Detecting Biased Language in Icelandic: A Named Entity Recognition Approach for Socially Responsible Text Analysis

Steinunn Rut Friðriksdóttir
University of Iceland
srf2@hi.is

Hafsteinn Einarsson
University of Iceland
hafsteinne@hi.is

Trigger warning: This paper contains examples of biased, offensive, and potentially harmful language, which are included for research purposes only.

Abstract

Bias research has been limited for Icelandic, a low-resource language with few NLP tools for socially aware text analysis. We address this gap by developing a publicly accessible web application that detects biased and stigmatizing vocabulary in Icelandic text and provides category-specific feedback to encourage reflection and more inclusive communication. The application is powered by the best-performing of three Named Entity Recognition (NER) models that we trained on automatically annotated data derived from a manually compiled lexicon of over 2,000 biased terms and phrases across 14 social categories, ranging from misogyny and queerphobia to religious and ethnic bias. All components, including the lexicon, annotated dataset, the three fine-tuned models and the web application code, are freely available online[1], offering a transparent and reproducible framework for bias detection in low-resource languages.

1 Introduction

Language can perpetuate social biases through the repeated use of stigmatizing or exclusionary vocabulary. Although the detection of such language has received increasing attention in natural language processing, most existing approaches are designed for high-resource languages such as English and often fail to account for the linguistic and cultural specificities of smaller language communities. In this work, we present a domain-specific Named Entity Recognition (NER) model designed to identify biased vocabulary in Icelandic, a low-resource language where such tools are scarce.

Our approach was inspired by the NBIAS framework [1], which introduces a taxonomy-driven system for identifying biased terms in English texts using NLP techniques. We extend this concept to include more fine-grained categorization and adapt it to the sociocultural landscape of the Icelandic language. Like NBIAS, our work emphasizes fine-grained categorization of bias types which we frame as a sequence labeling task, training a NER model to detect both single-word and multiword biased expressions using BIO tagging.

We constructed a lexicon of over 2,000 Icelandic words and expressions that may carry bias, covering 14 social categories: addiction, disability, origin (including both nationality and ethnicity), LGBTQIA+ identity, appearance, personal traits (i.e. vocabulary meant to attack the interlocutor's

[1]Lexicon, dataset, IceBERTBias, ScandiBERTBias, mBERTBiasIce, web application code on Github, model training and data curation code.

Socially Responsible Language Modelling Research (SoLaR) Workshop at NeurIPS 2024.

person without referencing protected traits), profanity, religion, sexually charged language, social status (including both age related terms, class division, and outdated job titles), vocabulary indicating the interlocutor's lack of intelligence, vulgar vocabulary (for instance threats of violence and references to bodily fluids), misogyny, and general bias (including politically inflammatory language, anti-abortion related phrases and negatively charged verbs such as *væla* (e. *whine*)). This lexicon is then used to automatically label text of approximately 6,3 million tokens, consisting of news from six media sources along with dialogue from three Icelandic web-based forums spanning 21 years (2000-2021). This serves as training data for our model and as the backbone for a web-based application aimed at helping users identify biased or inflammatory language in their own writing.

In its prototype version, the application enables users to input Icelandic text and receive categorized feedback on potentially biased or stigmatizing vocabulary. Trigger words are highlighted according to a color scheme corresponding to their category—for example, terms potentially reflecting misogyny appear in one color, while those associated with queerphobia appear in another. The aim is not to impose prescriptive norms, but rather to encourage reflection and promote more inclusive language use.

This work contributes to the socially responsible development of language technologies in several ways. First, it provides a new annotated resource for bias detection in Icelandic, a critically underrepresented language in fairness-oriented NLP. Second, it demonstrates how NER can be adapted to flag socially relevant language phenomena beyond traditional named entities. Finally, the prototype web application has been made publicly available on GitHub for local deployment, supporting transparency and encouraging reflection on inclusivity and fair language use.

1.1 Related Work

Bias detection in natural language processing (NLP) has traditionally focused on classification tasks, e.g., detecting hate speech, toxicity, or offensive content at the sentence or document level. However, recent work has explored Named Entity Recognition (NER) as a more fine-grained approach to identifying biased expressions within text. This framing allows systems to detect specific words or phrases that contribute to exclusionary or harmful discourse, offering greater transparency and interpretability.

One foundational example of this approach is NBIAS [1], a modular NLP pipeline comprising four stages: data collection, corpus construction, model development, and evaluation, designed to detect bias in text. The authors curate annotated datasets from diverse domains and employ a transformer-based token-classification model with a custom "BIAS" entity label, marking tokens or spans that reflect biased content. Similarly, the GUS framework [2] was used to train several NER-models to detect social bias through three BIO-categories: Generalizations, Unfairness, and Stereotypes. Their results show that encoder-only models like BERT outperform decoder-only LLMs in accuracy, especially for stereotypes, highlighting the value of discriminative models.

These approaches demonstrate the potential of using NER for socially meaningful NLP tasks, particularly in identifying fine-grained instances of bias. Our work builds on this foundation by adapting the NER-based bias detection paradigm to a low-resource setting and expanding the range of annotated categories. Inspired by NBIAS and GUS, we curate bias-related vocabularies across diverse social dimensions and fine-tune a BERT-based model (IceBERT) for token-level detection. In doing so, we extend the scope of NER for bias identification to an underrepresented language and offer a practical, interpretable tool for public use.

1.2 Dataset Construction

This work relies on two complementary resources: a manually curated lexicon of biased and inflammatory Icelandic expressions, and a dataset of Icelandic forum texts annotated automatically using this lexicon. The dataset was used to train three Named Entity Recognition (NER) models to detect biased language in context.

1.3 Lexicon Construction

We constructed a lexicon of 2,106 Icelandic expressions, including both single words and short multiword phrases, that reflect bias, stigma, or inflammatory language. These expressions were

manually compiled by the authors from a variety of public sources, including IceTaboo [3], a lexicon of offensive or sensitive Icelandic terms, comment sections on Icelandic blogs and discussion forums, and right-wing Icelandic media platforms, which frequently feature exclusionary or polarizing rhetoric. Each expression was categorized in one of 14 bias-related categories: Addiction (101), Disability (95), Origin (288), General (255), LGBTQIA (193), Looks (129), Personal (121), Profanity (86), Religion (209), Sexual (154), Social (175), Stupidity (98), Vulgar (44), and Women (158).

The aim of the lexicon is to comprehensively reflect inflammatory language in Icelandic. However, it is important to note that the vocabulary lists are not exhaustive. Bias in language is complex, culturally embedded, and context-dependent. These lists represent a first step toward systematizing patterns of bias in Icelandic, but they do not capture the full range of harmful or exclusionary expressions that may occur in real-world texts.

1.4 Annotated Training Dataset

To create training data for the NER models, we used the social media and news subcorpora of the Icelandic Gigaword Corpus (IGC) [4] to get a mixture of informal, user-generated texts and more formal, proofread texts. We specifically used text from the forums Hugi.is, Malefnin.com and Bland.is, and the media outlets Bylgjan, Vísir, RÚV, Kaffið, Fréttablaðið, and Mannlíf. The majority of our data comes from the forums as they offer conversational, real-life language, where bias is often expressed more directly than in formal writing, but adding the news texts is an attempt at making the models better at generalizing to other domains. Additionally, the texts span 21 years, ranging from the year 2000 to 2023. It should, therefore, span both short-lived phrases from specific periods, and problematic vocabulary that remains consistent throughout these two decades. Text examples containing only O-labels were removed.

The IGC is tagged and lemmatized, which enabled efficient automatic annotation. A Python script matched lemmatized token sequences in the corpus against entries in the vocabulary lists, while maintaining the inflected word forms in the resulting output. This process allowed for the inclusion of both single-token and multiword expressions. It is, however, important to note that the multi-word expressions are significantly fewer than the single-word tokens, potentially harming the recall of the resulting model. At this point, sentence-level classification is out of scope for this project but we hope to add it to future versions of the models.

Critically, the resulting annotations were not manually reviewed or corrected. As such, the training data may contain false positives (e.g., cases where a matched expression is not actually biased in context) and false negatives (expressions not in the vocabulary list but still biased). A known example is the word "grjón" (e. *rice*) which is used as a derogatory word for people of Asian descent. We assume that the number of examples where this word refers simply to the food product are more numerous than those referring to Asians, leading to a number of false positives. While this certainly introduces noise into the dataset, we opt to include terms of this nature to improve overall recall of biased terms. Automatic tagging is a limitation, but manual revision would be prohibitively expensive, and thus we conclude that this reflects a realistic constraint in low-resource language settings.

1.5 Tag Schema

The model uses the standard BIO (Beginning, Inside, Outside) tagging scheme, following the widely adopted CoNLL format[2] for sequence labeling tasks. Each token in the input text is assigned one of the following labels:

- B-CATEGORY: Marks the first token of a biased expression belonging to a specific category (e.g., B-SEXUAL, B-WOMEN).
- I-CATEGORY: Indicates a subsequent token of a multiword expression within the same category.
- O: Indicates that the token is outside of any annotated biased expression.

This format enables the model to detect both single-token and multi-token expressions that signal bias or harmful framing. For example, in the phrase "grenja eins og smástelpa" (e. *cry like a little girl*),

[2]Language-Independent Named Entity Recognition (II).

the first token would be tagged as B-WOMEN, and the subsequent tokens as I-WOMEN. Although most expressions in the training data are single-token matches (leading to a high frequency of B-tags), the CoNLL-style BIO schema supports extensibility and consistent evaluation.

2 Model Architecture and Training

We fine-tuned three models as bias-aware sequence labeling models for Icelandic. Firstly, we fine-tuned IceBERT [5], which uses the RoBERTa-base architecture and is pre-trained on Icelandic text exclusively. IceBERT provides strong linguistic coverage for Icelandic and serves as a robust foundation for domain-specific downstream tasks. Secondly, we fine-tuned ScandiBERT [6], a RoBERTa based model pretrained on concatenated corpora spanning all five Scandinavian languages, i.e. Danish, Norwegian, Swedish, Icelandic, and Faroese, designed to improve cross-lingual transfer to low-resource languages like Faroese by leveraging phylogenetic closeness instead of relying on massively multilingual models. Finally, we fine-tuned BERT-base-multilingual-cased (mBERT) [7], a massively multilingual model pretrained on Wikipedia text in 104 languages, including Icelandic.

2.1 Fine-Tuning Setup and Evaluation

All models were trained on the same dataset using an 80/10/10 split (153,816 training sentences, 15,381 development sentences, 15,383 test sentences). The dataset shows substantial class imbalance, with high-frequency categories such as B-PROFANITY (19,335 instances) and B-PERSONAL (19,151), and low-frequency categories such as B-RELIGION (2,901), B-ORIGIN (4,931), and B-LGBTQIA (3,506). Most bias expressions are labeled only with a B-prefix, with few spanning multiple tokens (e.g., I-GENERAL: 150 tokens). To mitigate imbalance, we used class-weighted loss functions (`scikit-learn`'s balanced class weights), inversely weighting classes by frequency.

To ensure comparability across models, we applied identical hyperparameters, chosen based on pilot experiments with the development set and established best practices for transformer-based NER. All models were fine-tuned using the AdamW optimizer (learning rate 2×10^{-5}, weight decay 0.01, label smoothing 0.1) with a per-device batch size of 16 and two gradient accumulation steps (effective batch size 32). Training used cosine learning rate scheduling with a 10% warmup ratio and early stopping (patience 5), both guided by the macro F1-score, which served as the primary evaluation metric due to severe class imbalance. Mixed-precision (FP16) training was enabled. Evaluation was performed every 1,000 steps using the `seqeval` library to compute entity-level precision, recall, and F1-scores (micro- and macro-averaged). All experiments were run for 8 epochs on a single NVIDIA RTX 3060 Laptop GPU (6 GB VRAM) with an AMD Ryzen 7 5800H CPU and 16 GB RAM, with total training time under three hours.

2.2 Results

Table 1 reports the macro- and micro-F1 scores of the three models on the held-out test set. To assess robustness, we additionally implemented a stratified sampling procedure to construct a balanced evaluation subset from sources within the Icelandic Gigaword Corpus excluded from our training, development, and test sets. These sources included judgments from the three levels of jurisdiction in Iceland, parliamentary speeches, various media outlets, three blogsites, and sport coverage. Further 55 sentences were manually selected and added to the sample to include more I-tagged examples.

		Test	**Gold**
Macro F1 (95% CIs)	IceBERT	0.970 (0.970-0.971)	0.868 (0.867-0.869)
	ScandiBERT	0.978 (0.978-0.978)	0.861 (0.859-0.862)
	mBERT	0.972 (0.972-0.973)	0.846 (0.845-0.848)
Micro F1 (95% CIs)	IceBERT	0.975 (0.974-0.975)	0.874 (0.872-0.875)
	ScandiBERT	0.982 (0.982-0.982)	0.865 (0.863-0.866)
	mBERT	0.976 (0.976-0.976)	0.848 (0.847-0.850)

Table 1: Performance of the models on the held-out test set and the manually reviewed additional dataset. Further analysis of precision, recall and F1-scores per category can be found in Appendix A.

For each "B-" category, up to five sentences were sampled in which the "B-" tag was immediately followed by its corresponding "I-" tag, and up to five in which it was not. To promote lexical diversity and reduce near-duplicate inflections, sentences were filtered so that the words carrying the B-tags differed by at least three characters in Levenshtein distance. This process should result in a class-balanced set with a broad, non-redundant vocabulary. All tagging was subsequently reviewed and corrected by the authors, resulting in a gold-standard set of 190 sentence examples.

McNemar's tests revealed statistically significant differences in model performance on both the test and gold datasets ($p < 0.001$ for all pairwise comparisons). On the test set, ScandiBERT achieved the highest macro F1, outperforming IceBERT by +0.0079 and mBERT by +0.0057. mBERT also surpassed IceBERT by +0.0021. On the gold set, IceBERT achieved the best results, surpassing ScandiBERT by +0.0073 and mBERT by +0.0220, while ScandiBERT outperformed mBERT by +0.0147. These results indicate that ScandiBERT is better suited to the broader, automatically labeled test set, whereas IceBERT has an advantage on the smaller, manually curated gold set, possibly reflecting a greater ability to adapt to new domains and data conditions.

2.3 Qualitative Error Analysis

While the quantitative results provide an overall picture of model performance, a closer inspection of individual errors reveals systematic patterns that explain many of the observed false positives (FP) and false negatives (FN). Table 2 presents six representative cases from the gold set, chosen to illustrate different error types.

Snippet (EN)	Gold label	Predicted	Error	Cause/Explanation
...ef ekki þa er það ***homma legt*** ("fruity", i.e. gay)	B-LGBTQIA, I-LGBTQIA	O, O	FN	Misspelling (added space, correct: "hommalegt")
...þetta helvítis ***útlenska pakk****...* (foreign riff-raff)	B-ORIGIN, I-ORIGIN	O, B-GENERAL	FN/MC	Category overlap (whole phrase vs. single word)
Ljótt *er að segja...* (it's bad (to say something))	B-LOOKS	O	FP	Okay in context (literally translates to "ugly")
...af ***olíuprinsum*** *skagans.* (oil sheikhs)	B-SOCIAL	O	FN	Rare, unlikely to appear in training data
...í hlutverki ***forfallins dópista****...* (deadbeat junkie)	B-ADDICTION, B-ADDICTION	B-ADDICTION, I-ADDICTION	FN	I-tag missed but correct class
...þessi pólski djöfull*...* (this Polish fucker)	O, B-ORIGIN, I-ORIGIN	B-ORIGIN, I-ORIGIN, I-ORIGIN	MC	Span beginning misplaced

Table 2: Representative errors made by IceBERT on the gold test set. FP = false positive, FN = false negative, MC = misclassification. Examples are shortened for space.

Our analysis shows that the most common errors on the gold dataset involve multi-word expressions, which are much less frequent than single-word tokens in the training data. Addressing this gap is a priority for future work, for example through data augmentation techniques.

3 Web Application

The prototype provides an interactive interface for detecting and highlighting potentially inappropriate or sensitive words within user-submitted text. Users can paste text into a text area, submit it, and receive annotated output in which words are highlighted according to predefined categories (e.g.,

profanity, misogyny, disability, etc.), each with distinct color coding. An example of this can be seen in Appendix B. The tokenization approach ensures that if any part of a word is identified as inappropriate (even at the subword level), the entire word is highlighted and labeled. This design choice avoids partial or fragmented highlights, making it easier for users to interpret the results in context.

The interface includes a category legend, a disclaimer about the automated nature and limitations of the detection, and a feedback option. It is implemented using HTML, CSS, and JavaScript, with a responsive design and user-friendly features such as tooltips and loading indicators. The tool is explicitly assistive and educational, not prescriptive: it does not censor, block, or suggest replacements, but rather encourages reflection and supports informed linguistic choices.

To promote trust and usability, the prototype performs no data retention and offers a contact option for suggestions or corrections. Accessibility considerations include screen reader-friendly labels, mobile responsiveness, and clear color coding with text labels for each category. Users with specific needs (e.g., color vision deficiencies or assistive technologies) are encouraged to provide feedback to improve inclusivity. By making both the models and code freely available, this prototype aims to empower Icelandic speakers with accessible NLP technology that fosters greater linguistic awareness and inclusion.

4 Discussion and Limitations

This work demonstrates that a bias-aware NER model can be trained for Icelandic using automatically annotated data. Nevertheless, several limitations must be acknowledged. First, the model relies on manually curated vocabulary lists that, while broad, are not exhaustive. They inevitably miss more subtle, euphemistic, or emerging forms of biased language, leading to under-detection of certain expressions and contexts.

Second, the training data was annotated automatically via lemmatized string matching without manual review. This introduces false positives (e.g., neutral uses of flagged words) and false negatives (e.g., bias expressed with terms absent from the vocabulary). While the models perform well overall, some errors are likely attributable to this noisy supervision. Third, the models operate at the word and phrase level, without modeling broader sentence or discourse context. This can result in mislabeling of terms used in critical or reclaimed contexts, particularly in journalistic or activist writing.

Finally, there is a risk that users may misinterpret flagged terms as inherently offensive, even when used appropriately, or assume unflagged text is free from bias. This underscores the importance of UI design in the prototype web application to convey that flagged terms are prompts for reflection rather than definitive judgments. Future work will focus on expanding the vocabulary, improving contextual modeling, and incorporating user feedback to validate model predictions in more nuanced real-world use cases.

5 Social Impact Statement

This work aims to promote inclusive and socially aware language use in Icelandic by detecting biased and potentially harmful expressions across diverse social categories. The models are available with a prototype web tool for local deployment, allowing users to input text and receive feedback on flagged terms by category (e.g., gender bias, religious bias, profanity). Designed to encourage reflection rather than enforce censorship, the tool provides transparent, non-judgmental guidance.

Focusing on Icelandic, a low-resource language with few bias detection tools, the freely available vocabulary and models aim to support both research and civic engagement. To mitigate risks, the interface emphasizes interpretability, avoids prescriptive suggestions, and does not retain user-submitted text. The resources are intended for research, education, and awareness-raising, not punitive monitoring or automated moderation. They are released under an OpenRAIL License that restricts harmful use. Future work will incorporate community feedback, improve contextual modeling, and add educational resources to promote nuanced understanding of bias in language.

Acknowledgments and Disclosure of Funding

Steinunn Rut Friðriksdóttir was supported by The Ludvig Storr Trust no. LSTORR2023-93030.

References

[1] Raza, S., Garg, M., Reji, D. J., Bashir, S. R., & Ding, C. (2024). NBIAS: A Natural Language Processing Framework for Bias Identification in Text. *Expert Systems with Applications, 237(Part B)*, 121542. https://doi.org/10.1016/j.eswa.2023.121542

[2] Powers, M., Raza, S., Chang, A., Mavani, U., Jonala, H. R., Tiwari, A., & Wei, H. (2025). The GUS Framework: Benchmarking Social Bias Classification with Discriminative (Encoder-Only) and Generative (Decoder-Only) Language Models. *arXiv*. https://arxiv.org/abs/2410.08388

[3] Sólmundsdóttir, A., Stefánsdóttir, L. B., & Ingason, A. K. (2021). IceTaboo: A Database of Contextually Inappropriate Words for Icelandic. In *Proceedings of CLARIN Annual Conference*, 39-43. https://epubl.ktu.edu/object/elaba:108748986/108748986.pdf#page=46

[4] Steingrímsson, S., Helgadóttir, S., Rögnvaldsson, E., Barkarson, S., & Guðnason, J. (2018). Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://aclanthology.org/L18-1690.pdf

[5] Snæbjarnarson, V., Símonarson, H. B., Ragnarsson, P. O., Ingólfsdóttir, S. L., Jónsson, H. P., Þorsteinsson, V., & Einarsson, H. (2022). A Warm Start and a Clean Crawled Corpus–A Recipe for Good Language Models. *arXiv*. https://arxiv.org/abs/2201.05601

[6] Snæbjarnarson, V., Simonsen, A., Glavaš, G., & Vulić, I. (2023). Transfer to a low-resource language via close relatives: The case study on Faroese. *arXiv*. https://arxiv.org/pdf/2304.08823

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*, 4171-4186. https://aclanthology.org/N19-1423.pdf

A Model Scores per Category

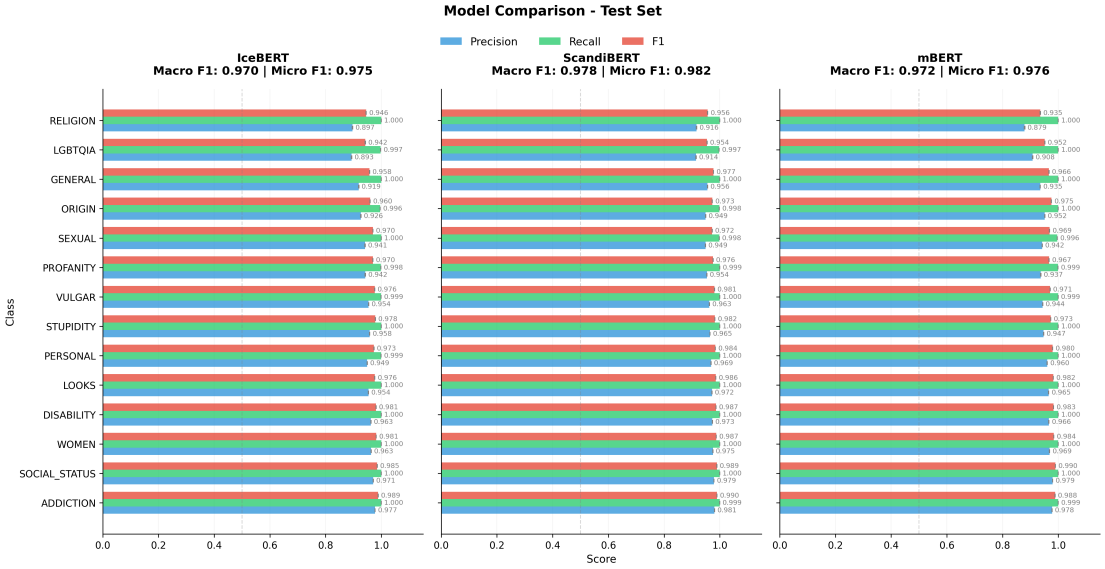

Figure 1: Precision, recall and F1 scores per category for each of the three models as measured on the held-out test set of 15,383 sentence examples.

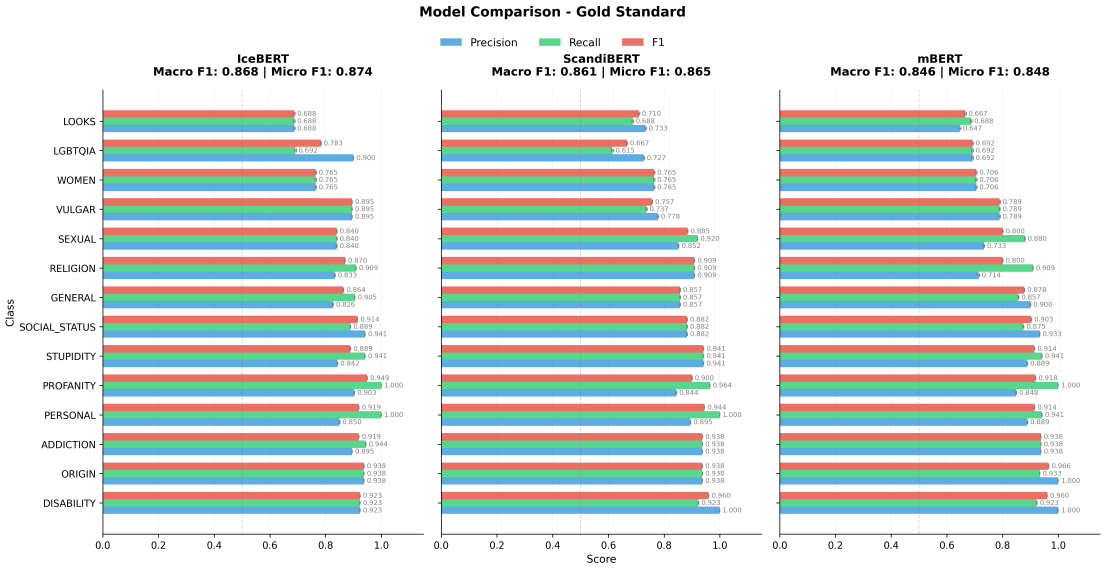

Figure 2: Precision, recall and F1 scores per category for each of the three models as measured on the manually reviewed gold set of 190 sentence examples. The gold set is composed of sentence examples taken from sources not included in the training, development or testing data.

B The Web Application

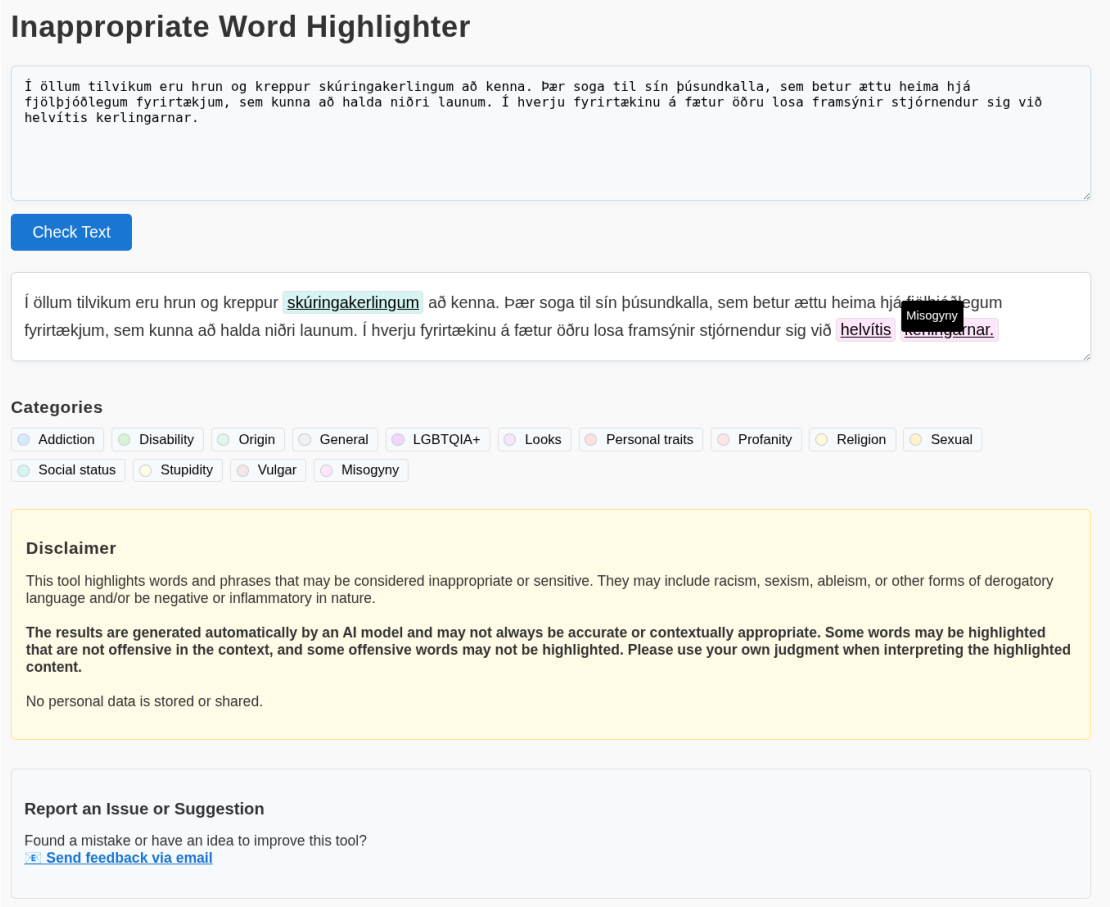

Figure 3: The prototype of the web-based application. Users input text into the text area and receive output where inappropriate words have been highlighted according to their category.

C NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: While the web-based tool that the paper mentions is still in development, its intended use is for the general public which should, if everything goes as planned, foster a more inclusive and socially aware communication for its users. As Icelandic is both a morphologically complex and a relatively low-resourced language, particularly in this domain, this work could very well inspire other researchers to create their own versions, hopefully contributing to bias detection efforts in lower-resourced languages.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Yes. Section 4 discusses the limitations of the work, particularly regarding the coverage of the data and model. It also highlights the potential for false positives and false negatives made by the model. It should additionally be noted that the paper currently only discusses a single fine-tuned model. If the paper is to be accepted, further analysis will be made, comparing several fine-tuned models on the discussed task.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Do to the anonymity of the reviewing process, no code or data has been attached to the paper. This will be added before the final publishing. However, information on the data collection process is included along with the hyperparameter settings used during fine-tuning of the model.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Note that this has NOT been included in this version due to anonymity but will be added before the final publishing of the paper.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Data splits and hyperparameters are discussed in the paper. Comments from reviewers on this matter would be appreciated if this has not been made clear enough.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: 95% confidence intervals and McNemar tests have now been added.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The hardware and time of execution are discussed in section 2.1.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: While the nature of the data is harmful in the sense that it includes toxic and biased language, its intended use is to guide users to mitigate their use of harmful language. Trigger warnings will be included with the published dataset and lexicon.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In section 4, along with the social impact statement, both positive and negative impacts of the model and web-tool are discussed. For instance, there is a risk that users will interpret any flagged words as inherently offensive, even when used appropriately. Similarly, there is a risk that users will use the tool to "prove" that potential harmful language that is missed by the model is not offensive. However, the potential positive impact outweigh the negative ones in the authors' opinion.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The published model and dataset are published under an OpenRAIL license that restricts harmful use.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The original papers presenting the Icelandic Gigaword Corpus, IceTaboo, and IceBERT are cited. All three are available with an open license.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: While we believe that the answer to this question is yes, comments from reviewers are appreciated if this is not the case.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]