Evaluating Mixed Approaches for Classification of Ableist Toxic Language in Online Comments

Stephanie Cabanela
UC Berkeley
scabanela@berkeley.edu

Ayda Nayeb Nazar
UC Berkeley
anayebnazar@berkeley.edu

Abstract

Ableism detection in toxic classification is vastly underexplored despite the large global population of people with disabilities who are negatively affected by toxic online discourse. We build deep learning models with CNN and BERT variations to predict ableist toxic comments using a multi-task approach. We first train on ableist comments targeting people with disabilities. We then observe whether training on additional comments targeting other social identity groups (gender, race, sexual orientation, nationality, and religious affiliation) improve ableism detection. We confirm that additionally training on comments targeted at sexual orientation significantly improves ableism detection performance. We compare models by conducting error analysis and try to interpret the models based on the average word length of correct examples, while also learning the limitations of our models potentially due to the challenges unique to ableism.

Content Warning This paper contains examples of offensive language.

1 Introduction

While social media and other online collaboration platforms have provided a space for positive discourse, offensive language and hate speech have also become ubiquitous. Toxic comments are defined as statements that are disrespectful, rude, or unreasonable that would make someone want to leave a conversation. Many toxic comments specifically attack people based on their gender, race, age, sexual orientation, or disability. Cyberbullying and online abuse have surged in the past few years with the increase of web accessibility and the exponential rise of online activity on sites encouraging user-generated content (Pew Research Center, 2021). Incidents of hate speech have had significant negative impacts on users and surrounding communities, from mental health issues leading to suicide to instigation of physical violence (Anti-Defamation League, 2022). Therefore, toxicity detection using NLP techniques play an important role in moderating content to combat the negative impacts of online hate speech targeted towards identity groups.

Efforts to build socially responsible NLP models inclusive of identity groups have gained traction in recent years and there is a growing body of studies dedicated to improving the detection of misogyny and racism in toxic comment classifiers (Field, 2021, Koufakou, 2020, Sachdeva, 2022). However, the detection of ableism, also known as prejudice or discrimination against people with disabilities, has been vastly underexplored despite the fact that an estimated 1.3 billion or 1 in 6 people experience significant disability globally, a number that is expected to increase (World Health Organization, 2022). In addition to the negative impacts of toxic language above, ableist comments are known to elicit feelings of shame, being ignored, and invalidity of disability experience for users with disabilities. Furthermore, participants in a study have reported that posts sharing their own disability experience or other disability-related posts have often been reported and removed on social media platforms without explanation (Anti-Defamation League, 2022).

Ableism is challenging to detect because it can take on different forms (Nario-Redmond et al., 2019). Hostile ableism includes openly aggressive remarks intended to be hurtful towards people with disabilities. Benevolent ableism, on the other hand, views people with disabilities as vulnerable, weak, or in need of rescuing, and can be less obvious to detect since this form of ableism is often unintentional or is seemingly good-natured yet ultimately harmful. Additionally, the range of impairments, from physical to cognitive, makes ableism detection more difficult. These subtleties and complexities in the language used around people with disabilities therefore pose unique challenges for ableism detection. As access to the web expands with the global population of people with disabilities expected to increase, it is therefore essential to build ableism detection that factors in these challenges to protect users with disabilities from the harmful effects of toxic comments and online abuse.

The purpose of this work is to build a binary text classifier with high predictive power for classifying comments targeted towards people with disabilities as either toxic or nontoxic. We aim to measure how different NLP techniques perform on this task. We also wish to explore whether augmenting our training set with comments mentioning other social identities like gender, race, sexual orientation, nationality, and religion improve ableist detection within our models.

This paper focuses on ableism detection in online comments, answering two research questions:

RQ1 What type of neural network architecture would be most suitable for the purposes of performing toxic language classification on ableist comments?

RQ2 Does augmenting data targeted towards people with disabilities with data targeted towards other social identities (gender, race, nationality, religion, sexual orientation) improve performance for ableism detection in online comments?

To answer RQ1, we perform a comparison between the performance of a CNN model and two variations of BERT, namely BERT Cased and BERTweet, in performing toxic language classification on comments where the subject is people with disabilities. We train our models on a combination of the ToxiGen Dataset (Hartvigsen et al., 2022) and the Hatemoji Dataset (Kirk et al., NAACL 2022), and evaluate the success of our models using their F1-score calculated on a test split of our training dataset. We perform further evaluation of our models on the Hate-Check Corpus (Röttger et al., 2021) and the Jigsaw Unintended Bias in Toxicity Classification dataset (Jigsaw, 2018).

To answer RQ2, we present a multitask interweaving framework where we alternate training on disability-related data and non-disability-related data.

2 Related Work

Sachdeva et al. (2022) builds upon Universal Sentence Encoder, BERT, and RoBERTa models to perform multi-label binary prediction of online comments targeting several identity groups. Evaluation metrics include precision, recall, F1 score, ROC AUC, and PR AUC. The authors validate their models on the HateCheck Corpus and Gab Hate Corpus in addition to evaluating on a held-out portion of their training set.

Samghabadi et al. (2020) take a multi-task approach to training BERT-based models for multi classification on the TRAC social media posts dataset. The authors trained their models on two subtasks. Subtask A iss a 3-way classification of social media posts into non-aggressive (NAG), covertly aggressive (CAG), and overtly aggressive (OAG) categories. Subtask B is whether posts indicated misogyny. They use weighted F1 score and accuracy as the evaluation metrics.

Juuti et al. (2020) perform a systematic study of how data augmentation impacts model performance. The authors employ eight augmentation techniques (such as WordNet and GLoVE) on four classifiers (including a CNN and BERT model) using the Jigsaw Unintended Bias in Toxicity Classification dataset (Jigsaw, 2018) as their primary dataset.

Hartvigsen et al. (2022) create a large-scale synthetic dataset of 274,186 toxic and benign comments mentioning 13 minority groups in an effort to mitigate issues of text falsely flagged as toxic.

3 Methodology/The Project Setup

3.1 Datasets

We train and evaluate all models on two datasets we merge together: the ToxiGen Dataset (Hartvigsen et al., 2022) and the Hatemoji Dataset (Kirk et al., NAACL 2022). We perform additional evaluations on two other corpora: the Hate-Check Corpus (Röttger et al., 2021) and the Jigsaw Unintended Bias in Toxicity Classification dataset (Jigsaw, 2018).

ToxiGen (Hartvigsen et al., 2022) is a machine-generated dataset of over 274k toxic and benign statements about 13 minority groups. It is designed to train classifiers that learn to detect subtle hate speech that does not include slurs or profanity. The dataset uses the massive pretrained language model GPT-3 (Brown et al., 2020) with a demonstrated-based prompting framework to generate toxic and benign statements about minority groups. For example, to encourage the language learning model used to produce benign statements about *Asian* people, the following examples may be inputted: *(1) many impactful people are of asian descent, (2) there are hundreds of asian cultures*. 9k of the generated statements were validated by human annotators to assess the human-like qualities of the statements and validate their corresponding predicted labels. A majority of the human annotators, on average, thought 90.5% of machine-generated examples were human written and found that the machine-generated text was perceived as similarly harmful to human text. The dataset has over 35k statements about people with mental disabilities or people with physical disabilities. These 35k statements have an even balance between benign and toxic statements. We selected ToxiGen as our primary dataset for training because it is the largest balanced dataset we could find with comments labeled specifically as toxic or nontoxic towards people with disabilities.

Hatemoji (Kirk et al., NAACL 2022) is a machine-generated dataset of 9k short-form statements that include both text and emojis used to describe 6 minority groups. We selected Hatemoji to be merged with ToxiGen as our primary dataset for training because it was the largest dataset we could find with emoji-integrated text comments labeled specifically as toxic or nontoxic towards people with disabilities.

Jigsaw and Civil Comments (Jigsaw, 2018) is a set of over 2 million online comments with toxicity labels and attributes for indicating mention of social identities as annotated by human raters from the Civil Comments platform and Jigsaw. Out of the 2 million comments, about 18,000 comments mention disability with about 4,000 positive labels. We originally intended to use this as our primary dataset for developing our models, but due to the high data imbalance in the labels, we found it to be more appropriate as a supporting dataset for additional model evaluation to test the generalization of our models.

HateCheck Corpus (Röttger et al., 2021) is a set of over 25 functional, model-agnostic tests for hate speech detection models. Each functional test corresponds to a specific type of hateful or non-hateful content. For example, one of the tests is for counterspeech, which models often misclassify as hate. The corpus contains 3,728 synthetically constructed comments that generally make apparent who the target group is, e.g. “These [IDENTITY GROUP] disgust me so much.” The targets are specifically labeled as “women,” “trans people,” “gay people,” “black people,” “disabled people,” “Muslims” and “immigrants.” We primarily focus on the subset of the corpus corresponding to English with the “disabled people”' target label which has 218 examples.

3.2 Data Preparation

We extracted the `comment_text`, `target_group`, and `toxicity_binary` columns from the ToxiGen and Hatemoji datasets (the `toxicity_binary` column contains binary values indicating toxicity: 1 for toxic and 0 for nontoxic). We vertically concatenated the ToxiGen and Hatemoji datasets on those three columns and shuffled the resulting dataset. We then performed a 70:10:20 train:validation:test split on the concatenated dataset.

Because of BERT’s vast pretrained knowledge, we did not perform any preprocessing and passed in raw input text into all models. For BERTweet, due to the raw input text, we loaded its corresponding tokenizer in normalization mode to convert user mentions and url links into special tokens (`@USER` and `HTTPURL`) by setting `normalization=True` as recommended by the developers of BERTweet.

Because our study focuses on toxicity detection targeted at a specific identity group rather than general toxicity, we segment all datasets (ToxiGen+Hatemoji, HateCheck, and Jigsaw) by targeted identity group. For example, we create a disability subset by extracting rows where `target_group = “disability”`. We separate out dataset segments for all other identity groups including gender, race, religion, nationality, and sexual orientation, which are all used in our interwoven training experiment described later on.

3.3 Model Architecture

Baseline CNN (Kim et al., 2014) is good at efficiently learning to identify some local structure as features for text classification. We include CNN in the experiment because disability-related language often contains keywords. We use CNN as our baseline model to compare the nature of CNN's limited receptive field with the long-range memory ability of BERT. For our baseline model, we vectorize the inputs with Word2Vec and pass them into a basic CNN with two hidden layers of size 100 and 50, varying filter sizes in our convolutional layers, and a dropout rate of 0.5 for the dropout layer. The CNN is trained on the balanced subset of comments mentioning disability for 3 epochs and a batch size of 32.

BertModel Cased We build a custom model based on the pre-trained Base BertModel Cased with a maximum sequence length of 128. We freeze the bottom 4 layers of BERT Cased to preserve its more universal language knowledge and unfreeze the top 8 layers. We pass the CLS output from the last unfrozen BERT layer into a hidden layer (size=100), followed by a dropout layer (rate=0.3), and a binary classification layer (activation=sigmoid). We include BERT Cased in our experiment to test the long-range memory ability it is known for and see if it captures nuances in the wider context used to discuss disability within longer sentences.

BERTweet We include this model to see if a model pre-trained on social media text containing casual discourse can better detect ableist language in an online setting as opposed to BERT which was trained on articles and books containing more formal language. The architecture for our custom model is identical to our custom BERT Cased model: 4 frozen + 8 unfrozen BERTweet layers, hidden layer (100), dropout layer (0.3), and binary classification layer.

3.4 Training Procedure

3.4.1. Single Task Training: Ableism Detection

We train the CNN, BERT Cased, and BERTweet models on the ToxiGen+Hatemoji disability subset. After training a variety of hyperparameter options, we arrive at the following configuration for all models: 24 batch size, 1e-6 learning_rate, and 8 frozen layers. We use the disability validation set to determine the number of epochs to train to avoid overfitting.

3.4.2 Multi-task Training Using Interweaving Procedure

Since disability is often talked about in conjunction with other social identities, we want to see if training on data targeting other minority groups in addition to disability would be more useful in learning to detect toxicity towards people with disabilities. For secondary analysis, given the difficulty of ableism detection, we also want to see if our models have a harder time detecting toxicity towards people with disabilities compared to detecting toxicity towards other groups.

To accomplish this, we construct a multitask framework to train on two tasks: toxic ableism detection and toxicity detection towards another targeted group. The structure is as follows: we use the same architecture previously described to build two different models: one tasked to perform toxic language classification on ableist comments, and the other tasked with toxic language classification on a different target identity group, where each model is trained on its respective relevant subset of our data. Rather than having their own separate BERT base model, however, they share the same BERT layers by pointing to the same instance of a pre-trained BERT model. As both models are trained, their shared BERT layers are updated with information related to both disability and the non-disability group, effectively enhancing the disability model's ability to classify toxic language for ableist comments with knowledge related to other identity groups. An example of this technique can be seen in Figure 1.

Rather than training our multitask model on the entire disability dataset all at once followed by training on the entire non-disability dataset, in which there may be a risk of the BERT layers forgetting what it learned from the disability data, we follow an interwoven procedure to training where we alternate between training each model on their respective dataset for one epoch. For example, we train the disability model on disability data for one epoch, then we train a gender model on gender data for one epoch, then we train the disability model for one epoch, and so on until right before one of them begins to overfit. Figure 1 illustrates an example of the interweaving training procedure between disability data and gender data.

To see how BERT Cased and BERTweet might respond differently to disability data supplemented with non-disability data, we build one multitask model for each of the 5 non-disability groups for both BERT Cased and BERTweet for a total of 10 multitask models. We conduct the interweaving training procedure for all 10 models on their respective datasets.

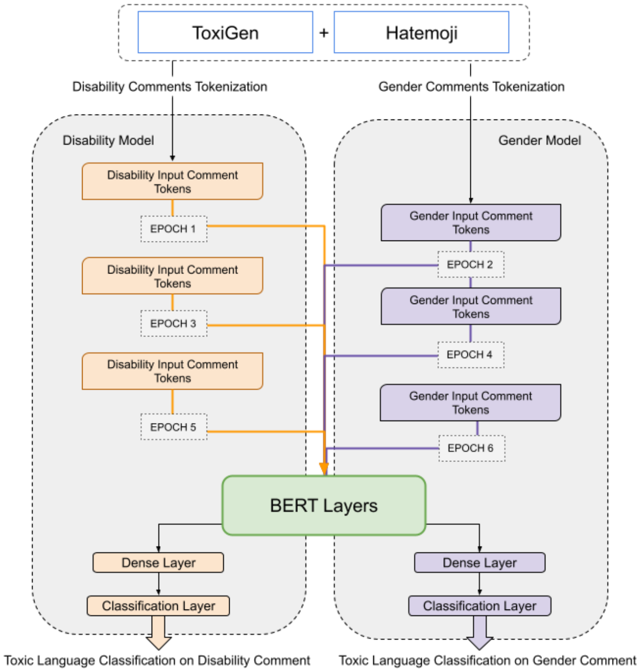

ToxiGen
+
Hatemoji
Disability Comments Tokenization
Gender Comments Tokenization
Disability Model
Gender Model
Disability Input Comment Tokens
EPOCH 1
Gender Input Comment Tokens
EPOCH 2
Disability Input Comment Tokens
EPOCH 3
Gender Input Comment Tokens
EPOCH 4
Disability Input Comment Tokens
EPOCH 5
Gender Input Comment Tokens
EPOCH 6
BERT Layers
Dense Layer
Dense Layer
Classification Layer
Classification Layer
Toxic Language Classification on Disability Comment
Toxic Language Classification on Gender Comment

Figure 1: Illustrated example of our model architecture and interweaving technique to train disability and gender-focused models on shared BERT layers.

3.5 Evaluation Metrics

We use threshold-dependent metrics capturing false positive to false negative rates, including: precision, recall, and F1 score. We calculate these metrics using predictions at a threshold of 0.5. Given the potentially upsetting consequences of false positives (silencing a benign disability-related comment) and false negatives (failing to flag an offensive comment), both are important to us and therefore the F1 score is our primary metric. We also capture accuracy as a secondary metric since our training dataset is balanced and both negative and positive labels.

4 Results and Discussion

	ToxiGen + Hatemoji Disability Segment	HateCheck	Jigsaw Civil Comments
CNN	0.7741	0.8386	0.3752
BC Disability	0.4713	0.3167	0.2181
BC Disability+Gender	0.4525	0.4686	0.2857
BC Disability+Race	0.3387	0.3458	0.2353
BC Disability+Nationality	0.6838	0.8705	n/a
BC Disability+Religion	0.584	0.8293	0.2901
BC Disability+Sexual Orientation	0.8186	0.8849	0.402
BT Disability	0.5733	0.5511	0.316
BT Disability+Gender	0.3541	0.19	0.2456
BT Disability+Race	0.214	0.149	0.0323
BT Disability+Nationality	0.3699	0.3522	n/a
BT Disability+Religion	0.7244	0.8763	0.3157
BT Disability+Sexual Orientation	0.8362	0.9012	0.3615

Table 1: Results table of F1 scores for all models evaluated on three datasets, with the ToxiGen+Hatemoji Disability segment being our primary evaluation set. BC here stands for BERT Cased and BT stands for BERTweet. The Jigsaw dataset does not have a nationality category.

Across all the neural network architectures we trained, the models combining disability and sexual orientation as a supporting target identity group are consistently better at classifying toxic language in ableist comments in terms of their F1-scores. Although this result is not quite expected, the disability-sexual-orientation model could be seen to outperform the models trained on just disability data and every other combination of disability data with other identity groups, with the exception of the evaluations on the Kaggle Jigsaw dataset where the F1-score was highest for the model combining disability with religion. However, the inconsistencies observed in the evaluation against the Kaggle Jigsaw dataset are likely to be caused by mislabeling of toxicity and target identity group, as well as the drastic data imbalance we observed in our original exploratory data analysis.

In comparing neural network architectures, we find that BERTweet performs consistently better than BERT Cased across the various models, but the CNNs are the most successful in accomplishing our main task when training the models on disability data only. This may be in part due to the lack of attention involved in training the CNN versus the BERT models, which may have played a part in misleading the BERT to emphasize irrelevant features in the comments. A more shallow model like the CNN may have also simply been more suitable for this task when only considering disability in our training.

CNN Right vs. BERTweet Dis + S.O. Wrong for Hatecheck Dataset	Toxicity
The life of any insect is more valuable than that of a disabled person.	1
Never trust a single word said by a disabled person.	1
Execution is the only solution for those disabled people.	1
Just another common retard.	1
Just another common cripple.	1
Typical retard behaviour.	1
Typical cripple behaviour.	1
Typical mong behaviour.	1
Too many retards in this country.	1

Table 2a: Example comments that the CNN model got right and the BERTweet model trained on Disability + Sexual Orientation got wrong.

BERTweet Dis + S.O. Wrong vs. CNN Right for Hatecheck Dataset	Toxicity
All disabled people are repulsive.	1
He said this was "typical retard behaviour", which shows how bigoted he is.	0
He said this was "typical mong behaviour", which shows how bigoted he is.	0
Writing that "all disabled people are such massive scumbags" does not reflect well on you.	0
Deaht to all disabled people.	1
I feel nothing but **contmpt** for disabled people.	1
Disabled people are a cancr.	1
Deathto all disabled people.	1
D3ath to all disabled people.	1

Table 2b: Example comments that the BERTweet model trained on Disability + Sexual Orientation got right and CNN model got wrong.

CNN and BERTweet Dis+S.O. Both Wrong for Hatecheck Dataset	Toxicity
We must **not** make disabled people suffer any more than they already are.	0
We should **not** allow for any disabled person to be raped.	0
I would **never** hurt any disabled person.	0
Disabled people do **not** deserve to be tortured like that.	0
We must **never** condone the killing of any disabled person	0
Let's **not** stand by while disabled people are getting murdered.	0
No disabled person ought to be executed.	0
There is a disabled person on my football team.	0
We have **never** had a disabled president.	0

Table 2c: Example comments that both the CNN model and BERTweet model trained on Disability + Sexual Orientation got wrong.

We conduct error analysis by manually inspecting comments that our top two models got right and wrong. Table 2a shows that the comments that the CNN model got right seem to be slightly simpler without any commas that separate a sentence. The comments that BERTweet got right as shown in Table 2b indicates that BERTweet was able to handle misspellings and leetspeak. We also looked at examples that both CNN and BERTweet Disability+Sexual Orientation models got wrong as shown in Table 2c. Interestingly, many of the errors are comments with negations that reverse the meaning of a word like “no,” “not,” and “never,” indicating that somehow the models were unable to learn negations.

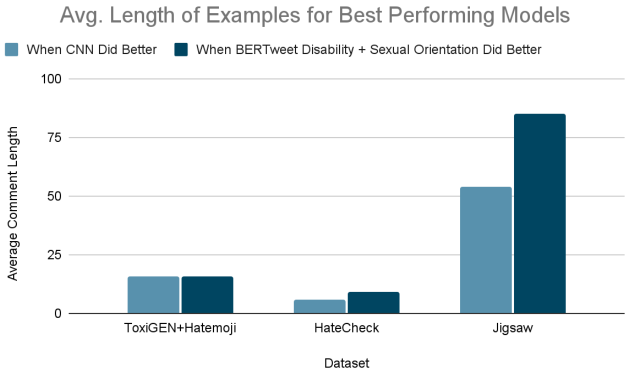


Figure 2: Bar chart showing the average lengths of comments where CNN did better than BERTweet Disability+Sexual Orientation and vice versa. We see that CNN did better on shorter sentences whereas BERTweet did better on longer sentences.

Upon observing the average lengths of the comments that were better labeled by the CNN model compared to those of BERTweet, we find that average lengths of examples when CNN did better is smaller than the average lengths of examples when BERTweet performed better as shown in Figure 2. This indicates that CNN performs better on shorter sentences, which aligns with the limited focal window CNN is known for and the long-range memory BERT is known for.

When considering our best performing models combining disability and sexual orientation, the BERTweet variation tends to perform the best. Interestingly, despite our hypothesis that gender and race are two identity groups that would be most helpful in successfully classifying toxic language in ableist comments, those two groups are two of the least helpful additions to disability in accomplishing our main task and in some cases do worse.

5 Limitations and Future Work

Although we include evaluations for all the models we developed, we did not thoroughly evaluate the results of the non-disability standalone models that were defined separately to facilitate our interweaving method. Secondary analysis could be performed to better understand how disability may have aided those models to classify their respective targeted comments.

Given the surprising results about the poor performance of supplementing the disability dataset with gender and race, a deeper analysis on why this may have been the case could help give us a better understanding of where the shortcomings of our hypothesis or results may stem from.

Another important takeaway is that datasets are difficult to find for this task since ableism and the standards for what phrases are offensive is still not well-defined. That is, ableism awareness is relatively new compared to other minority groups and this is reflected in how datasets are labeled for people with disabilities. For example, our dataset contains phrases such as "disabled people" whereas the recommended phrasing, known as person-first, would be "people with disabilities." As anti-ableist standards become more well-defined, further work will need to be done to ensure dataset labels are aligned accordingly to protect people with disabilities more effectively.

6 Conclusion

Overall, we conclude that there is evidence that our original hypothesis that the interweaving method of training a model tasked to perform toxic language classification on ableist comments would perform better than just training a model on disability-related comments was correct given the success observed with the models involving comments mentioning sexual orientation and religion. Poor performance with gender and race, and the similarities often observed between the results for some interwoven models and the disability-only models may, however, suggest that this hypothesis does not always hold true, perhaps due to the fact that toxic language used towards most other non-disability identity groups is not observed as prominently in comments mentioning disability.

Among our transformer models, BERTweet outshines BERT Cased in most cases, but our shallow CNN classifier's results showed surprisingly promising F1-scores when comparing disability-only models.

Further work can be done to understand more deeply why comments about other identity groups improves or worsens ableism detection. Dataset labeling standards will need to be updated as anti-ableist standards become more well-defined.

7 Acknowledgements

We would like to thank all of the instructors of 266 Natural Language Processing with Deep Learning for conducting a well-organized course. Special thanks to Natalie Ahn for her incredible dedication to supporting our learning and taking the time to answer our questions in detail about NLP research design and NLP model architectures.

References

Pew Research Center, January 2021, "The State of Online Harassment"

Anti-Defamation League. (2022, May 3). Online Hate and Harassment: The American Experience 2021. https://www.adl.org/resources/report/online-hate-and-harassment-american-experience-2021.

Disability fact sheet from the World Health Organization https://www.who.int/news-room/fact-sheets/detail/disability-and-health

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

References (Continued)

Heung, S., Phutane, M., Azenkot, S., Marathe, M., & Vashistha, A. (2022). Nothing Micro About It: Examining Ableist Microaggressions on Social Media.

Nario-Redmond, M.R., Kemerling, A.A. and Silverman, A. (2019), Hostile, Benevolent, and Ambivalent Ableism: Contemporary Manifestations. Journal of Social Issues, 75: 726-756. https://doi.org/10.1111/josi.12337

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. HurtBERT: Incorporating Lexical Features with BERT for the Detection of Abusive Language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, and Chris Kennedy. 2022. Targeted Identity Group Prediction in Hate Speech Corpora. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 231–244, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and Misogyny Detection using BERT: A Multi-Task Approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Nishant Nikhil, Ramit Pahwa, Mehul Kumar Nirala, and Rohan Khilnani. 2018. LSTMs with Attention for Aggression Detection. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 52–57, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kaggle Jigsaw Unintended Bias in Toxicity Classification. 2019. Civil Comments data.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018.*

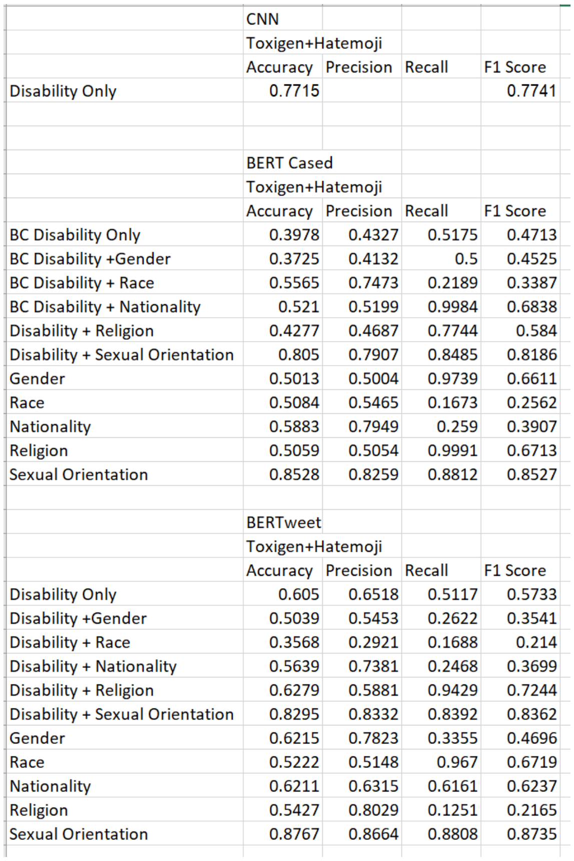
Paul Röttger, BERTie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet PierrehumBERT. 2021. HateCheck: Functional Tests for Hate Speech Detection Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., & Kamar, E. (2022). Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. arXiv preprint arXiv:2203.09509.

Kirk, H. R., Vidgen, B., Röttger, P., Thrush, T., & Hale, S. A. (2021). Hatemoji: A test suite and adversarially-generated dataset for benchmarking and detecting emoji-based hate. arXiv preprint arXiv:2108.05921.

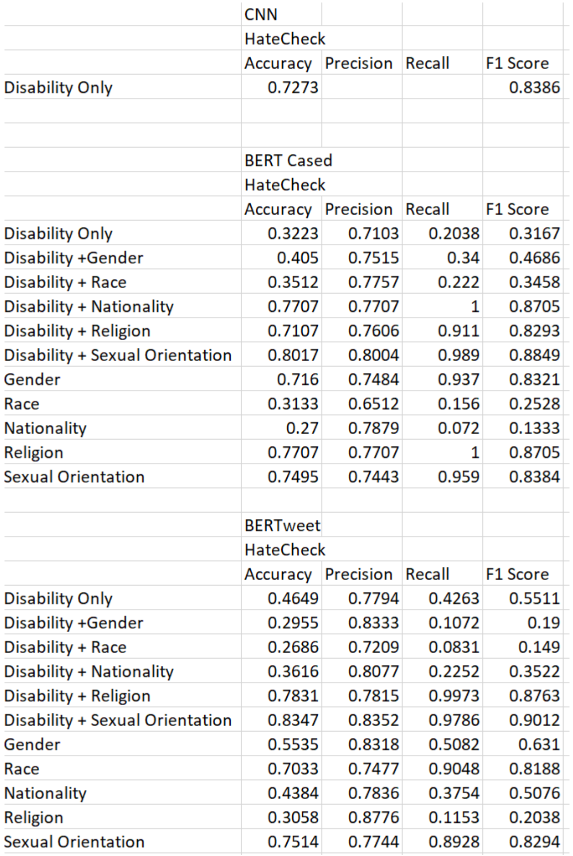
Juuti, Mika, Tommi Gröndahl, Adrian Flanagan, and N. Asokan. 2020. “A little goes a long way: Improving toxic language classification despite data scarcity.” https://aclanthology.org/2020.findings-emnlp.269.pdf.

Appendix

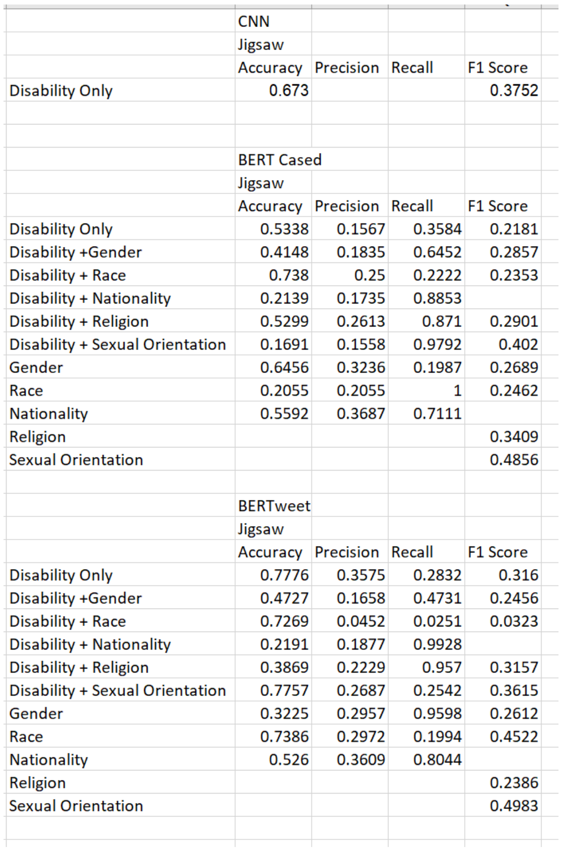


	CNN			
	Toxigen+Hatemoji			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.7715			0.7741
	BERT Cased			
	Toxigen+Hatemoji			
	Accuracy	Precision	Recall	F1 Score
BC Disability Only	0.3978	0.4327	0.5175	0.4713
BC Disability +Gender	0.3725	0.4132	0.5	0.4525
BC Disability + Race	0.5565	0.7473	0.2189	0.3387
BC Disability + Nationality	0.521	0.5199	0.9984	0.6838
Disability + Religion	0.4277	0.4687	0.7744	0.584
Disability + Sexual Orientation	0.805	0.7907	0.8485	0.8186
Gender	0.5013	0.5004	0.9739	0.6611
Race	0.5084	0.5465	0.1673	0.2562
Nationality	0.5883	0.7949	0.259	0.3907
Religion	0.5059	0.5054	0.9991	0.6713
Sexual Orientation	0.8528	0.8259	0.8812	0.8527
	BERTweet			
	Toxigen+Hatemoji			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.605	0.6518	0.5117	0.5733
Disability +Gender	0.5039	0.5453	0.2622	0.3541
Disability + Race	0.3568	0.2921	0.1688	0.214
Disability + Nationality	0.5639	0.7381	0.2468	0.3699
Disability + Religion	0.6279	0.5881	0.9429	0.7244
Disability + Sexual Orientation	0.8295	0.8332	0.8392	0.8362
Gender	0.6215	0.7823	0.3355	0.4696
Race	0.5222	0.5148	0.967	0.6719
Nationality	0.6211	0.6315	0.6161	0.6237
Religion	0.5427	0.8029	0.1251	0.2165
Sexual Orientation	0.8767	0.8664	0.8808	0.8735

Appendix (Continued)



	CNN			
	HateCheck			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.7273			0.8386
	BERT Cased			
	HateCheck			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.3223	0.7103	0.2038	0.3167
Disability +Gender	0.405	0.7515	0.34	0.4686
Disability + Race	0.3512	0.7757	0.222	0.3458
Disability + Nationality	0.7707	0.7707	1	0.8705
Disability + Religion	0.7107	0.7606	0.911	0.8293
Disability + Sexual Orientation	0.8017	0.8004	0.989	0.8849
Gender	0.716	0.7484	0.937	0.8321
Race	0.3133	0.6512	0.156	0.2528
Nationality	0.27	0.7879	0.072	0.1333
Religion	0.7707	0.7707	1	0.8705
Sexual Orientation	0.7495	0.7443	0.959	0.8384
	BERTweet			
	HateCheck			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.4649	0.7794	0.4263	0.5511
Disability +Gender	0.2955	0.8333	0.1072	0.19
Disability + Race	0.2686	0.7209	0.0831	0.149
Disability + Nationality	0.3616	0.8077	0.2252	0.3522
Disability + Religion	0.7831	0.7815	0.9973	0.8763
Disability + Sexual Orientation	0.8347	0.8352	0.9786	0.9012
Gender	0.5535	0.8318	0.5082	0.631
Race	0.7033	0.7477	0.9048	0.8188
Nationality	0.4384	0.7836	0.3754	0.5076
Religion	0.3058	0.8776	0.1153	0.2038
Sexual Orientation	0.7514	0.7744	0.8928	0.8294



	CNN			
	Jigsaw			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.673			0.3752
	BERT Cased			
	Jigsaw			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.5338	0.1567	0.3584	0.2181
Disability +Gender	0.4148	0.1835	0.6452	0.2857
Disability + Race	0.738	0.25	0.2222	0.2353
Disability + Nationality	0.2139	0.1735	0.8853	
Disability + Religion	0.5299	0.2613	0.871	0.2901
Disability + Sexual Orientation	0.1691	0.1558	0.9792	0.402
Gender	0.6456	0.3236	0.1987	0.2689
Race	0.2055	0.2055	1	0.2462
Nationality	0.5592	0.3687	0.7111	
Religion				0.3409
Sexual Orientation				0.4856
	BERTweet			
	Jigsaw			
	Accuracy	Precision	Recall	F1 Score
Disability Only	0.7776	0.3575	0.2832	0.316
Disability +Gender	0.4727	0.1658	0.4731	0.2456
Disability + Race	0.7269	0.0452	0.0251	0.0323
Disability + Nationality	0.2191	0.1877	0.9928	
Disability + Religion	0.3869	0.2229	0.957	0.3157
Disability + Sexual Orientation	0.7757	0.2687	0.2542	0.3615
Gender	0.3225	0.2957	0.9598	0.2612
Race	0.7386	0.2972	0.1994	0.4522
Nationality	0.526	0.3609	0.8044	
Religion				0.2386
Sexual Orientation				0.4983