On the definition of toxicity in NLP

Sergey Berezin, Reza Farahbakhsh, Noel Crespi

On the definition of toxicity in NLP

Sergey Berezin, Reza Farahbakhsh, Noel Crespi

SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris
91120 Palaiseau, France
sberezin@telecom-sudparis.eu

Abstract. The fundamental problem in toxicity detection task lies in the fact that the toxicity is ill-defined. This causes us to rely on subjective and vague data in models' training, which results in non-robust and non-accurate results: garbage in - garbage out. This work suggests a new, stress-level-based definition of toxicity designed to be objective and context-aware. On par with it, we also describe possible ways of applying this new definition to dataset creation and model training.

1 Introduction

The toxicity detection task's fundamental problem lies in the toxicity being ill-defined. Jigsaw, a unit within Google and one of the leaders in the field uses a definition of toxicity given by Dixon et al. [1] - "rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion". One can instantly see the issue with this definition, as it gives no quantitative measure of the toxicity and operates with highly subjective cultural terms. Despite all vagueness and flaws, this definition is widely used by many researchers [2].

More than that, lack of proper definition causes researchers to use different terms for similar work [3], thus producing labels "toxic", "hateful", "offensive", "abusive", etc. for effectively the same data [4].

The subjectivity of current definitions also causes disagreement in labelling, causing drastically different results in models' training [5] . Another problem lies in the assumption that toxicity labelling can be done without a proper understanding of the context of each message [6].

This causes racial, sexual, political, religious and geographical bias in datasets [2] [7] and also significantly lowers the performance of toxicity detectors, effectively turning them into profanity detectors [8] without any contextual awareness [9].

2 In search for a formal definition

First of all, we should be able to measure toxicity objectively.

In machine learning, we need to have a target metric, such as an amount of money or a category of an animal, that we approximate with some kind of loss function, which we use to solve the optimisation task. Without a clear and measurable target metric, we are unable to confidently tell if our loss function, and therefore our model, is any good at all.

To find an appropriate metric, first of all, we need to think of the objective we want to achieve. The main objective of toxicity detection is to save people from stress, which

is caused by disturbances in social interactions, undesirable social roles, criticism, self-criticism and unfair treatment [10]. The stress-evoking mechanism lies in the fact that insults directed at a person pose a severe threat to them and their reputation. More than that, even witnessing the infliction of harm on others causes stress since it shows potential group aggressors and signals social conflict in the vicinity [11].

So, the key metric for us is the level of toxicity-caused stress a person experiences.

However, what and why is considered an insult, criticism or any other undesirable interaction?

Let us follow the process of word understanding. After we hear a word, we step into the lexical retrieval phase, during which words are compared to the known social intention or emotional stance of the speaker using this word, the impact on listeners of this word, or the nature of the things referred to by this word. After understanding these correlations, the release of stored emotional meaning is used as a prediction of what the word means in this particular context [12].

So, our reaction to words is based on our knowledge of the meaning of the words and their acceptance in our morality and communication norms [13]. We, as a society, define norms to prevent what causes stress to us, and we, as a person, feel stress when those norms are violated [14] [15]. From that, we can see that toxicity causes stress by contradiction of accepted morality and norms of communication.

However, morality and norms of communication vary a lot across the world and even across the time [16]. [14]. While we should not ban a British person for saying that they are going to have a "fag break" when they are going to smoke a cigarette, we should definitely ban somebody who is using the word "fag" as a derogatory name for a homosexual person. In the same way, the word "nice" is perfectly fine in modern English but should be considered toxic when analysing archive letters from the 14th century since it meant "foolish" back then. This leads us to include situational and verbal context in our account.

Using that, let us formulate the definition: **Toxicity** - is a characteristic of causing stress by contradiction of accepted morality and norms of interaction with respect to the situational and verbal context of interaction.

From this: **Toxic speech** - is such a speech that causes stress by contradiction of accepted morality and norms of communication with respect to the situational and verbal context of interaction.

2.1 Measurements

The question arises: How do we determine the toxicity of something?

We can do so by measuring the level of stress. The first way to do so is to register increased levels of stress hormones (e.g. cortisol or catecholamine levels), which can be measured in blood, saliva or urine samples. However, this might not be easy to do en masse.

We can also use other, non-invasive methods, such as registration of changes in heart rate, blood pressure, pupil diameter, breathing pattern, galvanic skin response, emotion, voice intonation, and body pose [17]. It is shown that using a combination of such measurements can achieve very high quality ($> 90\%$ acc.) estimation of stress levels while keeping the testing procedure fast and simple [10].

An even more scalable method is to approach the question from the side of the violation of accepted morality and norms. It is shown that those violations cause stress (i.e., they can be used as a proxy in stress registering) and cause negative responses from other people. By noticing these responses, we can detect the fact of norm violation.

More than that, responses are typically intended to modify the violator's behaviour and to strengthen a violated norm - this gives us a way to not only detect the fact of norm violation but to pinpoint a specific norm and the way of its defyition.[18].

A negative reaction to norms' breaching is culture-universal and proportional to the inappropriateness of the triggering behaviour. People consider it more appropriate to use gossip, social isolation, and confrontation the more severe violation of norms is perceived to be. Thus, by registering a negative reaction, we can also estimate the harshness of a norm violation and, subsequently, the level of stress this violation causes [19].

Cross-cultural comparisons have also highlighted how idiosyncratic social norms are. What is considered appropriate in one culture can be seen as highly deviant in another culture. However, norms are consistent within countries and largely independent of the domain of a norm violation [16].

2.2 Application and further development

An automatic way of determining social norms and their violations is described in [20]. In this work, authors used language models to analyse 2.8M comments removed by moderators of 100 top Reddit sub-forums over ten months.

Following their idea, we suggest use LLM for creating a dataset based on the proposed definition. The steps are as follows:

1. Fine-tune a model such as Llama 2 or GPT-4 on corpora of fixed context (i.e. time, region, social group, etc.). An example of such might be a corpus of British literature of the 17th century.
2. Use this fine-tuned model to detect norms' breaches by negative reactions caused by them. Few-shot learning with hand-crafted examples from the corpora might be applied here to ease the task.
3. Extract the norms violated and label them.

This will allow us to build situational and verbal context-aware models for toxic speech detection and will help us solve existing drawbacks of toxicity detection, eliminating its biases and increasing its performance.

References

1. Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. pages 67–73, 12 2018.
2. Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s), jul 2023.
3. Bertie et al. Vidgen. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy, August 2019. Association for Computational Linguistics.

4. Paula et al. Fortuna. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, 2020. ELRA.
5. Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. The authors matter: Understanding and mitigating implicit bias in deep text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 74–85, Online, August 2021. Association for Computational Linguistics.
6. Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*, 2021.
7. Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.
8. Edwin Chen. Holy $#!t: Are popular toxicity models simply profanity detectors? https://www.surgehq.ai/blog/are-popular-toxicity-models-simply-profanity-detectors, 2022. Accessed: 2023-09-01.
9. Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online, August 2021. Association for Computational Linguistics.
10. Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, PP:1–1, 07 2019.
11. Marijn E Struiksma, Hannah NM De Mulder, and Jos JA Van Berkum. Do people get used to insulting language? *Frontiers in Communication*, 7:910023, 2022.
12. Bryor Snefjella, Nadia Lana, and Victor Kuperman. How emotion is learned: Semantic learning of novel words in emotional contexts. *Journal of Memory and Language*, 115:104171, 2020.
13. Robert B Cialdini and Noah J Goldstein. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621, 2004.
14. Gerben A Van Kleef, Michele J Gelfand, and Jolanda Jetten. The dynamic nature of social norms: New perspectives on norm development, impact, violation, and enforcement, 2019.
15. Robert M Sapolsky. *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping*. Holt paperbacks, 2004.
16. Kimmo Eriksson, Pontus Strimling, Michele Gelfand, Junhui Wu, Jered Abernathy, Charity S Akotia, Alisher Aldashev, Per A Andersson, Giulia Andrighetto, Adote Anum, et al. Perceptions of the appropriate response to norm violation in 57 societies. *Nature communications*, 12(1):1481, 2021.
17. Nandita Sharma and Tom Gedeon. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, 108(3):1287–1301, 2012.
18. Gerben A Van Kleef, Florian Wanders, Eftychia Stamkou, and Astrid C Homan. The social dynamics of breaking the rules: Antecedents and consequences of norm-violating behavior. *Current Opinion in Psychology*, 6:25–31, 2015.
19. Jörg Gross and Alexander Vostroknutov. Why do people follow social norms? *Current Opinion in Psychology*, 44:1–6, 2022.
20. Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25, 2018.