RECAST: Interactive Auditing of Automatic Toxicity Detection Models

Austin P Wright
Georgia Institute of Technology
apwright@gatech.edu

Muhammed Ahmed
Mailchimp
muhammed.ahmed@mailchimp.com

Omar Shaikh
Georgia Institute of Technology
oshaikh@gatech.edu

Stephane Pinel
Mailchimp
stephane.pinel@mailchimp.com

Haekyu Park
Georgia Institute of Technology
haekyu@gatech.edu

Diyi Yang
Georgia Institute of Technology
diyi.yang@cc.gatech.edu

Will Epperson
Georgia Institute of Technology
willepp@gatech.edu

Duen Horng Chau
Georgia Institute of Technology
polo@gatech.edu

Chinese CHI 2020, April 26, 2020, Honolulu, HI, USA
ACM 978-1-4503-8815-3/20/04.
10.1145/3403676.3403691

Abstract

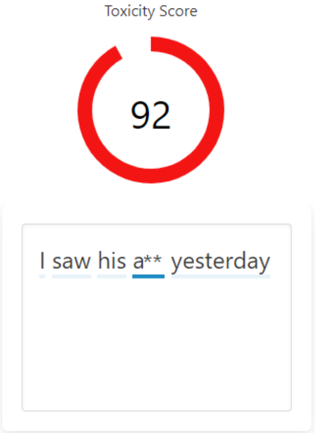
As toxic language becomes nearly pervasive online, there has been increasing interest in leveraging the advancements in natural language processing (NLP) to automatically detect and remove toxic comments. Despite fairness concerns and limited interpretability, there is currently little work for auditing these systems in particular for end users. We present our ongoing work, RECAST, an interactive tool for auditing toxicity detection models by visualizing explanations for predictions and providing alternative wordings for detected toxic speech. RECAST displays the attention of toxicity detection models on user input, and provides an intuitive system for rewording impactful language within a comment with less toxic alternative words close in embedding space. Finally we propose a larger user study of RECAST, with promising preliminary results, to validate it's effectiveness and useability with end users.

Author Keywords

Machine Learning Fairness; Interactive Visualization; Algorithmic Bias; Natural Language Processing

CCS Concepts

•**Human-centered computing** → **Visualization application domains;**



I saw his a** yesterday

Figure 1: A case where a user might notice and flag biases in their model with respect to dialects. The language shown above is non-toxic in the African American English (AAE) dialect, yet RECAST shows a fairly high toxicity score.

Introduction

There is a growing desire to moderate and remove toxic language from social media and public forums, as a result of increasing online interactions [4]. For example in a 2015 user survey of the online social network platform reddit, 50% of people who wouldn't't recommend reddit cited hateful or offensive content and community as the reason why, however on the other hand 35% of complaints from extremely dissatisfied users were about heavy handed moderation and censorship[2]. This balance of issues is further complicated by the high volume of interaction on these platforms, making manual moderation often not tractable and leading to the development of automatic toxicity detection models such as the Google Perspective API [6].

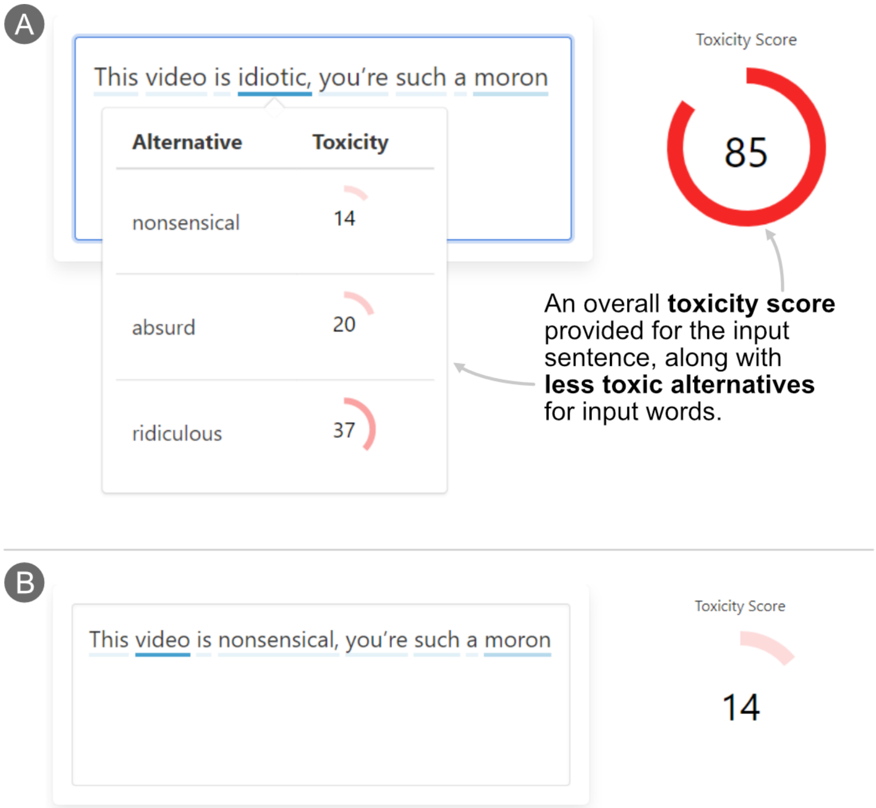
However, the existing body of research highlights potential flaws in the underlying language models for toxicity detection systems. For example, several word embedding models and their training language datasets exhibit biases towards certain subgroups such as gender and race [1, 8]. Some widely deployed NLP models, specifically BERT, also tend to overlook simple linguistic structures like negation, reducing their effectiveness [3]. A serious problem is that it is difficult to fix the potential erroneous model outputs due to a lack of the interpretability of NLP models.

These flaws in language models persist in toxic statement detection systems, especially for end users. Therefore users of online forums that use toxicity detection systems based on black-box NLP models might question how their language is being examined. Nonnative speakers may inadvertently write language that could confuse the system and be marked as toxic. Without tools designed for actual end users to audit what is being detected and make actionable changes to their language, people are disempowered to participate in discourse online. Furthermore, without an ability to detect when a model is falsely flagging language due to either linguistic limitations or social biases, the work of finding and correcting bias are left entirely to the unrepresentative population of machine learning researchers and software engineers. Therefore **the ability to audit these models must be provided to end-users affected by them.**

RECAST

We address these challenges by developing an interactive tool called **RECAST**, which allows for the interrogation of toxicity detection models through counterfactual alternative wording and attention visualization. This design does not require any expertise in machine learning, but enables users to visualize their sentence through the eyes of the algorithm. RECAST currently supports analysis of a fine-tuned BERT model on the Jigsaw Toxicity dataset [5]. Our ongoing work presents the following contributions and vision:

RECAST is an interactive system (Figure 2) that allows users to input text and view the toxicity of the overall sentence, along with which words most contribute to output score. RECAST displays a score between 0 and 100, which represents the probability that the model will classify the input as toxic. Users can view highly attended words, edit the text, select suggested alternate wordings, and watch the toxicity score dynamically update. Its design is purposefully simple, mirroring the text interaction techniques of underlining to note where editing is required and selecting listed alternatives which end users are familiar with from common software like Microsoft Word or Grammarly. This accessibility allows RECAST to effectively communicate the complexities of toxicity detection models using a visual language users are already fluent in.



This video is idiotic, you're such a moron

Alternative	Toxicity
nonsensical	14
absurd	20
ridiculous	37

An overall **toxicity score** provided for the input sentence, along with **less toxic alternatives** for input words.

This video is nonsensical, you're such a moron

Figure 2: A: Recast consists of a textbox and a radial progress bar. A color change on the radial progress, along with a score, indicate the toxicity of a sentence. Toxicity ranges from white (non-toxic) to red (very toxic). Users can hover over options to preview toxicity scores for replacing the selected word in the sentence. **B**: upon replacing the word (in the case of this figure, replacing “idiotic” with “nonsensical”), the main radial progress bar reflects the reduced toxicity score. However, the small attention on the other pejorative word "moron," compared to "video" in the alternative version, shows the idiosyncrasies of the model and underlying dataset.

Attention Visualization

We use attention to explain which words affect our model’s choices. Various visualisation concepts can also be used to show the relative importance of words, such as highlighting and text opacity [9]. However, we utilized an underline on every word, where the opacity of each underline would be controlled by attention placed on each word. We found that using an underline instead of adjusting the opacity or highlighting the word helped with legibility of the text, which is vital for users understanding differences in textual classifications.

Alternative Wording

Alternative wording provides users with options to swap or delete words in a sentence that are responsible for high toxicity scores. Figure 2 highlights such a use of Recast. The underline visualization draws the user’s attention to the most impactful words. When the user hovers over these words, suggested substitutions are shown and ranked by using the k-nearest words from Word2Vec embeddings [7]. Selecting one of these alternatives replaces the word and the new toxicity score is displayed to the right. This mode of interaction is easy and intuitive for users due to its similarity to familiar spellcheck or thesaurus tools and requires little retyping of edits. Furthermore it displays a range of options, which allows the end user agency in maintaining the original meaning as closely as possible. Finally, beyond the act of making the sentence less toxic, the technique allows users to learn which words tend to be highlighted, and what common synonyms the algorithm tends to suggest. This allows people to learn about the model and use this knowledge while writing future comments.

Vision for Future Work

Since the purpose of Recast is to provide power to end-users, an important feature to include is an ability to flag when the model gets it wrong. These examples can be used to provide researchers with data for retraining their models, and provide an avenue of recourse for people adversely affected by the errors in the model. Therefore the statements about the accessibility and usability of RECAST must be validated empirically by a user study. To this end, we plan on evaluating end-users' capacity to reduce toxicity in a sample text given RECAST. In preparation for a full study we have run a small pilot on 18 participants through Amazon Mechanical Turk with approval from the Institutional Review Board, where we found that users given RECAST rated the usefulness of the tool in reducing toxicity on a 5 point scale (higher being more useful) an average of 4.4, compared to 3.6 for a control group. While these results are not highly statistically significant, they help justify further exploration into this tool.

Conclusion

RECAST takes steps towards increased transparency for black-box NLP models that are responsible for moderating large swaths of the internet. By enabling users to interact with text input, view alternative wordings for toxic sentences, and identify potential biases, RECAST provides insights about models to those people actually affected by them, and allows everyone to participate online in both a **less toxic**, and **more fair**, environment.

Acknowledgements

This work was supported in part by NSF grants IIS-1563816, CNS-1704701, NASA NSTRF, gifts from Intel (ISTC-ARSA), NVIDIA, Google, Symantec, Yahoo! Labs, eBay, Amazon.

REFERENCES

[1] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. (2016).

[2] Catherine Buni and Soraya Chemaly. 2016. The secret rules of the internet. (Apr 2016). `https://tinyurl.com/ycl43nq7`

[3] Allyson Ettinger. 2019. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. (2019).

[4] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)* 51, 4 (2018), 85.

[5] Kaggle. 2017. Jigsaw Toxicity Dataset. `https://tinyurl.com/y3fbco5b`. (2017).

[6] Google LLC. 2017. Perspective API. `https://www.perspectiveapi.com/`. (2017).

[7] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 3111–3119.

[8] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 1668–1678. DOI: `http://dx.doi.org/10.18653/v1/P19-1163`

[9] Hernan Valdivieso, Denis Parra, Andres Carvallo, Gabriel Rada, Katrien Verbert, and Tobias Schreck. 2019. Analyzing the Design Space for Visualizing Neural Attention in Text Classification. (2019). `https://tinyurl.com/yzt866vb`