UNIVERSITY OF PRINCE EDWARD ISLAND

Identifying Insensitive Language about Disabled People Using Semantic Analysis and 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

February 2025

Declaration of Authorship

I, Roshna Roby, declare that this thesis titled, 'Identifying Insensitive Language about Disabled People Using Semantic Analysis and 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

TO DO →

Acknowledgements

This research was made possible by the assistance and through the support of my supervisors, Dr. Christopher Power and Dr. Paul Sheridan.

Contents

Declaration of Authorship i

Abstract ii

Acknowledgements iii

List of Figures vi

List of Tables vii

1 Introduction 1
- 1.1 Inclusive Language . . . 1
- 1.2 Purpose and research goals . . . 2

2 Background Information 4
- 2.1 Language around Disabilities and Disabled Persons . . . 4
  - 2.1.1 Different Models of Language . . . 4
- 2.2 Toxic Language detection . . . 4

3 Corpora Compilation and Annotation 5
- 3.1 Generation of the Lexicon List . . . 5
- 3.2 Automated Extraction and Initial Annotation . . . 7
  - 3.2.1 Abstract Compilation . . . 7
  - 3.2.2 Keyword Extraction Algorithm . . . 8
- 3.3 Manual Annotation and Dataset Refinement . . . 9
  - 3.3.1 Inter-annotator Reliability . . . 9

4 Experimental Setup 11
- 4.1 Annotation and Analysis Pipeline . . . 11
- 4.2 Dataset Analysis . . . 12
- 4.3 Data Augmentation . . . 13
  - 4.3.1 Step I: Random Selection of an Existing Example . . . 15
  - 4.3.2 Step II: Customizing Prompts and Sentence Generation . . . 15
  - 4.3.3 Step III: The Augmented Dataset - Combination of ChatGPT_generated and Abstracts Datasets . . . 15
- 4.4 Model Training . . . 17

4.4.1 Training on the Original Abstracts dataset 17
4.4.2 Training on the Augmented dataset 17

5 Results **18**
5.1 Result I: Models trained on the Unaugmented data 18
5.2 Result II: Model trained on Purely ChatGPT generated data 19
5.3 Result III: Models trained on Augmented data 20

6 Discussion **21**

7 Limitations and Future Work **22**

8 Conclusion **23**

Appendix A Model Specifics **27**
A.1 Bert Base Uncased 27

List of Figures

4.1 **The proposed pipeline to move from corpus and guideline documents to create an annotated dataset that can be used to train the baseline model and fine-tune Bert** 11

4.2 **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.** 12

4.3 **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.** 13

4.4 **Need to try adjusting too big→The bar chart illustrates the frequency of sentences for each term found within the Abstract dataset with the "Insensitive" and "Not Insensitive" categories side by side.** 14

4.5 **The bar chart illustrates the distribution of sentences with the "Insensitive" and "Not Insensitive" categories side by side. Each category shows contributions from the original Abstract dataset and ChatGPT, represented as stacked bars. The x-axis lists the terms, while the y-axis indicates the number of sentences.** 16

List of Tables

5.1 Performance Metrics of the Model Trained on Unaugmented Data 19
5.2 **Loss and Performance Trends Across Training Epochs** 19
5.3 Performance Metrics of the Model Trained on Pure GPT-4o Generated Data . . 19
5.4 Loss and Performance Trends Across Training Epochs (Pure GPT-4o Data) . . 19
5.5 Performance Metrics of the Model Trained on Augmented Data 20
5.6 Loss and Performance Trends Across Training Epochs (Augmented Data) . . . 20

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 underlines human language, leading 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, fostering a shift in the social outlook towards every individual. Inclusive language is also a clear indicator of the commitment to learning and changing marginalizing perspectives, cultural beliefs, and stereotypes [2, 3] by organizations or even individuals alone; hence, it plays a very important role in propagating respect across communities.

1.1 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, but new meanings may even arise within short time periods, even just 10 years [3]. 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 [3, 4] about the negative undertones and marginalizing nature of certain terms.

2. Individual preferences about how one wishes to be referred to may vary even within the same community. The Employment Canada Guide [5] 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 [4], 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 [5, 6]. 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 [6] and Canada [5]. This shows the lack of a universal standard that can be followed across contexts [5].

Inclusive language, as defined by the Government of British Columbia [7], emphasizes that it is a commitment to "using language that's free from prejudice, stereotypes, or discriminatory views of specific people or groups". Such an intentional commitment ensures that the different individuals in society are not purposefully or unconsciously excluded, disrespected, or marginalized [5][6]. It is important to note that inclusivity, as discussed throughout this paper, is studied and referenced only within the context of the English language.

1.2 Purpose and research goals

This research project studies and analyzes the different terms used to refer to disability 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 ADA National Network [8] and UN-specified guidelines [9].

The goals of this study are:

1. Compile a non-exhaustive list of guidelines and lexicons around insensitivity in disability-related language, specifically within written communication.

2. Extract sentences that use terms or phrases referred to by the 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 in using GPT-4o supported data augmentation where the dataset lacks sufficient examples.

4. Train a BERT model to recognize insensitivity in language used within academia using the compiled list of guidelines.

This study compiles 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 dataset that contains sentences extracted from the abstracts of all ASSETS conferences over the years - a body that places importance on accessibility and inclusivity. By manually compiling and annotating the dataset, this thesis aims to show the evolution of disability-related terms within the research context. Given the binary classification nature of the study, dataset 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 pre-trained BERT-base. This helps to evaluate the integrity of the dataset based on whether it will be able to generalize to new data or just overfit the existing sentences. **TO DO →** The dataset was found to be unbalanced; therefore, this thesis also describes a few methods taken to balance and augment the data. It compares the results from the initial models with those from the models trained on both fully synthetic ChatGPT-4o generated data and augmented data. We compare a total of five models:

- Two logistic regressors - one trained on unaugmentated data based on sentences extracted from ASSETS abstracts and another trained on augmented data (see section 4.3).

- Three BERT models - one for unagumented data, one for synthetic data generated by ChatGPT-4o through one-shot prompting, and finally, a model trained on augmented data

Future works could consider placing a focus on extending the initial lexicon list of 250 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 will continuously have more diverse data to learn from and could help handle the class imbalances within the current data set. Therefore, the BERT model's ability to generalize to different contexts can be improved significantly, which could ultimately be used as a foundation to build a more resilient and rombust model to find insensitivity and inclusivity in written content across domains.

TO DO → The rest of this study is set up as follows. Chapter 2

Chapter 2

Background Information

2.1 Language around Disabilities and Disabled Persons

TO DO →

2.1.1 Different Models of Language

Evolution,individual preferences, models

-
-

2.2 Toxic Language detection

TO DO → Existing research around toxic language detection

Chapter 3

Corpora Compilation and Annotation

3.1 Generation of the Lexicon List

To process sentences for insensitivity, there is a need to define concretely what words can be categorized as "insensitive" or, in the case of this study, "notInsensitive." [TO DO: Refer to other hate speech documents here] 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 used to refer to all terms that can be used within contexts that are neutral, acceptable for medical use, or reflective of people-first or identity-first approaches.

The first step of this study was to accumulate a list of words 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 and Employment Canada [5, 7, 9] and non-official written language recommendations [10]. 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 an Excel sheet, 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 excel lexicon list makes note 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 ultimately patronizing [TO DO quote stella Young] 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 [6, 11, 12].

The guidelines investigated also included some alternate language at times to be used for certain terms or phrases they defined as insensitive; the lexicon excel makes note of these alternatives, labeling them as “notInsensitive”. This list of lexicons, i.e., words and phrases, was then generalized just to being either insensitive or not insensitive - this Excel sheet acts as the Lexicon list throughout the experiment. The study of the guideline documents identified some 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 [2, 5, 6, 11–13].
2. Avoid using “living with a disability” because that can sound like the disability is a negative burden [5].
3. Avoid language that equates persons with their condition (e.g., epileptics, the deaf) [2, 6, 11–13].
4. Avoid language that has unnecessary negative overtones or projected feelings (e.g., stroke victim or sufferer) [6, 11, 12].
5. Avoid language that is regarded as a slur (e.g., cripple) [6].
6. Avoid trendy euphemisms like physically challenged, special, differently-abled, and handicapable, as they can be patronizing [6, 11, 12].
7. Use proper definitions of terms like “impairment,” “disability,” and “handicap” based on contextual differences[6, 11, 12].
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 [6, 11].
9. “Hearing impaired” and similar impairment language should only be reserved for medical writing [6, 11].
10. The term “Hard of hearing’ should be reserved for those not identifying with the deaf community [6, 11].
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” [6, 11].

12. Be concise and specific when referring to cognitive disabilities. For example, specific cognitive disability such as language processing difficulties [6, 11].

13. Use "person with a mental disability" to refer to different forms of clinical illnesses such as schizophrenia, depression, and emotional disorders [6, 11].

14. Use terms such as "neurotic" and "psychotic" only for clinical writing [6, 11].

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 [6, 11, 12].

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 [5, 12].

17. When referencing accommodations, place focus on the term "accessible" rather than the disability, i.e., "accessible parking" vs. "handicapped parking" [5].

18. Instead of "Person who has trouble/difficulties/challenges," use "Person who needs..." [5].

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? [6, 11].

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 [6, 11].

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).

3.2 Automated Extraction and Initial Annotation

A human-in-the-loop [TO DO define] style was tried to annotate sentences from ASSETS abstracts.

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 of 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 138 while 1994 had the lowest with only 22 proceedings.

3.2.2 Keyword Extraction Algorithm

An initial sample of 10 papers was used to test for a suitable rule-based keyword-matching approach to potentially carryout 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 [TO DO describe fuzzy matching and the algorithm] 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 [14] was very briefly explored to extract the sentences from abstracts; however, it captured significantly fewer sentences than the combined approach described above, i.e., it failed to extract certain sentences using terms/phrases from the Lexicon list. With access to more computational resources; a detailed comparison between the rule-based combined 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 an Excel sheet 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 Abstracts dataset throughout the rest of the study.

3.3 Manual Annotation and Dataset Refinement

A subset of 200 sentences - about 18.3 percent, was carefully selected from the Abstracts dataset 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 ”insensitve” or ”notInsensitive.” Detailed examples provided to the annotators can be found in Appendix A. The annotators separately labeled this subset within their respective Google sheets, 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 i.e.. [TO DO go to notes]

When terms that are explicitly marked as ”insensitive” are mentioned rather than used within a sentence 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” [15]. During the review meeting, the following discrepancies were discovered between annotator A 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 labels 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 case without more context. For the scope of this thesis, it was decided that the sentences

will be labelled as a standalone components following a sentence level granularity for the semantic 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 following the guideline ” [TO DO]” mentioned in the article while Annotator B used the recent ECG 2024 [5] 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 guideline 2024 is fairly recent and multiple sources still seem to accept ”hard of hearing” [TO DO]. Once this misunderstanding was resolved across Annotator B's labels; the final Cohen's κ reached 0.8223 with 16 mismatched records from an initial value of 0.6138.

The mismatch due to term hard-of-hearing is a clear example of how the change in language over time often makes it difficult to devise proper standards and guidelines.

Chapter 4

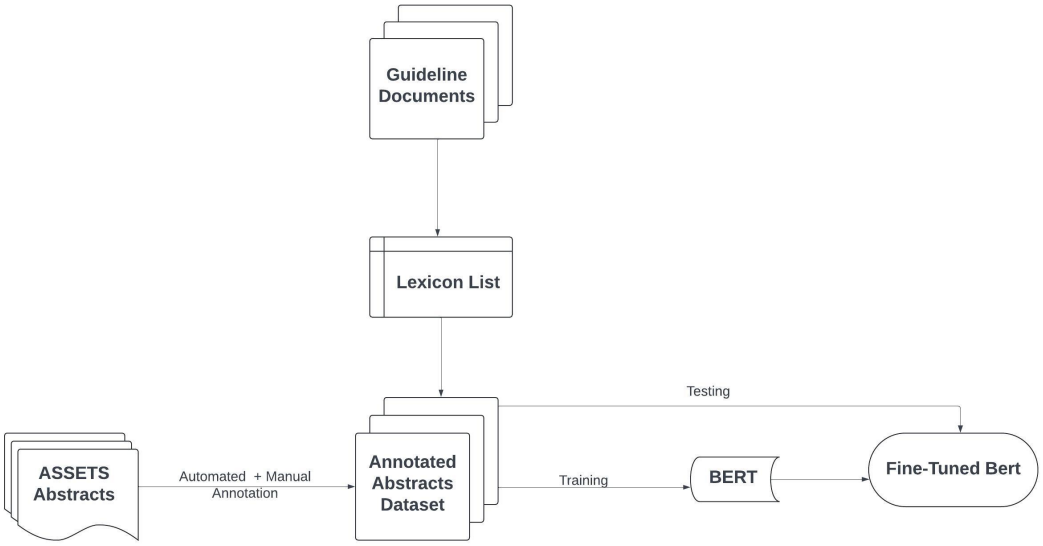
Experimental Setup

This chapter describes the preparation of the original dataset, the data augmentation process, and the classifier model selection and training processes in detail.

All tests were run on Google Colab using Python 3.10. Further details regarding the exact setup and example commands necessary to replicate results are described in Appendix [TO DO]

4.1 Annotation and Analysis Pipeline

Figure 4.1: The proposed pipeline to move from corpus and guideline documents to create an annotated dataset that can be used to train the baseline model and fine-tune Bert



4.2 Dataset Analysis

The Abstracts data set was analyzed to prepare for data augmentation and model training. We removed all duplicate records before running the analysis below.

Figure 4.2 visualizes the 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 as these years did not hold an ASSETS conference and hence, did not have any abstracts to extract sentences from. In the later years i.e 2019 onwards there seems to be an increase in the number of sentences that were annotated as "notInsensitive." Though it seems like the number of "insensitive" annotations are decreasing gradually from 2017 after an initially increasing trend throughout the previous years, there seems to be an upwards trend again from 2021 onwards. Figure 4.2 suggests 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 interpretead without Figure 4.3. Figure 4.3 normalizes the count of annotations based on the total number of sentences within each abstract of every conference paper per year.

Figure 4.2: 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.

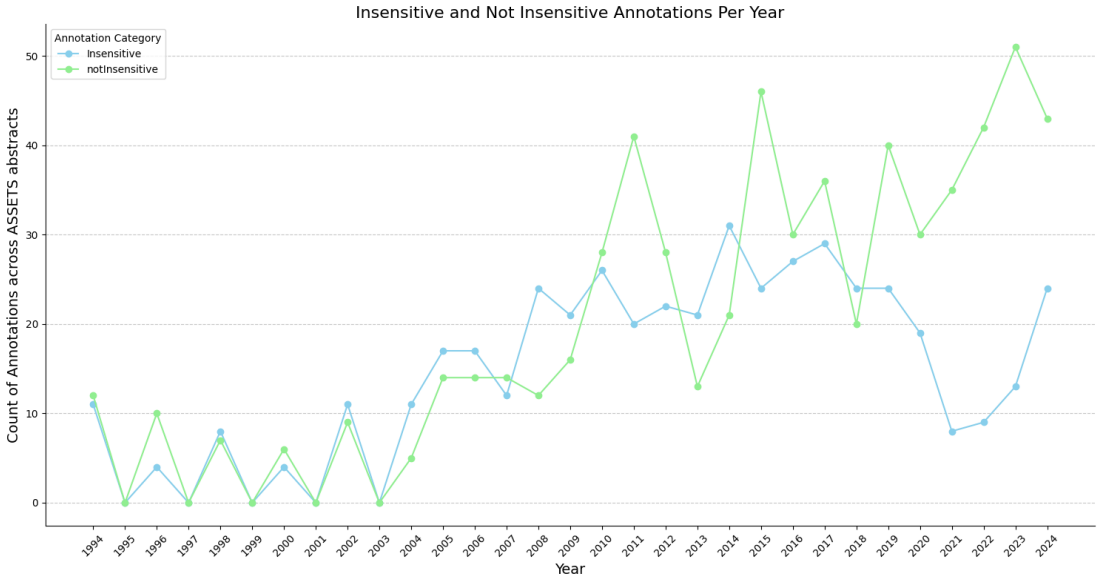
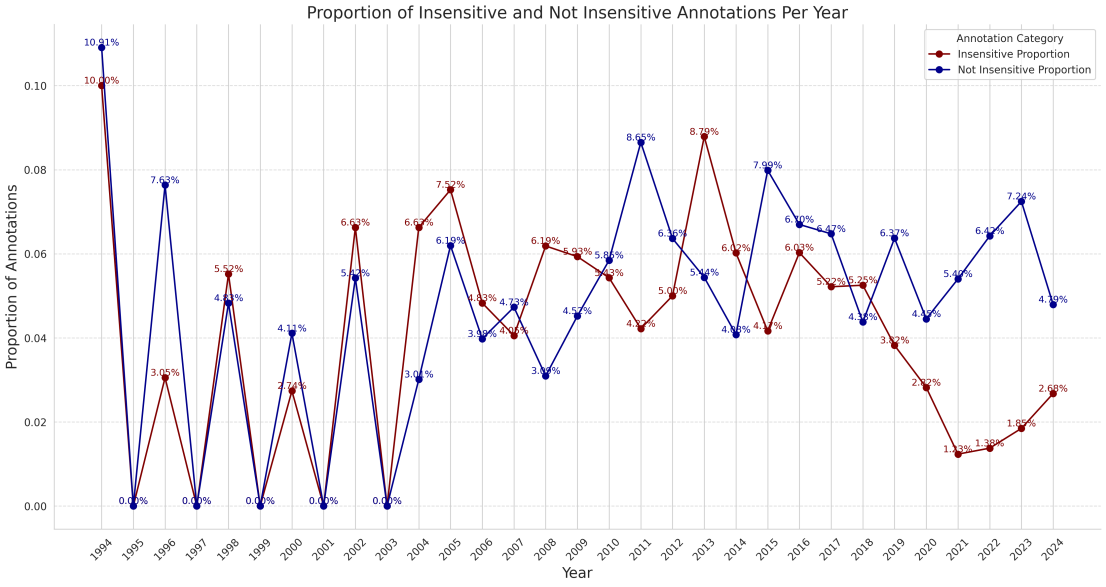


Figure 4.3: Yearly proportion of sentences annotated as "Insensitive" and "Not Insensitive" in ASSETS conference paper abstracts. The chart normalizes annotation counts by the total number of sentences within abstracts per year.



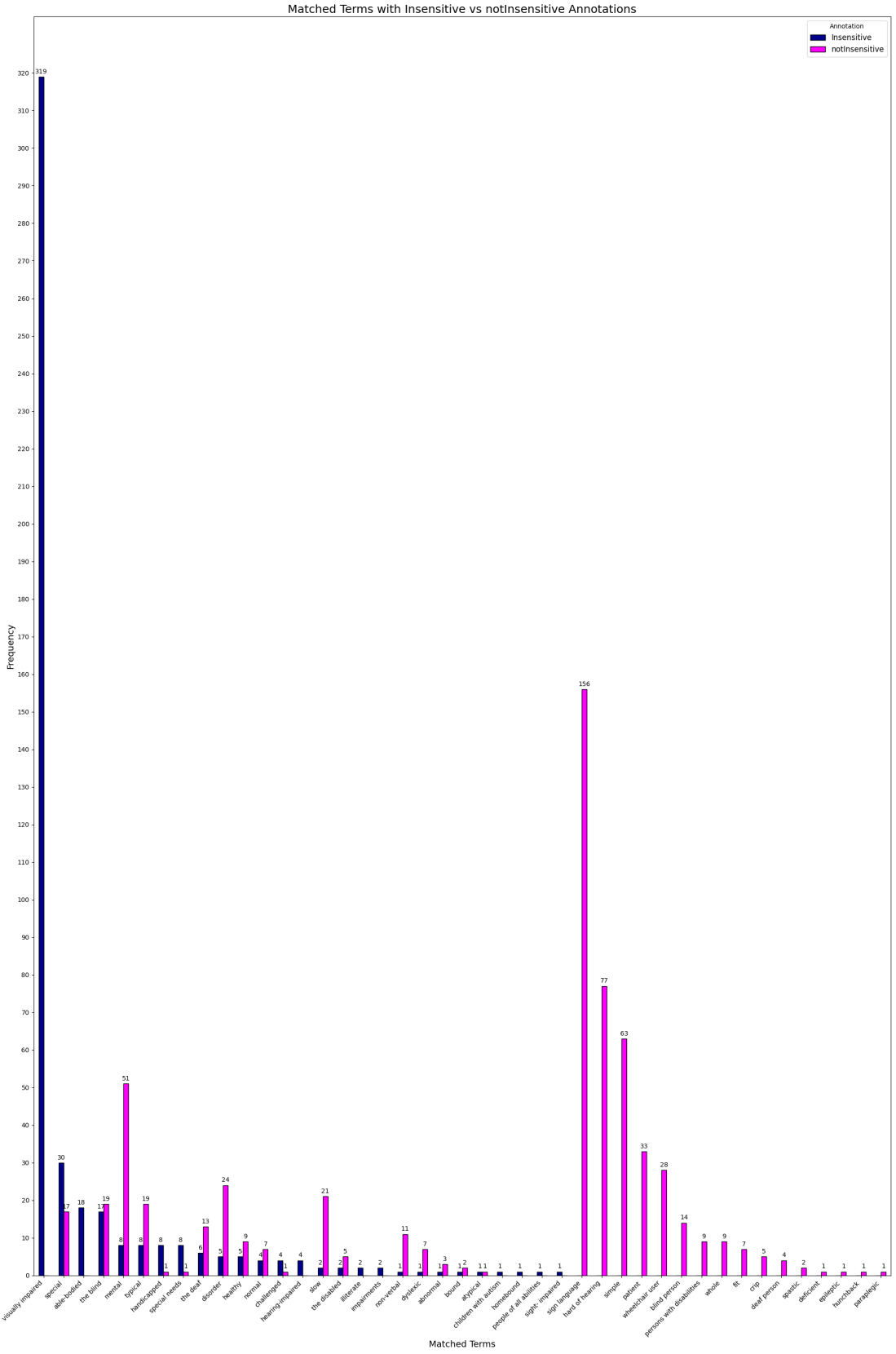
4.3 Data Augmentation

Given the obvious imbalance in the Abstracts dataset as analyzed in section 4.2, 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. For the purposes of this study and to ensure minimal use of resources and computational power, the data set was augmented to 50 sentences for each of the 43 terms or phrases that existed in the Abstracts dataset.

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 Abstracts dataset would mean that all the "insensitive" examples in the Augmented dataset 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 work could experiment by varying this number and comparing model performance for each. The final Augmentated dataset 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:

TO DO diagram for generation

Figure 4.4: Need to try adjusting too big→The bar chart illustrates the frequency of sentences for each term found within the Abstract dataset with the "Insensitive" and "Not Insensitive" categories side by side.



4.3.1 Step I: Random Selection of an Existing Example

The original Abstract dataset was used to select one sentence per term from the "insensitive" category ans another from the "notInsensitive" category at random. TO DO Explain one shot learning here

4.3.2 Step II: Customizing Prompts and Sentence Generation

Below is the general prompt structure followed for every term recorded in the Abstracts dataset.

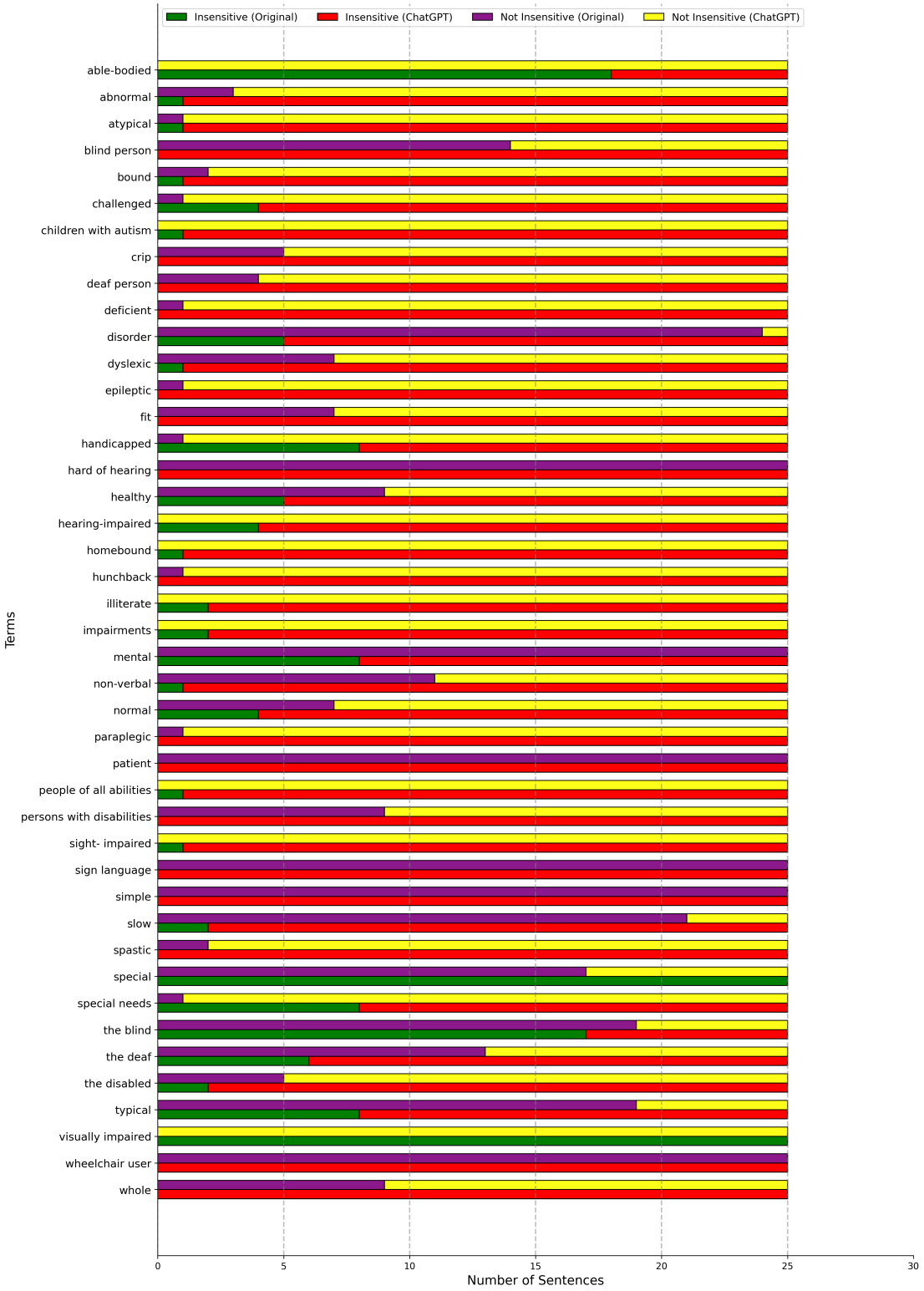
> Below are examples labeled as "Insensitive" and "Not Insensitive" that use the term ⟨Term⟩:
>
> **Special Note (Optional):**
> ⟨Provide any other information specific to the given term⟩
>
> **Insensitive example:**
> ⟨Insensitive example selected at random from the Abstracts dataset⟩
>
> **Not Insensitive example:**
> ⟨Not Insensitive example selected at random from the Abstracts dataset⟩
>
> 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.

Every generated example was manually verified and re-annotated using the same annotation codebook used for annotating the Abstracts dataset before being added to an Excel sheet named ChatGPT_generated with columns Term, Prompt, Sentence, and Label.

4.3.3 Step III: The Augmented Dataset - Combination of ChatGPT_generated and Abstracts Datasets

Figure 4.5 visualizes the contributions from each term TO DO

Figure 4.5: The bar chart illustrates the distribution of sentences with the "Insensitive" and "Not Insensitive" categories side by side. Each category shows contributions from the original Abstract dataset and ChatGPT, represented as stacked bars. The x-axis lists the terms, while the y-axis indicates the number of sentences.



already have compiled state of art for why BERT State of ART

4.4 Model Training

In order to build a model that is able to recognize "insensitive" terms based on the compiled datasets, we fine-tune BERT(Bidirectional Encoder Representations from Transformers). Developed in 2018 by researchers at Google AI Language, BERT is known as the state-of-the-art model for Natural Language Processing tasks [16, 17]. The original BERT paper [16] explains . In this study, we fine-tune BERT-base with parameters, Both the original Abstracts dataset and the augmented dataset were used to train logistic regressions that can be used to act as baseline classifier models to compare the BERT models against. BERT base param The study uses the Scikit-learn package [18] for both preprocessing and training algorithms. For ease of reproducibility, each stage of data processing and model training and testing uses 42 - a common random seed [19]. It is important to note that 42 was also 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 : "insensitive" and "notInsensitive."

4.4.1 Training on the Original Abstracts dataset

Before fine-tuning on Bert base, we did some preprocessing on the Abstracts dataset 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 default hyperparameters and model configurations from were used (See Appendix A for exact configurations).

4.4.2 Training on the Augmented dataset

Chapter 5

Results

This chapter outlines the results from training and evaluating the performances of the five models based on the unaugmented Abstracts dataset, ChatGPT-4o generated dataset, and finally, the augmented dataset (a combination of the two). The standard classification metrics from Scikit-learn, mainly accuracy, precision, F1 Score, and recall, are being used to evaluate each model. The standard cross-entropy loss function for binary classification is used to calculate the test and validation losses for both Bert and logistic regression classifiers below.

5.1 Result I: Models trained on the Unaugmented data

The BERT base model fine-tuned on the original Abstracts dataset achieved a test accuracy of 92.59 percent (see Table 5.1). However, the accuracy of the validation set was comparatively higher at 97.22 percent, bringing the difference in accuracies between the test and training set to 4.63 percent. Therefore, the model seems to be generalizing well to the training set but is struggling to predict unseen data from the test set - this is an indicator of potential overfitting to training data. The test precision is 0.9750, so it rarely has false positives where "nonsensitive" sentences are misclassified as "insensitive" [20]; in fact, the model 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. The imbalance between precision and recall explains the test F1-score of 0.90 (see Table 5.1). An ideal classifier would be expected to have a F1-score of 1 [21]; therefore, though this unaugmented Abstracts Bert model indicates somewhat positive metrics, it may still fail to identify a considerable amount of "insensitive" sentences.

Table 5.1: Performance Metrics of the Model Trained on Unaugmented Data

Dataset	Accuracy	F1 Score	Precision	Recall
Training	99.42%	0.9932	0.9919	0.9946
Validation	97.22%	0.9670	0.9778	0.9565
Test	92.66%	0.9069	0.9750	0.8478

Table 5.2 displays the loss trends across the three epochs that the model was trained for. In the first epoch accuracy was 0.9259 and ultimately reached 0.9722 in the third epoch. The training loss started at 0.2412 and dropped to 0.0405 by the third epoch, while the validation loss seems to be having a more gradual decline as it starts at 0.1693 and reaches just 0.1191 by the third epoch;.

Table 5.2: 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

The logistic regression model, trained on the same training, validation, and test sets, achieved

5.2 Result II: Model trained on Purely ChatGPT generated data

Table 5.3: Performance Metrics of the Model Trained on Pure GPT-4o Generated Data

Dataset	Accuracy	F1 Score	Precision	Recall
Training	100%	1.0000	1.0000	1.0000
Validation	97.67%	0.9772	0.9640	0.9907
Test	98.61%	0.9862	0.9727	1.0000

The BERT base model fine-tuned on the purely ChatGPT-generated data achieved a test accuracy of 98.61 percent (see Table 5.3). The accuracy of the validation set followed closely behind at 97.67 percent. The test precision is 0.9727, so it also rarely has false positives; three sentences in total were wrongly tagged as “insensitive”. Table 5.3 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, i.e., no false negatives. A perfect recall could also indicate that the model may be just memorizing the training set or that the training and testing sets could have been very similar to each other. Table 5.4 shows clear signs of overfitting as the validation loss is increasing while

Table 5.4: 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

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 - this is somewhat expected behavior when training on pure synthetic data [22].

5.3 Result III: Models trained on Augmented data

Table 5.5: Performance Metrics of the Model Trained on Augmented Data

Dataset	Accuracy	F1 Score	Precision	Recall
Training	99.42%	0.9942	0.9942	0.9942
Validation	94.88%	0.9502	0.9292	0.9722
Test	98.61%	0.9862	0.9727	1.0000

Note that in Table 5.5, the validation recall is lower at 0.97.

Table 5.6: Loss and Performance Trends Across Training Epochs (Augmented Data)

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1 Score
1	0.3105	0.2089	0.9302	0.9266	0.9352	0.9309
2	0.1166	0.2117	0.9395	0.9130	0.9722	0.9417
3	0.0595	0.1247	0.9488	0.9292	0.9722	0.9502

Chapter 6

Discussion

This chapter looks at the impact of the results achieved by all the models and how they compare.

For the first model trained on the unaugmented Abstract dataset, though its accuracy is almost 93 percent, it is clear from the analysis in Figure 4.5 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 Abstract dataset demerits such high values for the model's performance metrics. A drop in performance by almost 5 percent from the validation set to the test set indicates that the model could not generalize well on new data in the test set as described in section 5.1. Therefore, the model shows signs of onset of overfitting that can also be observed from **??** where the validation loss drops first till the second epoch and then seems to plateau without dropping significantly; meanwhile, the training loss continuous to drop considerably across epochs. The intensity of overfitting could be evaluated in future works by training this model for more epochs.

Example of fp. When the model does predict a sentence as "insensitive" it is very confident [20] but many incentive annotations do not get predicted at all.

Chapter 7

Limitations and Future Work

BERT keyword ext SHAP
unanswered question of alternatives around impairment model non exhaustive list. REAL world validation. try diff augmentation sizes more subcategories of insnsitveness. train on more epochs. keyword matching algotithm needs to be modified. Auto gen examples using gpt API

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 SHAP (SHapley Additive exPlanations) analysis to explain the model's decisions better. [TODO]

Chapter 8

Conclusion

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] 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, October 2022. doi: 10.1145/3517428.3544813.

[3] Hidden Disabilities. Disability-inclusive language – getting it right, n.d. URL https://hdsunflower.com/insights/post/disability-inclusive-language-%E2%80%93-getting-it-right. Accessed: January 2, 2025.

[4] 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. In consultation with Dr. Chelsea Jones (Ryerson University), Fran Odette (George Brown College), and Heather Willis (Ryerson University). Accessed: December 3, 2024.

[5] 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.

[6] ACM SIGACCESS. Accessible writing guides. ACM SIGACCESS Resource Page, 2024. URL https://www.sigaccess.org/welcome-to-sigaccess/resources/accessible-writing-guide/. Accessed: 2024-10-28.

[7] 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. Updated: January 2024, Accessed: January 2, 2025.

[8] E. W. Sayers et al. Database resources of the National Center for Biotechnology Information. *Nucleic Acid Research*, 50:D20–D26, 2022.

[9] 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.

[10] National Center on Disability and Journalism. Disability language style guide. 2021. URL https://ncdj.org/style-guide/. Accessed: January 2, 2025.

[11] 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: 10.1145/1452562.1452565. URL https://doi.org/10.1145/1452562.1452565.

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

[13] 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: 10.1145/3593013.3593989. URL https://doi.org/10.1145/3593013.3593989.

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

[15] Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica (Zagreb)*, 22(3):276–282, 2012.

[16] 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: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

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

[18] 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.

[19] 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.

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

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

[22] 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.

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.