Exploring Toxic Language on Stormfront: Machine Learning vs. Dictionary-Based Detection

Alec Dorkadesler
Department of Computer and Systems Sciences, DSV
Stockholm University
Stockholm, Sweden
alecdortkardesler@gmail.com

Ran Duan
Department of Computer and Systems Sciences, DSV
Stockholm University
Stockholm, Sweden
Owdran@gmail.com

Alem Woldeyohannes
Department of Computer and Systems Sciences, DSV
Stockholm University
Stockholm, Sweden
alemtibebuw@gmail.com

Xiaochun Li
Department of Computer and Systems Sciences, DSV
Stockholm University
Stockholm, Sweden
lixiaochuna@gmail.com

Houye Dong
Department of Computer and Systems Sciences, DSV
Stockholm University
Stockholm, Sweden
salahdhy@gmail.com

***Abstract*—This report examines toxic language detection on Stormfront using two approaches: machine learning and a custom dictionary. The study compares their effectiveness in identifying hate speech, racism, antisemitism, and misogyny. Machine learning showed moderate agreement (Krippendorf's Alpha = 0.791), while the dictionary method performed less consistently. Annotator biases and subtle toxicity detection challenges highlight the need for refined models.**

Keywords—toxic language, hate speech detection, machine learning, dictionary-based approach, Stormfront

I. Introduction

Toxic language, characterized by rudeness, disrespect, or unreasonable comments that provoke emotional responses, is a significant challenge in online discussions [1]. With the rise of far-right forums like Stormfront.org, identifying and mitigating toxic language is essential to foster healthier online communities. For this study, we define toxic language as comments likely to disrupt discussions or drive away users.

The primary goal of this assignment is to compare two approaches to detecting toxic language: machine learning through the HateScan API and a custom-built dictionary. These methods were applied to a dataset of 5,000 comments sourced from Stormfront.org, an online community known for its extremist ideologies and hate speech [2]. By examining how each approach identifies toxic comments, we aim to explore their effectiveness, the types of toxic language present, and the impact of annotator agreement on the final results.

Our research question focuses on understanding the manifestation of toxic language in online extremist forums and evaluating how machine learning and dictionary-based approaches differ in identifying toxicity. Additionally, we aim to analyze the specific groups targeted by toxic language and how the nature of this language shifts depending on the target group. This report presents a detailed analysis of the data, methods, results, and discussions surrounding the use of these tools to identify toxic discourse online.

This study not only compares two toxic language detection methods—machine learning and dictionary-based approaches—but also highlights the challenges of accurately identifying nuanced hate speech. By understanding these challenges, we aim to contribute to the broader conversation of improving tools for online content moderation and ensuring safer digital environments.

II. Method

A. Data

For this analysis, we sourced 5,000 comments from Stormfront.org, a known white nationalist forum. The dataset was processed to analyze toxicity through two distinct methods: machine learning via the HateScan API and a custom dictionary-based approach. The machine learning method leverages probabilistic models to detect toxic language based on patterns from pre-trained datasets, allowing it to catch nuanced cases that are less reliant on specific keywords. In contrast, the dictionary-based approach identifies toxicity strictly through the precedence of predefined terms, making it context-insensitive and prone to overlooking coded or subtle forms of hate speech [3]. Comments varied in target groups, including ethnicities, religions, and genders, reflecting the nature of the toxic discourse common to this forum.

B. Machine Learning Analysis

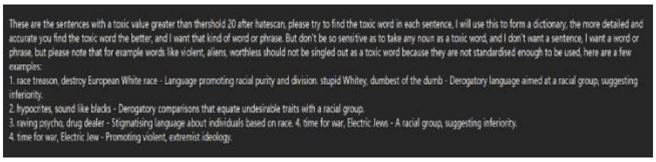
We employed the HateScan API to assess toxicity and threat levels for each comment. The API returns a probability score representing the perceived level of toxicity, and we set a threshold of 50% for classification. Comments scoring 50% or higher were labeled as toxic, while those below this threshold were considered non-toxic. The decision to set a 50% threshold was based on our observation that comments scoring above this level appeared consistently toxic, whereas scores below 50% required further discussion.

The Python script processed each comment through the API, and the resulting prediction scores were stored for later manual annotation.

Out of the 5,000 comments analyzed, 377 comments received a prediction score of 50% or higher, indicating toxicity. Similarly, 4,623 comments had a score below 50% and were considered non-toxic. To ensure a representative sample, we used a sample size calculator with the following specifications: a confidence level of 95%, a margin of error of 5%, and a population proportion of 50%. For the toxic comments, this yielded a sample size of 191 comments from the 377 toxic comments. For non-toxic comments, the same specifications resulted in a sample size of 355 comments from the 4,623 non-toxic comments.

C. Dictionary-Based Approach

Our dictionary was developed by compiling terms associated with toxic language, initially using a dataset of 600 entries flagged as toxic. To balance inclusivity and specificity, we adjusted the threshold to 20%, ensuring a broader sample size compared to the previous 50% threshold. The first draft of the dictionary was created by refining ChatGPT prompts to optimize efficiency, as shown in Figure 1 provided below.



These are the sentences with a toxic value greater than thershold 20 after hatescan, please try to find the toxic word in each sentence, I will use this to form a dictionary, the more detailed and accurate you find the toxic word the better, and I want that kind of word or phrase. But don't be so sensitive as to take any noun as a toxic word, and I don't want a sentence, I want a word or phrase, but please note that for example words like violent, aliens, worthless should not be singled out as a toxic word because they are not standardised enough to be used, here are a few examples:

1. race treason, destroy European White race - Language promoting racial purity and division. stupid Whitey, dumbest of the dumb - Derogatory language aimed at a racial group, suggesting inferiority.
2. hypocrites, sound like blacks - Derogatory comparisons that equate undesirable traits with a racial group.
3. raving psycho, drug dealer - Stigmatising language about individuals based on race. 4. time for war, Electric Jews - A racial group, suggesting inferiority.
4. time for war, Electric Jew - Promoting violent, extremist ideology.

Figure 1

We also incorporated terms from external data sources to broaden its scope. The dictionary was refined by removing irregular expressions, overly long phrases, and unnecessary punctuation. Initial scans flagged toxic language, but 9% of flagged comments were classified as non-toxic, signaling a need for further refinement. Each team member reviewed content, adding new terms and adjusting entries for better accuracy.

Following this iteration, we conducted a second round of scanning. The results were significantly improved, with the flagged instances categorized as follows:

- **Toxic**: 3,983 instances, with 235 entries selected for detailed examination.
- **Non-Toxic**: 1,017 instances, with 101 entries selected for validation.

This iterative approach enabled us to fine-tune the dictionary continuously and improve the system's capability to detect toxic language effectively. By incorporating diverse insights and systematically refining the dictionary, we strike a balance between sensitivity (minimizing false negatives) and specificity (reducing false positives) in identifying toxic language across various contexts. We continue to explore ways to enhance the dictionary, such as addressing contextual constraints, word combinations, negation handling, and tone indicators.

D. *Annotation and Agreement*

To ensure thorough analysis, five group members annotated each of the 191 toxic and 355 non-toxic comments. Each annotator classified comments as either toxic (1) or non-toxic (0). We used the mode function in Excel to determine the majority decision for each comment. Majority decisions were used to finalize classifications, while differences in annotator opinions were considered, especially regarding cultural and gender biases.

To measure inter-annotator agreement, we calculated Krippendorf's Alpha for both the machine learning and dictionary-based approaches. The machine learning analysis yielded a K-Alpha of 0.791, indicating moderate agreement [4]. The dictionary-based approach produced a lower K-Alpha of 0.607, which we attribute to over-reliance on keyword matching without considering the context of comments.

E. *Ethical Considerations*

Given the sensitive and offensive nature of the data, all group members were briefed on the potential emotional impact of reviewing toxic comments. We also ensured that comments flagged as toxic were treated with care, recognizing the broader cultural, racial, and gender biases present. Furthermore, we limited the dissemination of sensitive data outside of this project and ensured that it was handled ethically and securely.

III. Results

A. *Toxicity Prediction and Classification*

Using the HateScan API, we analyzed 5,000 comments from Stormfront.org to determine toxicity scores. We set a threshold of 50%, labeling comments with a score of 50% or higher as toxic and those below 50% as non-toxic. The machine learning model identified 377 toxic comments and 4,623 non-toxic comments. A sample size of 191 toxic and 355 non-toxic comments was generated using the specifications previously mentioned.

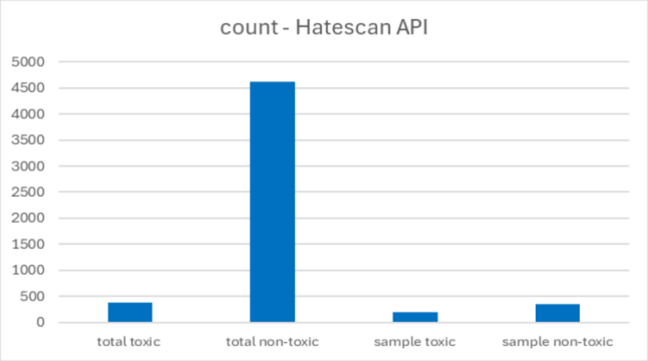


Figure 2

Using the dictionary model, 3,983 toxic instances were identified. A sample size of 235 entries were chosen randomly for closer examination, along with 1,017 non-toxic instances, and a sample size of 101 entries were selected randomly for validation.

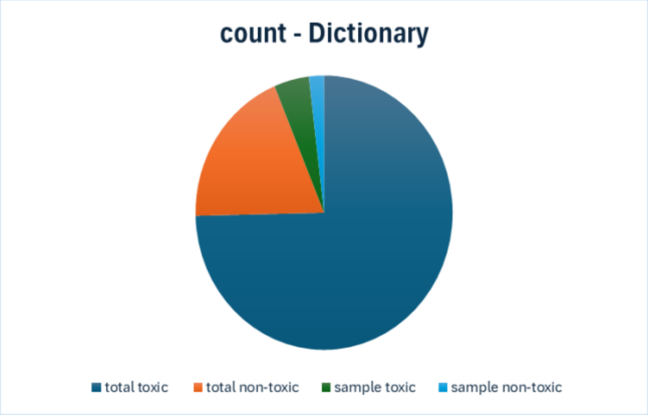


Figure 3

B. *Inter-Annotator Agreement*

Five annotators manually reviewed the comments and classified them as toxic or non-toxic. The machine learning method achieved a Krippendorf's Alpha of 0.791, indicating moderate agreement [4]. This relatively high level of agreement suggests that, despite individual biases, the majority of comments flagged as toxic by the machine learning model were perceived similarly by human annotators.

The dictionary-based method yielded a lower Krippendorf's Alpha of 0.607, indicating poor agreement [4]. We attribute this difference to reliance on keyword matching without taking context into account, resulting in misclassification of some comments that contained toxic terms but were not inherently toxic.

Method	K-Alpha	Description
Machine Learning	0.791	Moderate agreement; effective for overt toxicity, struggles with subtle cases.
Dictionary-Based	0.607	Lower agreement; struggles with context, relies on explicit keywords.

C. *Nature of Toxic Language*

Upon analyzing the toxic comments, several patterns of toxic language emerged:

1. **Racism and Xenophobia**: Many comments targeted Black individuals and other ethnic minorities using slurs, derogatory terms, and accusations of criminality or inferiority. For example, comments like *"blacks are a subhuman virus"* and *"Somali scum"* exemplify dehumanization and promote racial exclusion.
2. **Antisemitism**: Jewish individuals were frequently vilified, often through conspiracy theories (e.g., *"Jews control the media"*) and Holocaust denial. Antisemitic comments suggested Jews as social manipulators and threats to Western civilization, with some advocating for forced expulsion, underscoring deep-rooted anti-semitism.
3. **Islamophobia**: Islamophobic comments depicted Muslims as invaders or enemies (e.g., *"Islam is a disease," "Muslim horde"*), with some comments calling for violent action against Muslims or mockery of their religious practices.
4. **Misogyny**: Although less frequent, misogynistic language appeared in comments that insulted women, particularly by reducing them to their sexuality or advocating oppressive treatment (e.g., *"horrible Jewess feminist forced to wear burkahs"*).

D. *Non-Toxic Comments*

In contrast, non-toxic comments lacked inflammatory language, focusing on relative neutral topics such as current events, historical analysis, or logistical discussions. These comments, even when discussing sensitive issues, did not resort to slurs or dehumanizing rhetoric. The absence of personal attacks or derogatory language marked the clear distinction between toxic and non-toxic discourse.

E. *Targets of Toxic Language*

Toxic comments predominantly targeted specific racial and ethnic groups, such as Black individuals, immigrants, and non-White groups like Indians, Africans, Chinese, and Turks. Additionally, Jews and Muslims were frequent targets, reflecting entrenched racism, antisemitism, and Islamophobia. The comments often used dehumanizing language, framing these groups as societal threats or enemies.

F. *Differences in Toxic Languages Based on Target*

- **Ethnic/Racial Groups**: Toxic language toward racial groups involved dehumanization, slurs, and accusations of criminality or inferiority, such as calling Black people "subhuman" or describing immigrants as "invaders."
- **Religious Groups**: Antisemitic comments focused on conspiracy theories about Jewish control over global events, while Islamophobic comments framed Muslims as violent extremists.
- **Gender**: Misogynistic language often overlapped with racial or religious discrimination, particularly targeting feminists or women associated with Jewish or Muslim identities.

G. *Types of Toxic Language*

The comments can be categorized into three major types:

- **Racism/Xenophobia**: Primarily aimed at Black individuals, immigrants, and non-White groups, characterized by dehumanizing phrases such as *"subhuman"* or *"breeding"*.
- **Antisemitism**: Rooted in conspiracy theories blaming Jews for societal problems and invoking Holocaust denial.
- **Islamophobia**: Framing Muslims as violent or invasive, fostering fear and hostility.

H. *Disagreements in Toxicity and Annotator Bias*

One particular comment illustrates the complexity of detecting subtle toxicity. Despite receiving a toxicity prediction score of 0.34, the comment was flagged as inherently toxic by one annotator while the majority labeled it non-toxic. The comment in question read:

"Your parents are Greedy and selfish, they do not show any altruism for their own nation, instead they burrow themselves into another Nation, our Nation. Now you are going to grow up here, you are certainly not going to go back to your nation of your own will. You will breed over here, maybe 3/4 kids. Your kids will not even consider India as ancestral homelands, your kids will have 3/4 kids. Within 25 years you have produced 16 people, not counting the kids from your brothers and sisters, we could be looking at 64 people... Do you start to see why, we see your Parents as Greedy and Selfish? It's nothing personal, it's all of you Indians, Africans, Chinese, Turks... O.R.I.O.N. "No Surrender""

This comment exhibited several layers of toxicity:

- **Racial and Nationalistic Bias**: The comment targets entire ethnic and national groups (Indians, Africans, Chinese, Turks), labeling them as "greedy and selfish" and accusing them of undermining the native population by "burrowing" into another nation. This language promotes xenophobia and racial division by suggesting that the presence and reproduction of immigrants are inherently negative.
- **Dehumanization**: The statement reduces people from immigrant backgrounds to mere numbers and breeding units. The phrase "breed over here" is particularly dehumanizing, as it likens human

reproduction to animal behavior. The cold, mathematical calculation of descendants further strips away individual humanity, framing immigrants as a demographic threat.

- **Accusations of Exploitation**: By stating that the parents "do not show any altruism for their own nation" and accusing them of selfishness, the comment creates a false dichotomy, implying that immigrants are selfish for seeking better opportunities abroad. It perpetuates a harmful stereotype that immigrants take more than they give.
- **Collective Blame**: The comment claims it's "nothing personal" but then generalizes to "all of you Indians, Africans, Chinese, Turks," making broad, unfounded accusations about entire ethnic groups. This collective blame is a hallmark of toxic rhetoric because it unfairly targets people based on their heritage or background.
- **Militant Undertone**: The final statement, "O.R.I.O.N. 'No Surrender'", could imply allegiance to a white nationalist or supremacist ideology. This slogan adds an undertone of militancy and resistance, suggesting an "us vs. them" narrative that fuels hostility toward outsiders.

While it may not contain overtly aggressive words like slurs or direct threats, this comment is laced with xenophobic, dehumanizing, and divisive rhetoric. It promotes harmful stereotypes and generalizations about immigrants, reducing their worth to demographic statistics and positioning them as a problem for the host nation. These are classic markers of toxic language, even when expressed in a seemingly polite or "rational" manner.

This disagreement, where the majority labeled the comment as non-toxic despite clear indicators of racism and dehumanization, underscores how annotator biases or subtle toxic elements can lead to different interpretations. Such disagreements likely contributed to our K-alpha score of 0.791, just below the threshold for strong agreement (0.8) [4]. In this instance, the toxicity of the comment was understated by both the machine learning model and the majority of annotators, highlighting the challenges of capturing nuanced hate speech in large-scale data analysis.

I. *Distribution of Toxicity*

Overall, toxic language in Stormfront.org is both pervasive and diverse, targeting various groups with varying forms of hate speech. The machine learning model's predictions were generally aligned with human annotations for extreme cases of toxicity, while the dictionary-based method struggled with the contextual nuances of some comments. The toxic language found in this dataset predominantly targeted racial, ethnic, and religious minorities, frequently employing dehumanizing language, conspiracy theories, and calls for exclusion or violence.

The results of this analysis underscore the complexity of identifying toxic language in online spaces, where the line between overt hate speech and subtler forms of toxicity can be difficult to discern. These findings are further explored in the following discussion.

IV. Discussion

This study explored the identification of toxic language within Stormfront.org, an extremist forum, using both machine learning via the HateScan API and a custom dictionary-based approach. The results demonstrate that both methods captured a significant portion of the extreme hate speech, but there are notable differences in performance, particularly in terms of contextual sensitivity and annotator agreement.

A. *Effectiveness of Machine Learning vs. Dictionary-Based Methods*

The HateScan API proved effective at identifying extreme cases of toxicity, yielding a Krippendorff's Alpha of 0.791, reflecting moderate agreement among annotators. This suggests that machine learning can capture more blatant toxic comments that contain clear hate speech patterns (e.g., slurs, violent rhetoric). However, there were limitations, particularly with more subtle toxic comments. The API misclassified certain comments with more nuanced toxicity (e.g., those implying xenophobia or nationalism without explicit slurs), as reflected by the example where a comment scored 0.34 on the prediction scale despite clear dehumanization and racial bias.

The dictionary-based approach performed less consistently, with a K-alpha of 0.607, indicating less reliable annotator agreement. This discrepancy can be attributed to the reliance on keyword matching, which often fails to account for context [3]. For example, a comment containing a word from the dictionary might be flagged as toxic, even if used in a non-toxic or neutral context. Conversely, toxic comments that do not contain dictionary-listed words can be overlooked, such as those relying on coded language or more implicit forms of discrimination. This highlights the challenge of capturing toxic language solely through dictionary-based methods in environments with dynamic or evolving hate speech lexicons.

B. *Bias and Disagreement Among Annotators*

The analysis also revealed notable biases among annotators, with disagreements on the toxicity of certain comments. For example, a comment targeting immigrants for "breeding" and framing them as demographic threats received a low toxicity score from the API and a majority non-toxic vote from human annotators. This highlights the difficulty in detecting and classifying more implicit forms of toxicity, such as coded language or indirect hate speech, which may not be immediately obvious but still reinforce harmful stereotypes and social division. These types of disagreements likely contributed to the K-Alpha score of 0.791, which, while relatively high, reflects the complexity of interpreting subtle toxic language.

C. *Targets and Patterns of Toxic Language*

The nature of toxic language on Stormfront.org was deeply entrenched in racism, antisemitism, and Islamophobia, with comments targeting specific groups such as Black individuals, immigrants, Jews, and Muslims. Each group was targeted using language that varied in intensity and form. For example, racist and xenophobic comments often used dehumanizing language (e.g., comparing ethnic minorities to "vermin" or "subhuman"), while antisemitic comments focused on conspiracy theories (e.g., Jewish control of media and global events). Islamophobic comments tended to frame Muslims as invaders or violent extremists, often calling for exclusion or violence. Additionally, misogyny appeared in some comments, particularly those that reduced women to their gender or racial identity, sometimes linking them to broader conspiracy theories (e.g., feminists tied to Jewish or Muslim agendas).

D. Limitations and Weaknesses

There are a few key limitations to this analysis. Firstly, while the HateScan API captured blatant toxicity, its performance was less accurate when comments contained more implicit or coded forms of hate speech. The reliance on explicit markers of toxicity (such as slurs or extreme language) means more subtle but still harmful comments could be misclassified. Additionally, the dictionary-based method struggled to account for context, leading to some false positives (where non-toxic comments contained toxic keywords) and false negatives (where toxic comments did not contain dictionary-listed words). Lastly, annotator fatigue and bias likely influenced the manual classification process. While the sample size was manageable, fatigue from processing hateful content could lead to inconsistent classification, especially when annotators encountered less overt toxicity.

E. Future Directions

Future work could improve upon these methods by incorporating a more nuanced approach to toxicity detection. This might include updating the machine learning model to better account for context and expanding the dictionary to capture emerging hate speech terminology. Additionally, leveraging a larger and more diverse set of annotators could help reduce bias and fatigue, providing more consistent results. Finally, building models that can detect more subtle forms of toxicity (such as coded language or euphemisms) would allow for a more comprehensive understanding of toxic language in online environments.

References

[1] "Toxicity," *Jigsaw*. https://current.withgoogle.com/the-current/toxicity/

[2] Southern Poverty Law Center, "Stormfront," *Southern Poverty Law Center*, 2015. https://www.splcenter.org/fighting-hate/extremist-files/group/stormfront

[3] M. Reveilhac and D. Morselli, "Dictionary-based and machine learning classification approaches: a comparison for tonality and frame detection on Twitter data," *Political Research Exchange*, vol. 4, no. 1, Feb. 2022, doi: https://doi.org/10.1080/2474736x.2022.2029217.

[4] G. Marzi, M. Balzano, and D. Marchiori, "K-Alpha Calculator — Krippendorff's Alpha Calculator: A User-Friendly Tool for Computing Krippendorff's Alpha Inter-Rater Reliability Coefficient," *MethodsX*, vol. 12, pp. 102545–102545, Jun. 2024, doi: https://doi.org/10.1016/j.mex.2023.102545.