CommTox: Contextually-Aware Community Perceived Toxicity Classification

Ayan Chowdhury, Rhett Hanscom [⋆], Tamara Lehman, Qin Lv, and Shivakant Mishra

University of Colorado, Boulder CO 80309,
Rhett.Hanscom@Colorado.edu,
www.colorado.edu/center/demtech/

Abstract. CommTox, a community toxicity classifier, leverages machine learning and historical behavioral data in order to assign a score for the perceived toxicity within a community. Building contextual awareness on top of Perspective API, CommTox brings highly flexible and adaptable context-aware toxicity classification to developers. CommTox is deployed across the social media platform of YouTube, and evaluates perceived toxicity across a number of online communities (users interacting with similar content over time). Results indicate that while community reception of comments offers a fair amount of accuracy when predicting perceived toxicity, the most effective models need to include some level of text-dependent features (such as word-embeddings). Additionally, the inclusion of context-free toxicity scoring allows projects to use CommTox across platforms and communities. The goal of CommTox is to provide a standardized metric from which comparisons free of bias can be drawn for varying communities.

Keywords: Toxicity, hate-speech, censorship, machine learning

1 Introduction

Toxicity classification has long challenged researchers due to the inherently subjective and context-dependent nature of toxic language. Traditional systems often treat comments in isolation, ignoring conversational or community context [35, 1, 9, 4, 21]. This is further complicated by annotator bias [5], demographic variance in speech perception, and platform-specific manifestations of toxicity [6]. Without standardized definitions, many studies develop individual classification criteria, limiting cross-comparison.

Most social media platforms (SMPs) assess toxicity at the comment level using natural language processing (NLP), but few incorporate user behavior or community norms [10]. While the Perspective API offers a scalable context-free solution, it has notable limitations including a lack of contextual awareness and susceptibility to adversarial inputs [11].

In this work, we propose CommTox, a contextually-aware toxicity classifier that augments Perspective API scores with community behavior data. Rather

[⋆] The first two listed authors contributed equally to this work.

than focusing solely on how a comment is written, CommTox examines how it is received, capturing subtle signals of *perceived toxicity.* This enables measurement of three key indicators: (1) alignment between predicted toxicity and community judgment, (2) the resilience of communities to toxicity, and (3) broader trends in toxic discourse.

Despite an expanding body of literature, standardized tools for analyzing toxicity at the community level remain scarce [22–26]. As a result, researchers must develop custom tools, which hinder reproducibility and scalability. This gap is even more pronounced for general users, who lack tools to assess toxicity patterns in their online spaces without technical expertise.

Comment-level toxicity detection is now commonplace, but it fails to capture the nuances of community interaction. Not all negative comments are toxic, for instance rudeness can exist with productive discourse depending on context [7]. A deeper understanding of how communities engage with content enables better classification of toxicity, detection of effective de-escalation tactics, and the identification of resilient communities.

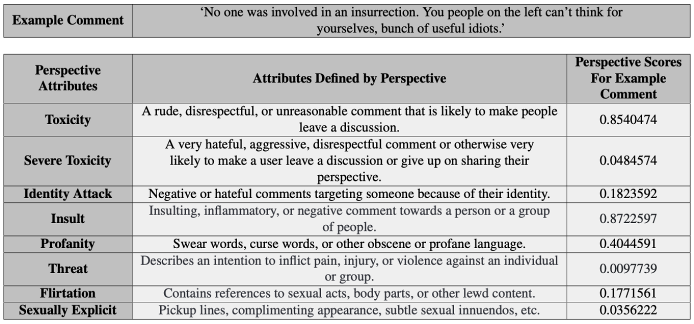
We define *toxicity resilience* as a community's ability to transform toxic content (as flagged by Perspective) into constructive dialogue. For example, minority groups often reclaim slurs as terms of empowerment [27]; while Perspective may flag these as toxic, the community may not perceive them as such. CommTox captures these discrepancies by considering response patterns rather than relying solely on content.

Prior work has shown the value of advanced machine learning models in toxicity classification. Zaheri *et al.* found that LSTMs outperformed traditional classifiers in identifying harmful content [32]. Udhayakumar *et al.* demonstrated the effectiveness of logistic regression when enhanced with context-aware recommendations [4]. Anuchitanukul *et al.* highlighted how conversational structure influences toxicity perceptions [10]. Beyond individual comments, several studies have explored user and community dynamics. Mall *et al.* classified users based on long-term toxic behavior, aiding moderation [33], while Almerekhi *et al.* investigated how Reddit users shift into toxicity across communities [34]. Despite these advances, few models are tailored to visually oriented platforms like YouTube, and fewer still offer transferable frameworks that quantify how toxicity is perceived across diverse communities.

CommTox bridges this gap by combining NLP-based toxicity scores with features derived from community reception, allowing for scalable and context-aware moderation across platforms. It also opens the door to implementing preemptive interventions when communities approach tipping points.

2 Methodology

To assign context-free toxicity scores, CommTox builds on the Perspective API developed by Jigsaw and Google [11]. This machine learning tool scores user-generated content across attributes such as *toxicity*, *severe_toxicity*, *identity_attack*, *insult*, *profanity*, and *threat.* While widely adopted for moderation on platforms like Reddit and The New York Times, its use as a foundation for community-aware models remains underexplored. Experimental attributes such as *flirtation* and *sexually_explicit* were omitted due to instability.



Example Comment	'No one was involved in an insurrection. You people on the left can't think for yourselves, bunch of useful idiots.'

Perspective Attributes	Attributes Defined by Perspective	Perspective Scores For Example Comment
Toxicity	A rude, disrespectful, or unreasonable comment that is likely to make people leave a discussion.	0.8540474
Severe Toxicity	A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.	0.0484574
Identity Attack	Negative or hateful comments targeting someone because of their identity.	0.1823592
Insult	Insulting, inflammatory, or negative comment towards a person or a group of people.	0.8722597
Profanity	Swear words, curse words, or other obscene or profane language.	0.4044591
Threat	Describes an intention to inflict pain, injury, or violence against an individual or group.	0.0097739
Flirtation	Contains references to sexual acts, body parts, or other lewd content.	0.1771561
Sexually Explicit	Pickup lines, complimenting appearance, subtle sexual innuendos, etc.	0.0356222

Fig. 1: Example of Perspective API scores alongside attribute definitions [11].

Perspective scores range from 0 to 1, with thresholds between 0.7 and 0.9 typically used to flag toxicity. We lower this cutoff to 0.65 to better capture borderline cases that often trigger community responses. However, Perspective's limitations, including a lack of contextual awareness, annotator bias, and vulnerability to adversarial inputs, necessitate additional behavioral modeling [12, 13, 5, 8, 1].

CommTox evaluates perceived toxicity using YouTube data from May through September 2024. We collected comments from trending videos across five content categories: Politics, Sports, Music, Movies, and Technology. Topics were identified using `pytrends` [37] and categorized via GPT-4o [38]. Additional search terms were generated when categories lacked sufficient coverage. This approach has been shown effective at similar tasks surrounding human-like text generation [36]. Video IDs were retrieved using the YouTube Data API, and only recent content (past six weeks) was scraped to ensure relevance.

Communities were defined as sets of comments on videos belonging to playlists aligned with the five categories. After removing duplicates, we filtered non-English comments, applied standard preprocessing (e.g., lemmatization, stop-word removal), and converted emojis to text to preserve sentimental nuances. Toxicity labels were generated using Perspective, and threads (nested sequences of comments and replies forming a distinct conversations both separate from and apart of the larger discourse on a video) were further categorized based on reply behavior.

Finally, comments above the toxicity threshold with replies were examined for response patterns. Toxic comments without replies were ignored as community reception would be difficult to infer using only the comment itself. CommTox labels a thread as toxic if replies also score highly, and resilient if replies are predominantly non-toxic. These interaction patterns form the behavioral context that informs our classifier.

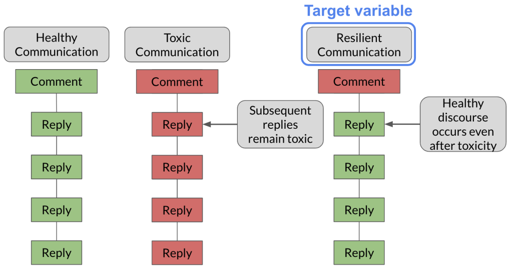


Fig. 2: Community behavior patterns: **Healthy**, **Toxic**, and **Resilient**. Red indicates Perspective-flagged toxicity; green indicates low toxicity. Blue circles mark classifier targets.

3 The CommTox Classifier

Perceived toxicity, or how online communities receive comments labeled as toxic by systems such as the Perspective API, is modeled and assessed though use of the Random Forest (RF) within CommTox. Unlike context-free classification, CommTox integrates user responses, enabling finer-grained detection of community-specific norms and *toxicity resilience*.

A comment is labeled as perceived toxic if (1) its Perspective toxicity score is ≥ 0.65, (2) it has at least one reply, and (3) $\geq 30\%$ of those replies are also toxic. Conversely, resilient communication occurs when replies remain largely non-toxic.

3.1 Model Architecture

CommTox employs a Random Forest classifier composed of decision trees trained with bootstrapped samples and random feature subsets [14, 2]. This approach reduces overfitting and improves generalizability. Each tree votes independently, and predictions are aggregated via majority vote. Feature importance is extracted post-training to assess which variables drive predictions [15].

3.2 Feature Sets

CommTox integrates three categories of features: (1) **CFT**: Context-free Perspective API scores [11], (2) **WEF**: Word embeddings from Google's Word2Vec, averaged per comment [3], (3) **Non-WEF**: Eleven engineered features capturing lexical and structural patterns (e.g., toxic word count [16], punctuation, emoji use, word counts, log-like counts). Together, these features capture both textual signals and engagement-driven context across platforms.

3.3 Model Variants

We evaluate two model configurations:

1. **CFT + Non-WEF**: Baseline numeric model with 80% accuracy.
2. **CFT + WEF**: In addition to previous features, adds embeddings and TF-IDF, improving performance to 83%.

To address class imbalance, we apply SMOTE [20], and augment training data via synonym substitution [18]. Community-specific classifiers are trained by splitting data per content domain (e.g., Politics, Sports), capturing interaction nuances in each community separately.

3.4 Context-Free vs. Perceived Toxicity

Context-free toxicity (CFT) offers standardized, scalable tagging but ignores discourse structure, speaker identity, or cultural re-appropriation (such as reclaimed slurs [27]). Perceived toxicity accounts for community reception: comments flagged as toxic by Perspective but received positively may be misclassified. CommTox bridges this gap by incorporating community behavior.

Communities also differ in *toxicity resilience*, or their ability to de-escalate toxic threads. High-resilience communities respond to toxicity with civility, while low-resilience groups may amplify it. CommTox quantifies these dynamics via reply-chain analysis.

3.5 Model Training

We focus training on top-level comments (not replies), since these are more visible within SMPs. Figure 2 shows three reply-chain patterns:

1. **Healthy**: Non-toxic comment followed by non-toxic replies.
2. **Toxic**: Toxic comment triggers toxic replies.
3. **Resilient**: Toxic comment receives non-toxic replies.

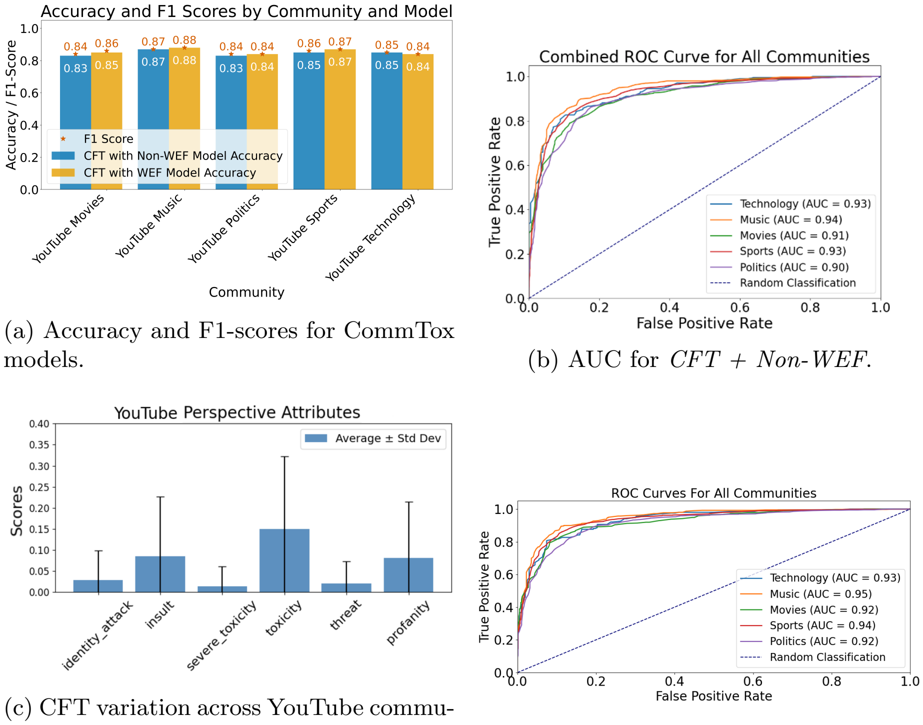
CommTox classifies comment threads into these categories based on Perspective scores and reply behavior.

3.6 Model Optimization

We use GridSearchCV to optimize RF hyperparameters (tree depth, estimators, etc.) [19]. Each model is trained and evaluated using an 80/20 temporal split (early data for training, later data for testing). Separate models are fine-tuned for each community to better capture domain-specific language and behavioral patterns. For each community, we produce a complete feature set by combining the TF-IDF vectorized text data with additional numerical features and embeddings.

4 Results and Discussion

CommTox is evaluated using data from the 14 oldest days for training and the rest for testing. Two model variants are compared: *CFT + Non-WEF* and *CFT + WEF*. Figure 3 shows that the WEF model consistently outperforms the non-WEF model across communities. Accuracy and F1 scores (Fig. 3a) are higher for the WEF model, and ROC curves (Figs. 3b, 3d) show AUC scores of 0.94 and 0.97, respectively. Figure 3c shows variation in baseline toxicity across domains.



(a) Accuracy and F1-scores for CommTox models.

(b) AUC for *CFT + Non-WEF*.

(c) CFT variation across YouTube communities.

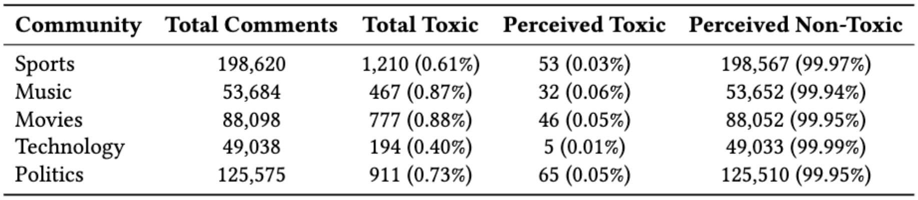
(d) AUC for *CFT + WEF*.

Fig. 3: Model performance metrics across communities.

4.1 Community Toxicity Resilience

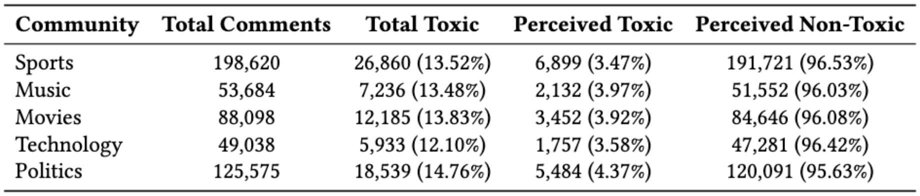
To assess community response patterns, we analyzed reply chains for head comments with high CFT scores (≥ 0.65). A thread is labeled *perceived toxic* if at least one reply is also toxic; otherwise, it is *perceived non-toxic*. Figures 4a and 4b summarize results across five communities.

While the WEF model detects more toxic comments (e.g., 13.5% in Sports), over 95% of these threads do not escalate—indicating strong community re-



Community	Total Comments	Total Toxic	Perceived Toxic	Perceived Non-Toxic
Sports	198,620	1,210 (0.61%)	53 (0.03%)	198,567 (99.97%)
Music	53,684	467 (0.87%)	32 (0.06%)	53,652 (99.94%)
Movies	88,098	777 (0.88%)	46 (0.05%)	88,052 (99.95%)
Technology	49,038	194 (0.40%)	5 (0.01%)	49,033 (99.99%)
Politics	125,575	911 (0.73%)	65 (0.05%)	125,510 (99.95%)

(a) Toxicity chain metrics: Non-WEF model.



Community	Total Comments	Total Toxic	Perceived Toxic	Perceived Non-Toxic
Sports	198,620	26,860 (13.52%)	6,899 (3.47%)	191,721 (96.53%)
Music	53,684	7,236 (13.48%)	2,132 (3.97%)	51,552 (96.03%)
Movies	88,098	12,185 (13.83%)	3,452 (3.92%)	84,646 (96.08%)
Technology	49,038	5,933 (12.10%)	1,757 (3.58%)	47,281 (96.42%)
Politics	125,575	18,539 (14.76%)	5,484 (4.37%)	120,091 (95.63%)

(b) Toxicity chain metrics: WEF model.

Fig. 4: Reply-chain analysis of perceived toxicity across communities.

silience. The Non-WEF model detects fewer cases but shows even higher non-escalation rates (99.9%). In 1,327 conflicting cases, the WEF model flagged comments as toxic but found replies remained non-toxic in 88% of cases, reinforcing the need for contextual interpretation.

4.2 Implications and Limitations

Despite variations in baseline toxicity, most communities demonstrate de-escalatory behavior. For example, in Politics (our most toxic community) 14.8% of comments are flagged, but only 4.4% of threads become perceived toxic. This suggests moderation efforts should focus not only on detection but also on fostering resilience.

CommTox highlights the benefits of integrating community response into toxicity classification. Behavior-aware models outperform content-only approaches in both sensitivity and interpretability. However, reliance on Perspective API introduces limitations: annotator bias, poor slang handling, and adversarial vulnerabilities [5, 8]. Pre-trained Word2Vec embeddings may also miss platform-specific language nuances. Data coverage can vary by topic due to API rate limits and user activity imbalance.

4.3 Conclusion

CommTox offers a context-aware framework for assessing toxicity by incorporating both linguistic features and community response. By distinguishing between context-free toxicity and perceived toxicity, it identifies not just harmful content but also the community's ability to de-escalate. Our findings underscore the importance of feature-rich, behaviorally grounded models for moderation, particularly on visually oriented platforms such as YouTube. CommTox is generalizable across SMPs and provides a scalable metric for comparing toxicity and resilience across communities.

References

1. Kumar, Deepak, Kelley, Patrick Gage, Consolvo, Sunny, Mason, Joshua, Bursztein, Elie, Durumeric, Zakir, Thomas, Kurt, Bailey, Michael: Designing toxic content classification for a diversity of perspectives.SOUPS (2021)
2. Altman, Naomi, Krzywinski, Martin: Ensemble methods: bagging and random forests. Nature Methods, vol. 14(10), pp. 933–935 (2017)
3. Hugging Face, fse: word2vec-google-news-300. (2025)
4. Udhayakumar, S, Silviya Nancy, J, UmaNandhini, D, Ashwin, P, Ganesh, R: Context aware text classification and recommendation model for toxic comments using logistic regression. ICBDCC 2019 (2021)
5. Sap, Maarten, Swayamdipta, Swabha, Vianna, Laura, Zhou, Xuhui, Choi, Yejin, Smith, Noah A. “Annotators with attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection”. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. (2022)
6. Singh, Ashwini Kumar, Ghafouri, Vahid, Such, Jose, Suarez-Tangil, Guillermo: Differences in the Toxic Language of Cross-Platform Communities. Proceedings of the International AAAI Conference on Web and Social Media (2024)
7. Sheth, Amit, Shalin, Valerie L, Kursuncu, Ugur: Defining and detecting toxicity on social media: context and knowledge are key. Neurocomputing (2022)
8. Rieder, Bernhard, Skop, Yarden: Studying the technical, normative, and organizational structure of Perspective API. Big Data & Society, vol. 8 (2021)
9. Eke, Christopher Ifeanyi, Norman, Azah Anir, Shuib, Liyana: Context-based feature technique for sarcasm identification in benchmark datasets using deep learning and BERT model. IEEE Access, vol. 9 (2021)
10. Anuchitanukul, Atijit, Ive, Julia, Specia, Lucia: Revisiting contextual toxicity detection in conversations. ACM Journal of Data and Information Quality (2022)
11. Google Jigsaw: Perspective API Documentation. (2025)
12. Hosseini, Hossein, Kannan, Sreeram, Zhang, Baosen, Poovendran, Radha: Deceiving google’s perspective api built for detecting toxic comments arXiv:1702.08138 (2017)
13. Jain, Edwin, Brown, Stephan, Chen, Jeffery, Neaton, Erin, Baidas, Mohammad, Dong, Ziqian, Gu, Huanying, Artan, Nabi Sertac: Adversarial Text Generation for Google’s Perspective API. CSCI (2018)
14. Ali, Jehad, Khan, Rehanullah, Ahmad, Nasir, Maqsood, Imran: Random forests and decision trees. IJCSI (2012)
15. Archer, Kellie J, Kimes, Ryan V: Empirical characterization of random forest variable importance measures. Computational statistics & data analysis (2008)
16. Luis von Ahn. ”List of Bad Words”. (2025). cs.cmu.edu/~biglou/resources
17. Ramos, Juan. ”Using tf-idf to determine word relevance in document queries.” Instructional conference on machine learning, vol. 242, no. 1, pp. 29-48. 2003.
18. Shorten, Connor, Taghi M. Khoshgoftaar, and Borko Furht. ”Text data augmentation for deep learning.” Journal of big Data 8, no. 1 (2021): 101.
19. Scikit-Learn Developers: Grid Search - Scikit Documentation. (2025)

20. Chawla, Nitesh V, Bowyer, Kevin W, Hall, Lawrence O, Kegelmeyer, W Philip: SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research (2002)
21. Dadvar, Maral, Trieschnigg, Dolf, Ordelman, Roeland, De Jong, Franciska: Improving cyberbullying detection with user context. ECIR (2013)
22. Wang, Wenjie, Feng, Fuli, Nie, Liqiang, Chua, Tat-Seng: User-controllable recommendation against filter bubbles. SIGIR conference on research and development in information retrieval(2022)
23. Salminen, Joni, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. "Developing an online hate classifier for multiple social media platforms." Human-centric Computing (2020).
24. Saveski, Martin, Roy, Brandon, Roy, Deb: The structure of toxic conversations on Twitter. Proceedings of the web conference (2021)
25. Rupapara, Vaibhav, Rustam, Furqan, Shahzad, Hina Fatima, Mehmood, Arif, Ashraf, Imran, Choi, Gyu Sang: Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. IEEE Access (2021)
26. Akuma, Stephen, Lubem, Tyosar, Adom, Isaac Terngu: Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. International Journal of Information Technology (2022)
27. Worthen, Meredith GF: Queer identities in the 21st century: reclamation and stigma. Current Opinion in Psychology (2023)
28. Madhu, Hiren, Satapara, Shrey, Modha, Sandip, Mandl, Thomas, Majumder, Prasenjit: Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments. Expert Systems with Applications (2023)
29. Reimers, Nils, and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing. 2019.
30. Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." Neural computation 9, no. 8 (1997): 1735-1780.
31. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Computational Linguistics: Human Language Technologies (2019)
32. Zaheri, Sara, Leath, Jeff, Stroud, David: Toxic comment classification. SMU Data Science Review (2020)
33. Mall, Raghvendra, Nagpal, Mridul, Salminen, Joni, Almerekhi, Hind, Jung, Soon-Gyo, Jansen, Bernard J: Four types of toxic people. Nordic Conference on Human-Computer Interaction (2020)
34. Almerekhi, Hind, Kwak, Haewoon, Jansen, Bernard J: Investigating toxicity changes of cross-community redditors from 2 billion posts and comments. PeerJ Computer Science (2022)
35. Pavlopoulos, John, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. "Toxicity Detection: Does Context Really Matter?." Annual Meeting of the Association for Computational Linguistics, pp. 4296-4305. 2020.
36. Orrù, Gabriele, Biancofiore, Francesco, Qian, Wenjie, Ji, Yang, Sapienza, Andrea, Cambria, Erik: Human or not? Information Fusion (2023)
37. Hogue, John. "pytrends: Pseudo API for Google Trends." GitHub repository, https://github.com/GeneralMills/pytrends. Accessed July 8, 2025.
38. OpenAI. "GPT-4o Technical Report." OpenAI, May 2024. https://openai.com/research/gpt-4o