A Probabilistic Approach to Measuring Online User Extremism: A Case Study of a Novel Dataset from Breitbart News

Sharad Varadarajan[1[0000−0002−9686−8984]], Aaron Holm[1[0000−0001−9239−3988]], Sieu Tran[1[0000−0003−0017−4329]], Nathen Huang[2[0000−0003−3589−0501]], and Johanna Jan[1[0000−0002−3200−3631]]

[1] Accenture, 800 N. Glebe Rd., Arlington, 22209, USA
sharad.varadarajan@accenturefederal.com
aaron.holm@accenturefederal.com
sieu.tran@accenturefederal.com
johanna.jan@accenturefederal.com
[2] nathuan329@gmail.com

Abstract. The Internet has seen a dramatic increase in online toxicity commonly associated with far-right and extremist movements. In this paper, we mined a novel dataset of article information, alongside user comments from the right-wing political news site, Breitbart News Network. Toxicity levels of the article comments were assessed using an original, multi-regression-based metric called the "extremism score." Mean extremism scores among top commenters revealed (1) overlapping peaks of extreme comments over time and (2) an increase in the prevalence of extreme comments as the number of these top commenters increased. Subsequent entity recognition analysis found that specific Democratic figures were referenced more often than their Republican counterparts in articles with one or more extreme comments, while general references to Republicans increased across the same subset of articles compared to Democrats.

Keywords: measuring extremism · toxicity online · logistic regression · Breitbart data · hate speech · natural language processing

1 Introduction

An increasingly dangerous phenomenon, online toxicity has become a staple across digital spaces — including online news articles and bulletin boards — where hate speech has flourished [3]. Platforms like YouTube, 4chan/8chan, and Reddit have become popular congregating sites for the gamut of toxic users from bullies to the radicalized netizens who appeared at the 2017 Unite the Right rally in Charlottesville and the rioters at the U.S. Capitol in 2021 [5].

While online toxicity's dangers may be easily understood, anticipating how and when it arises is not nearly as simple. Measuring when toxicity mutates into extremism and subsequently becomes a public safety threat remains a challenge. As toxic behavior online evolves, researchers have increasingly applied quantitative methods to study this subject. Machine learning has already been applied to study online toxicity, though it has been primarily focused on correctly classifying comments according to various gradations of toxicity [4]. Online toxicity research has often studied gamers[2], while machine learning has often been applied to public comments datasets from social media platforms. We seek, however, to understand the behaviors of toxic and extremist users in socially relevant contexts outside of those typically studied; one such area is conversation threads for news articles. Machine learning literature is sparse in understanding the behavior of toxic online users on news forums; for this reason, our research seeks to study certain aspects of toxic user behavior in this domain, as well as the major political entities that drive toxic engagement.

Our research team scraped article comment threads from the right-wing news site, Breitbart News Network. Breitbart, which has been linked with right-wing extremist figures, is a fertile ground for online toxicity research [6]. By utilizing machine learning to assess the toxicity of comments scraped from Breitbart, we aim to infer some insights about toxic user behavior in political contexts.

2 Methodology

2.1 Novel Breitbart Dataset

Combining API requests with Selenium via Python, our research team scraped 4.3 million comments across more than 60,000 articles from 2014 to 2021. For this first iteration of analysis, we limited our scope to comments from the top 100 most engaged Breitbart commenters, defined as those with the most comments across all scraped articles. After filtering our data for comments from only the top 100 most engaged Breitbart commenters, we proceeded with approximately 187,000 comments.

2.2 Training Dataset

In order to classify toxic comments for our novel, unlabeled Breitbart dataset, we leveraged a labeled multi-class dataset[8] consisting of interactions between users on Wikipedia that community members deemed to be counterproductive [1]. The comments were categorized as belonging to one or more of 6 different classes: `toxic`, `severe_toxic`, `obscene`, `threat`, `insult`, `identity_hate`.

2.3 Toxicity Predictive Modeling

To predict the toxicity of a single comment, we trained six binary, logistic regression models, one for each type of Wikipedia comment toxicity class. We selected logistic regression for both its popularity in binary classification and its robustness when training on relatively small datasets. We also vectorized our collection of raw comments using both word and character level Term Frequency-Inverse Document Frequency features (TF-IDF features) [7]. This series of TF-IDF vectorized logistic regression classifiers achieved a 98% AUC score on a test set of Wikipedia comments.

2.4 The Extremism Score

Our paper offers a unique metric, referred to as the *extremism score*, to assess the level of extremism for each comment. This score is the maximum probability output from the six regression models. A qualitative analysis showed that comments whose extremism scores fell between 0 and 0.4 were mainly neutral while those between 0.4 and 0.6 were more provoking in nature; content with an extremism score above 0.6, however, was clearly extreme (Table 1).

Table 1. Example Breitbart comment with model-output extremism score.

Extremism Score	Example Breitbart Comment
0	“Please cite Article, Section, etc. Thanks.”
0.2	“So, all Americqns are white guys?”
0.4	“Amazing what a fool this woman is!”
0.6	“Looks like somebody figured out a way to fix stupid.”
0.8	“Stupid comment from a brain addled American.”
1.0	“STUPIDER? More Stupid, f**king brainless idiot.”

Henceforth, any comment with a score of at least 0.4 is considered extreme in this report.

3 Exploratory Analysis

3.1 Extremism by User

After calculating each comment’s extremism score, we aggregated each user’s total comments over time to determine their respective average monthly extremism scores from 2014-present. We noticed considerable fluctuation in extremism when visualizing these values over time (Fig. 1). While these fluctuations could just be one-time emotional reactions to the article content itself, current events, or existing article comments, our analysis suggests that individual users tend to become more extreme over time. Since other users become similarly more extreme together, this trend is likely not isolated. The research team speculated whether increased exposure to or engagement with potentially toxic conversation threads within Breitbart may have an effect on users.

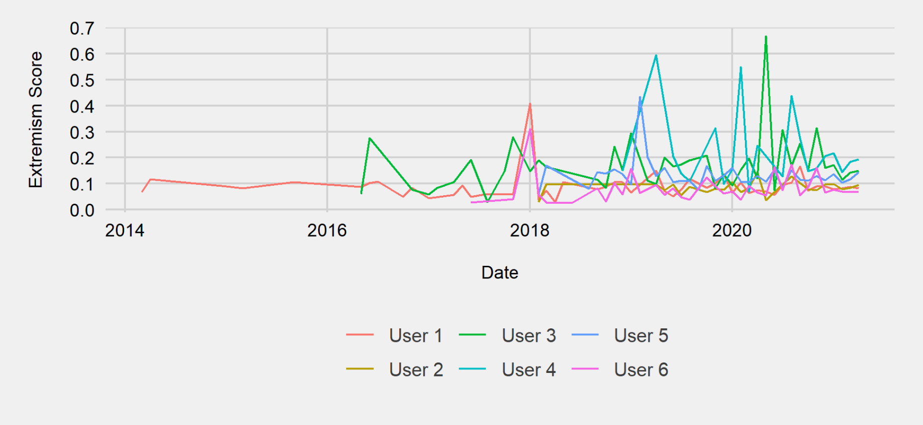

Fig. 1. Evolution of extreme commenting behavior among most engaged users.

However, the observed increased mean extremism score over time itself is insufficient for indicating that users are emotionally contaminating each other or that group dynamics are solely responsible for affecting how scores rise or fall. To study this possibility, we must first differentiate the number of users that appear under an article to see whether the article may in fact be culpable for inciting extremism.

3.2 Article Extremism with Engaged Commenters

Our team discovered that when the most engaged users commented on an article, the article's comment thread similarly exhibited higher mean extremism scores. This finding offers another perspective on peaks in extremist attitudes that supplements the previous speculation that extremist attitudes emerge from group dynamics (Fig. 2). Specifically, our finding suggests that not only can people develop more extreme views by being exposed to like-minded individuals, but also certain types of content may elicit their extremist views.

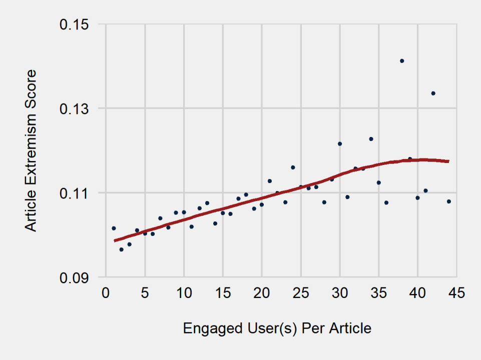

Fig. 2. Article extremism score by number of most engaged users.

After visualizing the most engaged users commenting on an article against the articles' extremism scores, we detected a potential positive correlation. Therefore, we conjecture that the topics discussed in the articles may significantly moderate the extremist behavior seen in the article's comments, particularly as such issues evolve over time to become more controversial.

3.3 Entities in Extreme Articles

The disproportionate prevalence of extreme comments across all articles suggest that specific article themes and topics may attract extreme commenting behavior. To identify some of these polarizing themes, we used entity recognition to investigate the major political entities referenced across all articles since Breitbart is an American far-right syndicated news website with political content dominating the majority of posts. To assess how often both toxic and non-toxic comments referenced polarizing political terms, we compared the percentage of references across all articles that contained contrasting entities, specifically: (1) "Biden" vs "Trump" and (2) "Democrat" vs "Republican." We calculated the raw frequency of these entities, along with commonly used inflections or synonyms, across four different sub-groups of articles: (1) All articles, (2) Articles with no extreme comments, (3) Articles with at least one extreme comment, and (4) Articles with multiple extreme comments.

Table 2. Named entity and mention frequency.

Named entity	Article group	Number of mentions
”Biden”	All articles	12,830
	Articles with no extreme comments	8,553
	Articles with at least one extreme comment	4,277
	Articles with multiple extreme comments	1,564
“Trump”	All articles	12,442
	Articles with no extreme comments	8,499
	Articles with at least one extreme comment	3,943
	Articles with multiple extreme comments	1,462
“Democrat”	All articles	7,376
	Articles with no extreme comments	5,290
	Articles with at least one extreme comment	2,086
	Articles with multiple extreme comments	701
“Republican”	All articles	6,686
	Articles with no extreme comments	4,503
	Articles with at least one extreme comment	2,183
	Articles with multiple extreme comments	852

“Biden” vs “Trump” The first pair of named entities of interest was: “Biden” and “Trump.” Across all four sub-groups, the percentage difference of references to Biden was notably higher in articles with extreme comments compared to all articles and those without any extreme comments (Fig. 3). This analysis revealed that users on Breitbart, a right-wing news site, tended to post more extreme comments on articles referencing a high-profile Democratic figure like Joe Biden compared to his Republican counterpart, Donald Trump.

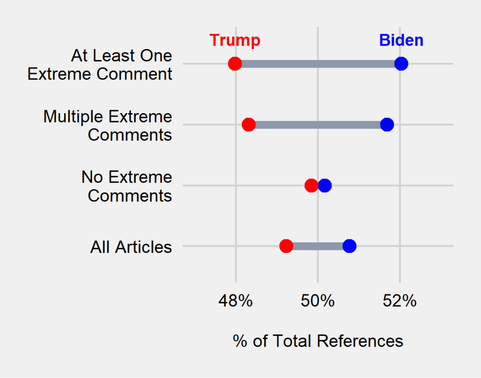

Fig. 3. Article extremism score by number of most engaged users.

“Democrats” vs “Republicans” The second pair of terms of interest was: “Democrats” and “Republicans” (Fig. 4). Unlike the named political entities, there was a higher share of references to the term “Republicans” compared to “Democrats” for articles with extreme comments. For articles with no extreme comments, however, “Democrats” was mentioned more often than “Republicans.” We previously hypothesized that mentions of “Biden” in articles coincided with extreme comments because these comments were critical of the Democratic figure in question. In an ostensible reversal, we interestingly find the opposite result with references to “Democrats” vs “Republicans” in articles. Since Breitbart is a right-wing site, we conjecture more mention of the successes or failures of the Republican party may generate tribalist attitudes in support of the party in aggregate compared to the critical tone that articles referencing specific political figures may engender.

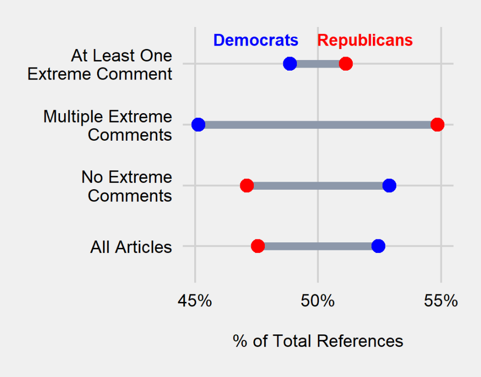

Fig. 4. Article extremism score by number of most engaged users.

4 Conclusion

To examine the role of online toxicity in political discourse, we mined a novel dataset of comments and article content published on Breitbart since 2014. Our team leveraged six binary logistic regression models trained on multi-label Wikipedia discussion comments to predict the probability that our scraped comments could belong to any of the six toxicity classes; the maximum value across these six probabilities constitutes the “extremism score.” Among the most engaged users on Breitbart, comments have on average become increasingly extreme over time, and there is a slightly positive correlation between an article’s average comment extremism score and the number of engaged users appearing in the comments. Our findings reveal that articles with a greater share of references to specific Democratic figures generate more extreme comments compared to their Republican counterparts though – by contrast – articles with a higher share of references to the term “Republicans” tend to draw more extreme comments than those with references to the term “Democrats.” Overall, these findings suggest that there may be some emotional contagion effects from both Breitbart article content and the community of users appearing in the comments section.

Though these tentative findings lack causal inference, they warrant further study of online toxicity to better understand the political implications of hardline ideological articles and commenters. Though we applied a reliable, interpretable model when classifying comments' toxicity classes, we can – through further study – try other machine learning models and benchmark their performance to better isolate extremist comments. Furthermore, after collecting the full comments dataset on all Breitbart articles published since 2014, we could train a neutral network-based supervised model (e.g., BERT-based, RoBerta-based, etc.) which may improve the predicted extremism scores. Additionally, the team is interested in applying word embeddings to the vectorization approach, as a replacement to the TF-IDF method we employed in this analysis. Finally, when determining which subjects attract extremist comments in articles, we aim to improve our current named entity recognition approach that relies on spaCy – an open source library for identifying entities – with more expansive NLP tools (like TextRazor) to identify relevant entities. Our findings encourage further research into these matters to better study, quantify, and anticipate the dangers of online extremism.

References

1. Hanu, L., Thewlis, J., Haco, S.: How AI is learning to identify toxic online content (2021), https://www.scientificamerican.com/article/can-ai-identify-toxic-online-content/
2. Kwak, H., Blackburn, J., Han, S.: Exploring cyberbullying and other toxic behavior in team competition online games. Association for Computing Machinery (2015). https://doi.org/10.1145/2702123.2702529, https://doi.org/10.1145/2702123.2702529
3. Molnar, C.: Interpretable machine learning. Lulu. com (2020)
4. Noever, D.: Machine learning suites for online toxicity detection (2018), https://arxiv.org/pdf/1810.01869.pdf
5. Papadamou, K., Zannettou, S., Blackburn, J., De Cristofaro, E., Stringhini, G., Sirivianos, M.: "How over is it?": Understanding the incel community on youtube. arXiv preprint arXiv:2001.08293 (2020)
6. Posner, S.: How Donald Trump's new campaign chief created an online haven for white nationalists (2016), motherjones.com/politics/2016/08/stephen-bannon-donald-trump-alt-right-breitbart-news/
7. Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting TF-IDF term weights as making relevance decisions. ACM Trans. Inf. Syst. **26**(3), 37 (2008). https://doi.org/10.1145/1361684.1361686, https://doi.org/10.1145/1361684.1361686
8. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: Personal attacks seen at scale. cs.CL arXiv:1610.08914 (2017)