.7 shows that the resulting accuracy is 70.64%, and the F1 score is also quite low at 48.39, with 31 false negatives and one false positive.

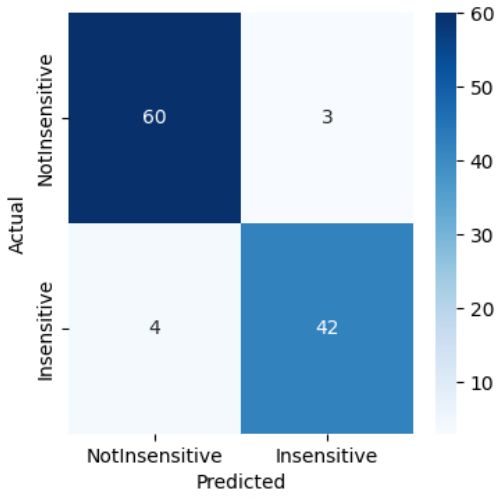


Figure 5.5: Unaugmented Test Set on Augmented BERT

5.3.3.2 Revaluating the BERT Model Trained on Unaugmented Corpus Using the Augmented Test Set

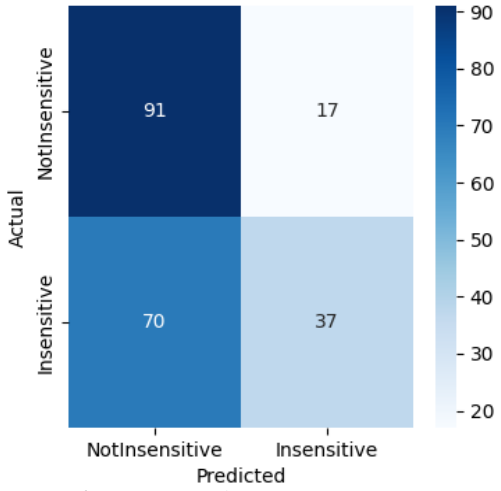


Figure 5.6: Augmented Test Set on Unaugmented BERT

Table 5.8: Unaugumented BERT Model's Performance on its Unugmented Abstracts Test Set versus Augmented Test Set

Metric	Unaugmented Test Set	Augmented Test Set
Accuracy	0.9266	0.5953
Precision	0.9069	0.6252
Recall	0.9750	0.5942
F1 Score	0.8478	0.5681

To evaluate the generalizability of the unaugmented BERT model, we switched the original test data from the unaugmented corpus with the test data used by the augmented model (see Section 5.3), i.e., 10% of the augmented corpus. Table 5.8 provides a comparison of the model's performance on the original unaugmented test set versus the augmented test set. The accuracy

of the replaced test data is very low at 59.53% compared to the model's training accuracy of 99.42%. This difference of almost 40% indicates that the model significantly fails to predict the label of a given sentence within the augmented test set in spite of performing well on the initial unaugmented test set. Figure 5.6 shows that the model predicted 17 false positives and 70 false negatives. Therefore, the BERT model trained on unaugmented corpus fails to generalize well even to data that is just a supplemented version of its training corpus.

Table 5.9: Summary of Model Performance With Different Training and Testing Datasets

Model	Training Set	Test Set	Accuracy	F1-Score	Precision	Recall
Unaugmented BERT	Unaugmented	Unaugmented	0.9266	0.9069	0.9750	0.8478
Unaugmented BERT	Unaugmented	Augmented	0.5953	0.5681	0.6252	0.5942
Unaugmented Logistic Regression	Unaugmented	Unaugmented	0.8899	0.8500	1.0000	0.7391
Unaugmented Logistic Regression	Unaugmented	Augmented	0.5814	0.3382	0.7931	0.2150
Synthetic BERT	Synthetic	Synthetic	0.9861	0.9862	0.9727	1.0000
Synthetic BERT	Synthetic	Unaugmented	0.7156	0.6984	0.7125	0.6953
Augmented BERT	Augmented	Augmented	0.9628	0.9626	0.9626	0.9626
Augmented BERT	Augmented	Unaugmented	0.9358	0.9340	0.9354	0.9327
Augmented Logistic Regression	Augmented	Augmented	0.8930	0.8900	0.9118	0.8692
Augmented Logistic Regression	Augmented	Unaugmented	0.7064	0.4839	0.9375	0.3261

The next chapter discusses insights drawn from the exploratory studies in Chapter 4 and the findings from the augmented model presented in this chapter, and reflects on what these results mean for the overall goals of the study.

Chapter 6

Discussion

This chapter looks at the impact of the results achieved by all the models and how they compare with each other, as shown in Table 5.9. The augmented BERT model outperforms all other models studied except the BERT model tested on pure synthetic data, with significantly better overall generalization on unseen data than the other models.

Though the accuracy of the first BERT model trained on the unaugmented corpus is almost 93%, it is clear from the analysis in Figure 5.3 that there are specific terms/phrases like "visually impaired" and "sign language" that may be affecting the model's learning unfairly. The obvious imbalance in the unaugmented corpus demerits such high values for the model's initial performance metrics. This suspicion is further reinforced by the results of Subsection 5.3.3.2. The model should have theoretically shown good performance on the augmented test set, as even the synthetic sentences in it were generated to be similar to the source sentences that the model was trained on. Upon closely studying the false positives, it becomes clear that most of the sentences that were predicted to be "insensitive" had terms like "visually impaired" or "able-bodied" which had no "notInsensitive" example within the unaugmented corpus that this model was trained upon. The rest of the false positives were sentences that used phrases like "special needs" to refer to the name of an organization, for example, "Special Needs Research Association." Therefore, this model was limited by the size and the type of examples available for each term in the training stage. The number of false negatives was comparatively higher (see Figure 5.6) with all of them being synthetic sentences.

The BERT model trained solely on synthetic data achieved the highest accuracy on its corresponding test set among all the models evaluated; however, this performance comes with the risk of overfitting to the synthetic corpus (see Section 5.2). This study places importance on any model's accuracy on the unaugmented test set, which consists entirely of source sentences from ASSETS abstracts—i.e., real-world examples. Despite its seemingly strong performance, the synthetic BERT model achieves only 70% accuracy on the unaugmented test set (see Table 5.9).

Therefore, it does not generalize well to new non-synthetic data, and it is possible that this model is learning unintended patterns introduced during the synthetic sentence generation process.

Meanwhile, the augmented BERT model showed consistently high metrics for both the augmented and unaugmented test sets. With a misclassification rate of only 6.42% (7 false predictions on 109 total sentences) on the unaugmented test set in Subsection 5.3.3.1, the augmented BERT model made correct predictions much more often than the unaugmented BERT model, which had a misclassification rate of 40.47% (87 of 215) in Subsection 5.3.3.2. The number of false positives (3) and false negatives (4) on the unaugmented test set is too small to identify any meaningful patterns. However, the performance of this model on data outside the academic context is very likely to be lower. The differences in language style, different contextual slang, and even the form of disability that is being discussed could be vastly different from what the model has learned from the ASSETS conference paper abstracts.

The logistic regression model trained on the augmented corpus had an accuracy that was lower than the augmented BERT model by just 6.98% on the initial augmented test set. However, in the experimental test using the unaugmented test set, the difference in accuracy increases to 22.94% (see Table 5.7), indicating a significant drop in the logistic regression model's performance. This shows that while the model performed reasonably well when evaluated on data similar to what it was trained on, its generalizability to unaugmented data was much weaker than that of the augmented BERT model.

Across all models, there is a consistent pattern in which classifiers are hesitant to predict a given sentence as "insensitive," i.e., there are more false negatives than false positives in each of the evaluations studied. This pattern is also seen in [15] that followed a similar workflow in which sentences were first annotated by humans and then used to train classifiers where their model was "biased" towards classifying sentences as less "hateful or offensive" than human annotators. This could have happened in models trained on the unaugmented corpus due to the class imbalance between the "insensitive" and "notInsensitive" sentences. Future research could explore techniques to adjust the recall-precision balance of the BERT model to favor recall so that more "insensitive" examples are identified; this is commonly done in safety-critical systems [44] where identifying all possible cases outweighs the drawback of having a few false positives.

Chapter 7

Limitations and Future Work

This chapter focuses on the limitations of this study and suggests possible modifications that future works in this domain can consider adjusting to counter these limitations.

A crucial limitation is within the foundational lexicon list itself; since the list is non-exhaustive and non-definitive, it is subject to change very quickly and is currently highly dependent on the ground truth guidelines used. The lack of consistent standards affecting terms like "hard of hearing" as explained in Subsection 3.3.1, may raise concerns about the authenticity of this manually compiled list. It would be necessary for future works to consider possible changes around the sensitivity of the terms or phrases provided by the list.

Another key limitation is within the keyword matching/extraction process, the algorithm is not resilient and may fail to catch multiple variations of certain terms within the sentences; for example, "vision impaired" will not be tagged by the current algorithm. Future works could place focus on building a better algorithm or using existing machine learning techniques to find all possible occurrences and variations of the listed terms. The study sought out data augmentation to handle the dataset imbalance problem; however, the feasibility of extracting sentences directly from the PDFs of the conference proceedings used in the study could also be tested. This may not only resolve the issue of limited data and allow the training of a better model but could also be used to reevaluate the performance of the current model.

The distribution of terms across train, validation, and test datasets in both the augmented and unaugmented corpus is not guaranteed to be proportional as the `stratify` parameter used for the splits only considered the target labels and not the 43 terms. Therefore, it is possible that the model may have had better chances of predicting a particular test set that was disproportionately made up of examples of terms that it had already learned the most from during the training stage.

There is a possibility of data leakage that may have occurred while using the entire unaugmented corpus to provide examples for the ChatGPT prompts. The augmented BERT model may have

indirectly seen some sentences from the unaugmented test set used in Subsection 5.3.3.1 through certain synthetic sentences in its training data. The impact of this limitation on the accuracy of the model is currently uncertain; however, any future work built upon this augmented corpus would need to handle this possibility first by having a set of unaugmented source sentences from the abstracts that remain out of the workflow until the evaluation stage.

There is some uncertainty regarding what the model is truly learning as it is currently unclear whether the final BERT model is making its predictions based on the actual terms or if it is just recognizing other underlying patterns. This raises concerns about the extent to which the model recognizes contextual insensitivity based on terms in comparison to other correlations that could have been introduced during training. A possible source of this issue could be the augmented data using GPT-4o generated sentences, which could have introduced unrelated patterns within the dataset that the model may be using to learn from but is not apparent right away. Given that the aim of the study is to build a model that recognizes insensitiveness based on the compiled terms and guidelines, this is a clear limitation. Future works could use Local Interpretable Model-agnostic Explanations (LIME) analysis that can provide insights as to which words in the sentence contributed more to the prediction made in order to explain better the model's decisions [19].

Research focused on solving issues around dataset imbalance in cases where there is very little training data available has recognized that the impacts of data augmentation and available techniques, such as paraphrasing or the use of LLMs such as GPT, need to be extensively studied to assess the possible risks and biases that could seep into the model [45, 46], especially when using generative models like GPT-4o that are prone to hallucinations.

Chapter 8

Conclusion

There is a research gap in the domain of toxic language detection for contexts outside social media, particularly regarding insensitive language related to disability. The lack of sufficient training examples that are representative of the diversity of disability communities is one major reason for this gap. Arango et al. [14] shows that finding a balanced dataset even within social media networks is a surprisingly difficult task such that even the popularly used dataset for toxic language detection [15] had only 4839 hateful tweets (racist or sexist), and the rest, i.e., 10,110 tweets were labeled as not hateful. The ground-truth guideline documents used in this study could act as a starting point for building a corpus dedicated to disability language alone.

Out of the five models trained and experimented upon, the augmented BERT outperforms all in terms of accuracy and generalization, achieving a test accuracy of 93% for the unaugmented corpus. However, its performance on data outside the academic domain should be evaluated with diverse real-world test sets to understand the possible biases that may have occurred due to the type of training data used, sentence-level granularity, and the LLM-based augmentation process. As discussed in Chapter 7, using generative models for data augmentation comes with a fair share of risks, like misleadingly high-performance metrics if the model learns some unintentional patterns resulting from synthetic data generation. However, the performance of this augmented BERT model is consistent with studies [46, 47] that compare the different augmentation techniques, such as back translation, word substitution, etc., and achieve comparatively better results with the GPT-based techniques. This model could act as a foundation to build a more reliable model that could be further trained on a curated corpus that extends beyond conference abstracts.

The definition of sensitive and inclusive language will inevitably change over the years; the development of more standardized guidelines built in close collaboration with people with disabilities would certainly help lay the foundation for more inclusive academic writing. However, it remains challenging to find alternative terms to those that have been in use for years, such as "visual impairments" and other terms rooted in the impairment model, particularly when there

is no consensus on what the "ideal" model of disability is. Since whether a term is considered insensitive often depends on the individuals and communities being described, their preferences must take precedence over everything else when writing about people with disabilities [3, 9, 21]. This study hopes to have set the stage to explore how we might support the scholarly community in respecting these preferences by proposing and evaluating a methodology for detecting potentially insensitive language within academic texts, using guideline-informed annotation. While our approach is not definitive, it is a step toward building tools and processes that support scholars to learn from and adapt to the dynamic nature of disability-inclusive language.

References

[1] Angela Friederici, Michael Skeide, and Verena Müller. Language characterizes humans, 2016. URL https://www.mpg.de/9982862/language-characterizes-humans. Accessed: January 2, 2025.

[2] Mike Karapita. Inclusive language in media: A Canadian style guide. *Humber*, 2017. URL https://www.humber.ca/makingaccessiblemedia/modules/01/transript/Inclusive_Language_Guide_Aug2019.pdf. Accessed: December 3, 2024.

[3] Ather Sharif, Aedan Liam McCall, and Kianna Roces Bolante. Should I say "disabled people" or "people with disabilities"? Language preferences of disabled people between identity- and person-first language. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22. ACM, 2022. DOI: https://doi.org/10.1145/3517428.3544813.

[4] ADA Knowledge Translation Center. Guidelines for writing about people with disabilities, 2017. URL https://adata.org/factsheet/ADANN-writing. Accessed: March 26, 2025.

[5] United Nations Geneva. Disability-inclusive language guidelines, 2021. URL https://www.ungeneva.org/sites/default/files/2021-01/Disability-Inclusive-Language-Guidelines.pdf. Accessed: January 2, 2025.

[6] Erin E. Andrews, Robyn M. Powell, and Kara Ayers. The evolution of disability language: Choosing terms to describe disability. *Disability and Health Journal*, 15(3):101328, 2022. ISSN 1936-6574. DOI: https://doi.org/10.1016/j.dhjo.2022.101328.

[7] Maria Berghs, Karl Atkin, Hilary Graham, Chris Hatton, and Carol Thomas. *Implications for Public Health Research of Models and Theories of Disability: A Scoping Study and Evidence Synthesis*, volume 4.8 of *Public Health Research*. NIHR Journals Library, Southampton (UK), 2016. URL https://www.ncbi.nlm.nih.gov/books/NBK378951/.

[8] Z. Zaks. Changing the medical model of disability to the normalization model of disability: Clarifying the past to create a new future direction. *Disability & Society*, 39(12):3233–3260, 2023. DOI: https://doi.org/10.1080/09687599.2023.2255926.

[9] Government of Canada. A way with words and images: Guide for communicating with and about persons with disabilities, 2024. URL https://www.canada.ca/en/employment-social-development/programs/disability/arc/words-images.html. Accessed: October 4, 2024.

[10] Disability Rights UK. Social model of disability and language, n.d. URL https://www.disabilityrightsuk.org/social-model-disability-language. Accessed: 2024-10-24.

[11] ACM SIGACCESS. Accessible writing guides. ACM SIGACCESS Resource Page, 2024. URL https://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/. Accessed: October 28, 2024.

[12] Government of British Columbia. Inclusive language and terms, 2024. URL https://www2.gov.bc.ca/gov/content/home/accessible-government/toolkit/audience-diversity/inclusive-language-and-terms. Accessed: January 2, 2025.

[13] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In Jacob Andreas, Eunsol Choi, and Angeliki Lazaridou, editors, *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California, June 2016. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N16-2013.

[14] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584, 2022. ISSN 0306-4379. DOI: https://doi.org/10.1016/j.is.2020.101584.

[15] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, 2017. DOI: https://doi.org/10.1609/icwsm.v11i1.14955.

[16] Drew Boyd. Research summary document: hatebase - AI for hate speech monitoring, November 2022. URL https://creativecommons.org/licenses/by/4.0/legalcode. Accessed: January 2, 2025.

[17] Ravi Gupta, Avanish Jha, Rudraksh Singh, Ratneshwar Kumar Bharti, and Ranjith R. From logistic regression to BERT: Benchmarking sentiment analysis models on E-commerce

data. In *2024 International Conference on Futuristic Technologies in Control Systems & Renewable Energy (ICFCR)*, pages 1–6, 2024. DOI: https://doi.org/10.1109/ICFCR64128.2024.10763143.

[18] Sindhu Abro, Sarang Shaikh, Zahid Hussain, Zafar Ali, Sajid Khan, and Ghulam Mujtaba. Automatic hate speech detection using machine learning: A comparative study. *International Journal of Advanced Computer Science and Applications*, 11, 01 2020. DOI: https://doi.org/10.14569/IJACSA.2020.0110861.

[19] Hind Saleh, Areej Alhothali, and Kawthar Moria. Detection of hate speech using bert and hate speech word embedding with deep model. *Applied Artificial Intelligence*, 37(1), February 2023. ISSN 1087-6545. DOI: https://doi.org/10.1080/08839514.2023.2166719.

[20] Walter Cronkite School of Journalism and Mass Communication Arizona State University. Disability language style guide. *National Center on Disability and Journalism*, 2021. URL https://ncdj.org/style-guide/. Accessed: January 2, 2025.

[21] Anna Cavender, Shari Trewin, and Vicki Hanson. General writing guidelines for technology and people with disabilities. *SIGACCESS Access. Comput.*, page 17–22, September 2008. ISSN 1558-2337. DOI: https://doi.org/10.1145/1452562.1452565.

[22] Vicki L. Hanson, Anna Cavender, and Shari Trewin. Writing about accessibility. *Interactions*, 22(6):62–65, October 2015. ISSN 1072-5520. DOI: https://doi.org/10.1145/2828432.

[23] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. "i wouldn't say offensive but...": Disability-centered perspectives on large language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 205–216, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. DOI: https://doi.org/10.1145/3593013.3593989.

[24] Kurtis Pykes. Fuzzy string matching in python tutorial, 2025. URL https://www.datacamp.com/tutorial/fuzzy-string-python. Accessed: 2024-10-09.

[25] William J. Mattingly. keyword-spacy: A spacy pipeline component for extracting keywords from text using cosine similarity, 2025. URL https://github.com/wjbmattingly/keyword-spacy. Accessed: 2024-10-11.

[26] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica (Zagreb)*, 22(3):276–282, 2012. DOI: https://doi.org/10.11613/BM.2012.031.

[27] A. S. Kolesnyk and N. F. Khairova. Justification for the use of cohen's kappa statistic in experimental studies of nlp and text mining. *Cybernetics and Systems Analysis*, 58(2):280–288, March 2022. ISSN 1573-8337. DOI: https://doi.org/10.1007/s10559-022-00460-3.

[28] National Association of the Deaf. Community and culture - frequently asked questions, 2025. URL https://www.nad.org/resources/american-sign-language/community-and-culture-frequently-asked-questions/. Accessed: 2024-11-17.

[29] Canadian Association of the Deaf - Association des Sourds du Canada. Terminology, 2022. URL https://cad-asc.ca/issues-positions/terminology/. Accessed: 2024-11-17.

[30] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/N19-1423.

[31] Hugging Face. Bert 101: An intuitive introduction to bert, 2023. URL https://huggingface.co/blog/bert-101. Accessed: February 4, 2025.

[32] Richa Singh, Rekha Kashyap, and Vikrant Sharma. Toxic comment analyzer using bert: A deep learning approach for toxicity detection. In *2023 Second International Conference on Informatics (ICI)*, pages 1–6, 2023. DOI: https://doi.org/10.1109/ICI60088.2023.10421672.

[33] V7 Labs. How to split your datasets into training, validation, and test sets, 2025. URL https://www.v7labs.com/blog/train-validation-test-set. Accessed: 2025-01-20.

[34] Yan Xu and Royston Goodacre. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analytical Testing*, 2(3):249–262, 2018. doi: 10.1007/s41664-018-0068-2. Epub 2018 Oct 29.

[35] Hugging Face. Bert for sequence classification - transformers documentation, 2025. URL https://huggingface.co/docs/transformers/en/model_doc/bert#transformers.BertForSequenceClassification. Accessed: February 4, 2025.

[36] 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.

[37] Daniel Godoy. Random seeds and reproducibility: Setting up your experiments in python, numpy, and pytorch, May 2022. URL https://medium.com/towards-data-science/random-seeds-and-reproducibility-933da79446e3.

[38] Xiaoyi Jiang and Chang Xu. Deep learning and machine learning with grid search to predict later occurrence of breast cancer metastasis using clinical data. *Journal of Clinical Medicine*, 11(19):5772, Sep 2022. DOI: https://doi.org/10.3390/jcm11195772.

[39] Scikit-Learn Developers. sklearn.metrics.f1score, 2024. URL https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html. Accessed: February 4, 2025.

[40] Elena Shushkevich, Mikhail Alexandrov, and John Cardiff. Improving multiclass classification of fake news using bert-based models and chatgpt-augmented data. *Inventions*, 8(5), 2023. ISSN 2411-5134. doi: 10.3390/inventions8050112. URL https://www.mdpi.com/2411-5134/8/5/112.

[41] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[42] OpenAI. Best practices for prompt engineering with the OpenAI API, 2023. URL https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-the-openai-api. Accessed: 2024-11-15.

[43] Jie Chen, Yupeng Zhang, Bingning Wang, Wayne Xin Zhao, Ji-Rong Wen, and Weipeng Chen. Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models, 2024. URL https://arxiv.org/abs/2406.12397.

[44] Scikit-learn. Precision-recall, 2025. URL https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html. Accessed: 2025-02-08.

[45] Frédéric Piedboeuf and Philippe Langlais. Is ChatGPT the ultimate data augmentation algorithm? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15606–15615, Singapore, December 2023. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2023.findings-emnlp.1044.

[46] Fahim Sufi. Generative pre-trained transformer (gpt) in research: A systematic review on data augmentation. *Information*, 15(2), 2024. ISSN 2078-2489. DOI: https://doi.org/10.3390/info15020099.

[47] Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Yihan Cao, Zihao Wu, Lin Zhao, Shaochen Xu, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Lichao Sun, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. Auggpt: Leveraging chatgpt for text data augmentation, 2023. DOI: https://doi.org/10.48550/arXiv.2302.13007.

[48] Claire Andrasegyedi and Karen Margaret Walker Chevalier. Inclusive language and clear writing, 2021. https://commission.europa.eu/.

[49] California Courts. Working with interpreters: Legal terminology glossary, 2022. URL https://www4.courts.ca.gov/partners/documents/7-terminology.pdf. Accessed: 2024-10-22.

Appendix A

Model Specifics

This appendix outlines the configuration details of the pre-trained models used throughout the study.

A.1 BERT-base Uncased

The `bert-base-uncased` model was used with the following default configurations from HuggingFace. Below are the key parameters:


```
BertConfig {
  "_attn_implementation_autoset": true,
  "_name_or_path": "bert-base-uncased",
  "architectures": ["BertForMaskedLM"],
  "attention_probs_dropout_prob": 0.1,
  "classifier_dropout": null,
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 512,
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 0,
  "position_embedding_type": "absolute",
```

```
  "transformers_version": "4.47.1",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30522
}
```


All models use the default loss function for binary classification from the BertForSequenceClassification model initialization, i.e. CrossEntropyLoss.

A.2 Dependencies for Model Training and Evaluation

The following Python libraries and versions were used during model training and evaluation:

`pandas` – 2.2.2
`scikit-learn` – 1.6.1
`huggingface_hub` – 0.29.3
`wandb` – 0.19.7
`transformers` – 4.50.0
`torch` – 2.6.0+cu124
`numpy` – 2.0.2
`seaborn` – 0.13.2
`matplotlib` – 3.10.0

A.3 System Environment (Logged via wandb)

The following system configuration was automatically recorded using **Weights & Biases (wandb)** during model training in Google Colab:

Platform: Google Colab (Free tier)
Python version: 3.11.11
Operating System: Linux-6.1.85+-x86_64-with-glibc2.35
GPU type: NVIDIA Tesla T4
GPU count: 1
CPU count: 1 physical / 2 logical cores

Appendix B

Annotation Guidebook: Detailed Examples

These are the examples provided to annotators in the annotation guidebook.

```
Detailed Examples
Example 1:
Sentence: "The software is designed to assist the hearing-impaired."
Annotated Term: "hearing-impaired"
Category: Insensitive
Reason: This term is outdated and follows the impairment model; "hard of
↪  hearing" or "Deaf" is preferred depending on the context.

Example 2:
Sentence: "The use of the term hearing-impaired is not recommended"
Annotated Terms:
"hearing-impaired" → Not_Insensitive
Note: This example is a special case of use vs mention; therefore, it is highly
↪  recommended to highlight the data row.

Example 3:
Sentence: "The tool improves accessibility for persons with disabilities."
Annotated Term: "Persons with disabilities"
Category: Not_Insensitive
Reason: This phrase aligns with inclusive language guidelines.

Example 4:
Sentence: "Many victims of stroke experience difficulties with mobility."
Annotated Terms:
```

"victims" → Insensitive (frames individuals in a deficit perspective).
"difficulties" → Insensitive (emphasizes limitations).

Appendix C

Lexicon List

Terms	Alternative	Possible classification	Source
Quadriplegics	people who use a wheelchair	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[11, 21]
The deaf	‘deaf people’ or ‘people who are deaf’	Insensitive-Noun (Alt. as sensitive-IdentityFirst vs sensitive-PeopleFirst)	[11, 21]
the blind	‘blind people’ or ‘people who are blind’	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[11, 21, 22] [9, 48]
abnormal		Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[11, 21, 22] [9, 48]
normal	persons without disabilities	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[11, 21, 22] [9, 48]
The disabled	people with disabilities	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[11, 21]
Victim of . . .	He has. . . ; she has. . . ; they have. . . ; a person with...	Insensitive (Insensitive-NegativeOvertone)	[11, 21, 22, 48]
Suffering from. . .	He has. . . ; she has. . . ; they have. . . ; a person with...	Insensitive (Insensitive-NegativeOvertone)	[11, 21]

Continued below

Terms	Alternative	Possible classification	Source
Afflicted with. . .		Insensitive (Insensitive-NegativeOvertone)	[11, 21]
Defective		Insensitive (Insensitive-Slur)	[11, 21]
Physically challenged		Insensitive (Insensitive-Patronizing)	[5, 11, 21]
Special		Insensitive (Insensitive-Patronizing)	[11, 21]
Differently abled		Insensitive (Insensitive-Patronizing)	[11, 21]
Patient		NEUTRAL (based on the definition)	[11, 21]
Sight deficient		Insensitive (Insensitive-NegativeOvertone)	[11, 21, 22]
people with sight problems or unsighted		Not Recommended (based on the definition)	[11, 21, 22]
deaf-mute	‘deaf people’ or ‘people who are deaf’	Insensitive-Slur (Alt. as sensitive-IdentityFirst vs sensitive-PeopleFirst)	[11, 21, 22]
deaf and dumb	‘deaf people’ or ‘people who are deaf’	Insensitive-Slur (Alt. as sensitive-IdentityFirst vs sensitive-PeopleFirst)	[11, 21, 22]
restricted to a wheelchair	‘person who uses a wheelchair’ or ‘wheelchair user’	Insensitive-NegativeOvertone (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[11, 21, 22]
confined to a wheelchair	‘person who uses a wheelchair’ or ‘wheelchair user’	Insensitive-NegativeOvertone (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[11, 21, 22, 48]
wheelchair-bound	‘person who uses a wheelchair’ or ‘wheelchair user’	Insensitive-NegativeOvertone (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[11, 21, 22]

Continued below

Terms	Alternative	Possible classification	Source
deformed	'person who uses a wheelchair' or 'wheelchair user'	Insensitive-Slur (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[9, 11, 21, 22]
crippled	'person who uses a wheelchair' or 'wheelchair user'	Insensitive-Slur (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[9, 11, 21, 22]
Retarded	Person with a psychiatric disability	Insensitive (Insensitive-Slur)	[11, 21, 22]
Deranged		Insensitive (Insensitive-Slur)	[11, 21, 22]
Deviant		Insensitive (Insensitive-NegativeOvertone)	[11, 21, 22]
Demented		Insensitive (Insensitive-Slur)	[11, 21, 22]
Deficient		Insensitive-NegativeOvertone [A gray area as "visual processing deficit" is a medical term]	[11, 21, 22]
People with deficits		Insensitive (Insensitive-NegativeOvertone)	[11, 21, 22]
Insane	Person with a psychiatric disability	Insensitive (Insensitive-Slur)	[9, 11, 21, 22]
Slow or slow learner		Insensitive-NegativeOvertone	[11, 21, 22]
Mad		Insensitive (Insensitive-Slur)	[11, 21, 22]
Crazy	Person with a psychiatric disability	Insensitive (Insensitive-Slur)	[9, 11, 21, 22]
Paranoid		Insensitive (Insensitive-Slur)	[11, 21, 22]
mongoloid	person or child with Down Syndrome	Insensitive (Insensitive-Slur)	[11, 21, 22]

Continued below

Terms	Alternative	Possible classification	Source
able-bodied	persons without disabilities	Insensitive (Insensitive-NegativeOvertone)	[9]
Special ed		Insensitive (Insensitive-Patronizing)	[11, 21, 22]
Birth defect	Person born with a disability	Insensitive (Insensitive-NegativeOvertone)	[9]
Congenital defect	Person born with a disability	Insensitive (Insensitive-NegativeOvertone)	[9]
Handicapped	Person with a disability	Insensitive (Insensitive-NegativeOvertone)	[9, 48]
hunchback			[20]
vegetable			[20]
homebound			[49]
paraplegic			[20, 49]
Epileptic	Person with epilepsy	Insensitive (Alt. as sensitive-PeopleFirst)	[9, 48]
Fit	Seizure		[9]
Attack	Seizure		[9]
Spell	Seizure		[9]
Caretaker	Supporter; support person; attendant		[9]
Hard of hearing	'people who are deaf' or 'deaf people'	Insensitive-NegativeOvertone (Alt. as sensitive-PeopleFirst vs sensitive-IdentityFirst)	[9]
High-functioning autistic	Autistic person with low support needs	Insensitive-Noun	[9]
Low-functioning autistic	Autistic person with high support needs	Insensitive-Noun	[9]
hearing-impaired		Insensitive (Focuses on impairment rather than disability)	[9]

Continued below

Terms	Alternative	Possible classification	Source
Inarticulate	Person with a communication disability	Insensitive (Focuses on impairment rather than disability)	[9]
impairments		Insensitive (Focuses on impairment rather than disability)	[9]
Incoherent	Person with a communication disability	Insensitive (Focuses on impairment rather than disability)	[9]
Invalid	Person with a disability	Insensitive-NegativeOvertone	[9]
Learning disabled	Person with a learning disability	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[9]
disorder	Person with a learning disability	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[9]
dyslexic	Person with a learning disability (e.g., person with dyslexia)	Insensitive-Noun(Alt. as sensitive-PeopleFirst)	[9]
Living with a disability	Person with a disability	Neutral (Common PeopleFirst language)	[9]
Living with autism	Autistic person	Insensitive (Insensitive-NegativeOvertone)	[9]
mental	Person with a mental health disability	Insensitive (Negative tone)	[9]
Mentally ill	Person with a mental health disability	Insensitive (Negative tone)	[9]
Schizophrenic	Person with a psychiatric disability	Insensitive-Noun(Alt. as sensitive-PeopleFirst)	[9]
Psychotic	Person with a psychosocial disability	Insensitive-Noun (Alt. as sensitive-PeopleFirst)	[9]
Lunatic	Person with a psychiatric disability	Insensitive-Slur	[9]

Continued below

Terms	Alternative	Possible classification	Source
Neurotic	Person with a psychosocial disability	Insensitive-Slur	[9]
Psycho	[5, 9]	Insensitive-Slur	[9]
challenged	Person with an intellectual disability	Insensitive-Patronizing	[9]
Mentally challenged	Person with an intellectual disability	Insensitive-Patronizing	[9]
Developmentally delayed	Person with a developmental or cognitive disability	Insensitive-Noun	[9]
Midget	Person of short stature	Insensitive-Slur (Derogatory and diminutive)	[9]
Dwarf	Person of short stature	Insensitive-Slur (Derogatory when not used medically)	[9]
Non-verbal	Person who does not use words or signs; person who uses alternate communication technology (if applicable)	Insensitive-Noun	[5, 9]
Non-communicative	Person who does not use words or signs; person who uses alternate communication technology (if applicable)	Insensitive-Noun	[9]
special needs	Accommodation requirements; person who requires supports	Insensitive -Patronizing	[5, 9]
Profoundly disabled	Person with a high need for support	Insensitive-NegativeOvertone	[9]

Continued below

Terms	Alternative	Possible classification	Source
Severely disabled	Person with a high need for support	Insensitive-NegativeOvertone	[9]
Seeing eye dogs	Guide dogs; dog guides	Neutral (Specific language correction)	[9]
Blind dogs	Guide dogs; dog guides	Neutral (Specific language correction)	[9]
Dogs for the blind	Guide dogs; dog guides	Neutral (Specific language correction)	[9]
sign language	American Sign Language; langue des signes québécoise; Indigenous sign language (including Plains Sign Language; Plateau Sign Language; Inuit Sign Language)	Neutral (Specificity preferred)	[9]
Spastic	Person who has spasms	Insensitive-Noun (Neutral when Medical condition)	[9]
Stricken with...	He has. . . ; she has. . . ; they have. . . ; a person with...	Insensitive (Insensitive-NegativeOvertone)	[9]
iliterate		Insensitive (Insensitive-NegativeOvertone)	[9]
Afflicted by...	He has. . . ; she has. . . ; they have. . . ; a person with...	Insensitive (Insensitive-NegativeOvertone)	[9]
sight- impaired		Insensitive	[2]
Developmentally Impaired		Insensitive	[2]

Continued below

Terms	Alternative	Possible classification	Source
visually impaired		Insensitive	[2, 9]
Bound		Insensitive	[2]
crip		Insensitive	[2]
Handicapable		Insensitive	[5]
Atypical		Insensitive	[5]
Person living with a disability		Insensitive	[5]
People of all abilities		Insensitive	[5]
People of determination		Insensitive	[5]
Normal		Insensitive	[5]
Healthy		Insensitive	[5]
typical		Insensitive	[5]
Whole		Insensitive	[5]
Of sound body		Insensitive	[5]
Troubled with		Insensitive	[5]
simple		Insensitive	[5]
Slow		Insensitive	[5]
Afflicted		Insensitive	[5]
Of unsound mind		Insensitive	[5]
Maniac		Insensitive	[5]
Hypersensitive		Insensitive	[5]
Panicked		Insensitive	[5]
Agitated		Insensitive	[5]
Subnormal		Insensitive	[5]
Partially-sighted		Insensitive	[5]
Lame		Insensitive	[5]
Person with physical limitations		Insensitive	[5]
Limp		Insensitive	[5]
Confined/restricted to a wheelchair		Insensitive	[5]
Stunted		Insensitive	[5]
Special person		Insensitive	[5]

Continued below

Terms	Alternative	Possible classification	Source
Leper		Insensitive	[5]
Leprosy patient		Insensitive	[5]
Can't talk		Insensitive	[5]
Handicapped bathroom		Insensitive	[5]
Of sound mind		Insensitive	[5]
person with disability		notInsensitive	[5]
persons with disabilities		notInsensitive	[5]
Person with an intellectual disability		notInsensitive	[5]
Person with an intellectual impairment		notInsensitive	[5]
Person with a psychosocial disability		notInsensitive	[5]
deaf person		notInsensitive	[5]
Person who is deaf		notInsensitive	[5]
Person with a hearing disability		notInsensitive	[5]
Person with a hearing impairment		notInsensitive	[5]
Person with hearing loss		notInsensitive	[5]
Hard-of-hearing person		notInsensitive	[5]
Deafblind person		notInsensitive	[5]
blind person		notInsensitive	[5]
Person who is blind		notInsensitive	[5]
Person with a vision/visual disability		notInsensitive	[5]

Continued below

Terms	Alternative	Possible classification	Source
Person with a vision/visual impairment		notInsensitive	[5]
Person with low vision		notInsensitive	[5]
Person with a physical disability		notInsensitive	[5]
Person with a physical impairment		notInsensitive	[5]
Wheelchair user		notInsensitive	[5]
Person who uses a wheelchair		notInsensitive	[5]
Person with a mobility disability		notInsensitive	[5]
Person with a mobility impairment		notInsensitive	[5]
Person using a mobility device		notInsensitive	[5]
Person of short stature		notInsensitive	[5]
Little person		notInsensitive	[5]
Person with achondroplasia		notInsensitive	[5]
Person with Down syndrome		notInsensitive	[5]
Person with trisomy-21		notInsensitive	[5]
Person with albinism		notInsensitive	[5]
Person affected by leprosy		notInsensitive	[5]

Continued below

Terms	Alternative	Possible classification	Source
Person who uses a communication device		notInsensitive	[5]
Person who uses an alternative method of communication		notInsensitive	[5]
Accessible parking		notInsensitive	[5]
Parking reserved for persons with disabilities		notInsensitive	[5]
Accessible bathroom		notInsensitive	[5]