# Measuring Internet Toxicity through Deep Learning: A Case Study of Breitbart News

**Sieu Tran,[1] Sharad Varadarajan,[1] Aaron Holm,[1]**
**Nathen Huang, Johanna Jan[1]**
[1]Accenture Federal Services
800 N Glebe Rd Suite 700,
Arlington, VA 22203
sieu.tran@accenturefederal.com
sharad.varadarajan@accenturefederal.com
aaron.holm@accenturefederal.com
nathuan329@gmail.com
johanna.jan@accenturefederal.com

## Abstract

In recent years, the Internet has seen a dramatic increase of online toxicity commonly associated with far-right political movements. Political comment threads are frequently inundated with unfiltered opinions that range in severity on a toxicity spectrum – often promulgating misinformation that promulgates political conspiracies and vaccine misinformation. In response to this challenge, our research team mined a novel dataset of article comment threads from the right-wing political news site, Breitbart News Network. We assessed toxicity levels of each article comment using a probabilistic metric called the "Internet Toxicity score (maxTOX)." Using a semi-supervised learning method, we trained a RoBERTa model on a popular Wikipedia comments dataset classified according to six toxicity labels. The model consistently revealed high maxTOX scores on Breitbart following key U.S. political events between 2020 and 2021. We empirically infer that Breitbart's news platform is a conduit for toxic online behavior and discuss the implications of our approach for measuring online toxicity.

## Introduction

Since right-wing rioters stormed the U.S. Capitol on January 6, 2021, lawmakers and citizens alike have been alarmed by social media's role in radicalizing netizens. Rioters were persuaded by digitally disseminated messages that claimed to expose corrupt political officials for wielding their influence to commit electoral fraud in 2020; these messages have since been replaced by ones amplifying misinformation about COVID-19 vaccine efforts. Once common among gaming communities, online toxicity has also become a staple across other digital spaces — particularly online news articles and bulletin boards — where hate speech has flourished (Miller, 2019). In response to the epidemic of toxic online discourse, e-commerce companies have sought to distance themselves from advertising with publishers and media platforms that fail to curb inflammatory attacks on or from within a platform. For instance, ad tech companies like OpenWeb purport resolve this problem by providing a seamless algorithmic content moderation platform that filters out toxic comments (Shoval, 2020). OpenWeb promises advertisers that their wares will only be promoted on sites where truthful, civil conversations occur-yet the tool has failed to filter out hyperpartisan remarks about election fraud, anti-vaxxer myths, and COVID-19 conspiracies (Wodinsky, 2021).

OpenWeb's shortcomings in sifting out misinformation demonstrates a key challenge with algorithmic content moderation: when do comments cross a threshold of being merely toxic to dangerously extreme? To study the evolution of online toxicity, researchers have increasingly leveraged machine learning methods. Prior machine learning research has focused on iterative improvements to classifying toxic comments with higher degrees of accuracy- using both shallow and deep learning methods (Rybinski et al., 2018, 329-343); prior research has also frequently used Wikipedia talk page edits as a common data source for baseline toxicity classification performance (Chakrabarty, 2019, 183-193). Unlike previous work, however, we seek to understand the behaviors of toxic and extremist users in socially relevant contexts outside of those typically studied: 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 and when they emerge. 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 (Posner 2016). By utilizing machine learning to assess the toxicity of comments scraped from Breitbart, we aim to infer some useful insights about toxic user behavior that have presaged key political events in the last year.

## Target Data

Combining requests with Selenium via Python, our research team scraped 4,186,509 comments across 65,211 Breitbart articles between 2014 and 2021. As a novel dataset, the Breitbart news comments were not assigned toxicity labels after scraping; thus, to assess the toxicity levels of each article comment, we utilized a semi-supervised learning approach that required first acquiring pre-classified toxic text comments for training our model.

## Training Data

We used Wikimedia's Toxicity Data Set (Wulczyn, Thain, and Dixon 2016, 2017) which contains approximately 223K annotated examples from Wikipedia Talk pages. Many recent studies have also used this dataset to perform toxic comment classification online due to its breadth and versatility (Gunasekara and Nejadgholi 2018; D'Sa, Illina, and Fohr 2020; Merayo-Alba et al. 2019). To annotate the dataset, Kaggle asked 5000 crowd-workers to rate Wikipedia comments according to their toxicity (evaluated based on how likely they were to make others leave the conversation). The multi-label dataset contains six classes: (1) toxic, (2) severe toxic, (3) obscene, (4) threat, (5) insult, and (6) identity hate.

### Label Imbalance

Table 1 shows the number of samples for each of the six classes. Of the 159,571 comments (consisting of 221,342 unique words) in the Wikipedia training data, 16,225 (10%) comments fall into one of the six categories whereas 143,346 (90%) fall into none. Of the 63,978 comments (consisting of 158,598 unique words) in the test data, 6,243 (10%) comments fall into one of the six categories while 57,735 (90%) fall into none.

| Class Label | Training Data | Test Data |
|---|---|---|
| Toxic | 15,294 | 6,090 |
| Severe Toxic | 1,595 | 367 |
| Threat | 478 | 211 |
| Insult | 7,877 | 3,427 |
| Obscene | 8,449 | 3,691 |
| Identity Hate | 1,405 | 712 |

Table 1: Data classification descriptions

The dataset is highly imbalanced so measures are needed to increase the positive class labels. Therefore, we performed several well-known data augmentation techniques — including back-translation of training data and pseudo-labeling of a sample of Breitbart data — to enhance the model's performance and transferability.

Three of the classes — obscene, toxic and insult — are highly correlated (Fig. 1) which might lead to bias in model training and evaluation. Since our goal is to find any indications of toxicity in our target data (i.e., the maximum probability of the six classes), however, this should not significantly affect our model performance.

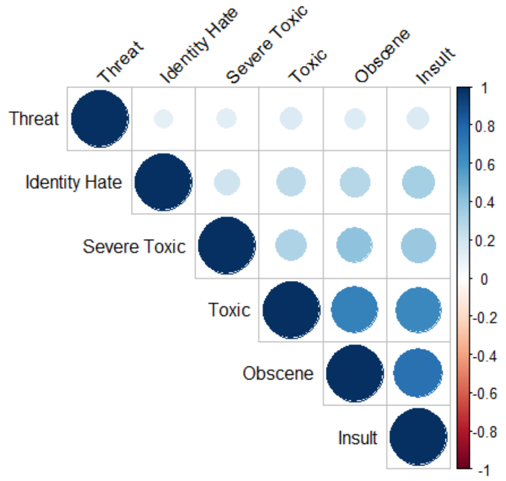

Figure 1: Correlation between the six classes of internet toxicity.

## Internet Toxicity Score (maxTOX)

While internet toxic behaviors are diverse, they mostly fall under one of the six categories introduced above: (1) toxic, (2) severe toxic, (3) insult, (4) threat, (5) obscene, and (6) identity hate. Our goal is not only to identify internet toxic behavior but also to measure the severity of such behavior in each online comment. Our paper offers a unique metric, referred to as the *maxTOX score*, to assess the level of toxic behavior exhibited in each comment. Multi-class semi-supervised learning of the aforementioned Wikipedia comment dataset will output the probability that a comment belongs to one or more of the six classes of internet toxic behavior. The maxTOX score is the maximum of those probability outputs.

## Model Building

### Transformer Architectures and Pre-trained Models

In this work, we utilized the distilled version of the RoBERTa model. Bidirectional Encoder Representation Transformer (BERT) is a transformer-based architecture that was introduced in 2018 (Devlin et al. 2018; Turc et al. 2019). BERT has had a substantial impact on the field of NLP, and achieved state of the art results on 11 NLP benchmarks at the time of its release; subsequently, RoBERTa, introduced by (Liu et al. 2019), modified various parts of BERTs training process. These modifications include more training data, more pre-training steps with bigger batches over more data, removing BERT's Next Sentence Prediction, training on longer sequences, and dynamically changing the masking pattern applied to the training data (Williams, Rodrigues, and Novak 2020). In this research, we used DistilRoBERTa-base (Sanh et al. 2019), which has 6 layers, 768 dimensions, and 12 heads, totaling 82 million parameters (compared to 125 million parameters for RoBERTa-base). On average DistilRoBERTa is twice as fast as RoBERTa-base, and the

English DistilRoBERTa model contains 50,265 WordPieces.

For RoBERTa, inspired by success in Williams, Rodrigues, and Tran (2021), we added an additional mean-pooling and dropout layers prior to the final classification layer. Adding these additional layers has been shown to help prevent over-fitting while fine-tuning (Williams, Rodrigues, and Novak 2020). Finally, to generate predictions, our team applied a Softmax layer with 6 output nodes to the model's outputs, one for each class of toxicity. The difference between the positive and negative class likelihoods were then used to score comments. We utilized an Adam optimizer with a learning rate of 1.5e-5 and an epsilon of 1e-8; then, we trained the model for 2 epochs — each with a batch size of 32 — and sought to minimize binary cross-entropy loss.

## WordPiece Analysis

Transformer models utilize WordPiece tokenization schemes that are dependant on the model being evaluated. During pre-training, the WordPiece algorithm determines which pieces of words will be retained and which will be discarded. An **UNK token** is utilized as a placeholder in the lexicon and used to represent WordPiece tokens received in novel input that did not get utilized at model creation. A large amount of tokens processed as UNK may suggest potential poor performance. The training set contains 29,876 unique WordPieces, while the test set contains 27,279 unique WordPieces based on the DistilRoBERTa tokenizer. Roughly 260 UNK tokens were found in the training data and 56 in testing data, suggesting DistilRoBERTa's performance will not be greatly hindered.

## Data Augmentation

Due to the imbalanced nature of our training set, data augmentation is critical. Back-translation is an effective technique for enhancing the performance of a model with limited and imbalanced training data (Feldman and Coto-Solano 2020; Xie et al. 2021), while pseudo-labeling of target data has demonstrated promising improvement in a model's transferability (Dopierre et al. 2020; Cascante-Bonilla et al. 2020).

With the roughly 20,000 comments falling into at least one of the six internet toxic classes from the Wikimedia Toxicity Data Set, we performed back-translation of the English comments with three target languages: French, German, and Spanish. After data de-duplication, we acquired roughly 58,536 more positive samples for the corpus.

Then, after extensive hyperparameter tuning, we used our best performing models to predict toxicity class labels on a random sample of roughly 187,000 comments from the Breitbart data set. Through this process, we acquired 116,737 pseudo-labeled comments (roughly 5,969 positive class labels) from the sample to add to our training data: only comments that achieved a 0.9 probability in at least one of the six toxic behavior classes or below a 0.1 probability in each class — deemed confident predictions from our model(s) — were labeled and added to our training data. The team incorporated this semi-supervised learning approach into our final model due to the aforementioned improvements in model transferability.

| Class Label | Backtranslation | Pseudo-labeling |
|---|---|---|
| Toxic | 55,632 | 5,969 |
| Severe Toxic | 5,051 | 1 |
| Threat | 1,799 | 26 |
| Insult | 29,413 | 4,427 |
| Obscene | 31,501 | 2,719 |
| Identity Hate | 5,496 | 348 |

Table 2: Data augmentation count per class

## Model Evaluation

Since our primary goal is to acquire the maximum predicted probability of the six classes, true performance of the model and the effectiveness of the maxTOX score may naturally be more optimistic than individual class performance. However, to benchmark DistilRoBERTa's performance on this task, we compared its performance with an established Bidirectional Long Short Term Memory (biLSTM) model framework. Though word embeddings are a semantic representation of words, bidirectional neural networks are nevertheless known for generating robust semantic representations for a given sequence of words. Previous research points to BiLSTM architectures performing well in Multi-label Toxicity Identification of online content (Gunasekara and Nejadgholi 2018; D'Sa, Illina, and Fohr 2020; Merayo-Alba et al. 2019).

| Model | Classification | Acc. | Prec. | Rec. |
|---|---|---|---|---|
| DistilRoBERTa | Toxic | 0.9377 | 0.6131 | 0.9362 |
| | Severe Toxic | 0.9933 | 0.4410 | 0.6213 |
| | Threat | 0.9985 | 0.7706 | 0.7962 |
| | Insult | 0.9757 | 0.7412 | 0.8409 |
| | Obscene | 0.9731 | 0.7167 | 0.8826 |
| | Identity Hate | 0.9950 | 0.7655 | 0.7978 |
| biLSTM | Toxic | 0.9073 | 0.5078 | 0.8608 |
| | Severe Toxic | 0.9925 | 0.3544 | 0.3815 |
| | Threat | 0.9967 | 0.5130 | 0.2796 |
| | Insult | 0.9485 | 0.5137 | 0.7406 |
| | Obscene | 0.9468 | 0.5268 | 0.7719 |
| | Identity Hate | 0.9910 | 0.6287 | 0.4565 |

Table 3: Model performance on the positive class

Equipped with back-translation and pseudo-labeling, our DistilRoBERTa model considerably outperformed our biLSTM model in both precision and recall. While the accuracy is high across all classes for both models, the label imbalance (even after data augmentation) led the team to pay more attention to other performance metrics. Specifically, the team believes recall is one of the most important metrics in toxic comment detection, as false negatives can have far more negative implications compared to false positives. While it was both reassuring and informative to measure model performance on the Wiki data test set, the team was more concerned with the model's transferability and how it performed on the Breitbart dataset.

Since Breitbart News is a different subject domain, and our data was previously unlabeled, we performed a qualitative analysis with 1,000 randomly sampled comments where the two models differ in hard label maxTOX predictions; 500 comments with a positive class for DistilRoBERTa and negative class for biLSTM, and 500 comments with a positive class for biLSTM and negative class for DistilRoBERTa. Three data-annotators separately assessed all 1,000 comments and documented which model they agreed with for each comment. On average, the three data-annotators agreed with each other 53.7% of the time and with DistilRoBERTa's classification scheme 58.78%. Based on DistilRoBERTa outperforming biLSTM on the Wikipedia test data and our results from a qualitative analysis on a random sample from the Breitbart dataset, we are optimistic about its overall performance.

## Model Application

Equipped with toxicity classifications for our target dataset, we collected descriptive statistics on toxic comments across Breitbart and studied the following questions for each Breitbart article's comment threads:

- Are Breibart articles culpable for inciting the toxicity of the comments?
- Are comments consistently toxic, or are there notable spikes in toxicity near major polarizing events?

Our model predicted that 550,732 of the 4,186,509 scraped comments have a maxTOX score greater than 0.5, and thus fall into at least one of the six internet toxicity classes. Table 4 shows a breakdown of the number of comments classified to each class.

| Class Label | Number of Comments | Percentage |
|---|---|---|
| Toxic | 545,725 | 13.0% |
| Severe Toxic | 49 | 1.2e-3% |
| Threat | 3,305 | 7.9e-2% |
| Insult | 245,222 | 5.9% |
| Obscene | 105,026 | 2.5% |
| Identity Hate | 33,672 | 0.8% |

Table 4: Breitbart comment classification

On average, the model deemed roughly 15% of the comments on each Breitbart article to be toxic based on their maxTOX scores. Figure 2 shows a considerable right skew in toxic comments per article; 17,428 articles have more than 15% internet toxic comments while 1,001 articles have over 90% internet toxic comments.

We identified a few potentially polarizing political events and investigated the evolution of average maxTOX scores on Breitbart for daily comments near such events; for this specific case study, we analyzed comments made around the death of George Floyd (Fig. 3), the announcement of Kamala Harris as Vice Presidential nominee (Fig. 4), the U.S. 2020 Presidential Election (Fig. 5), and the U.S. Capitol Riots (Fig. 6).

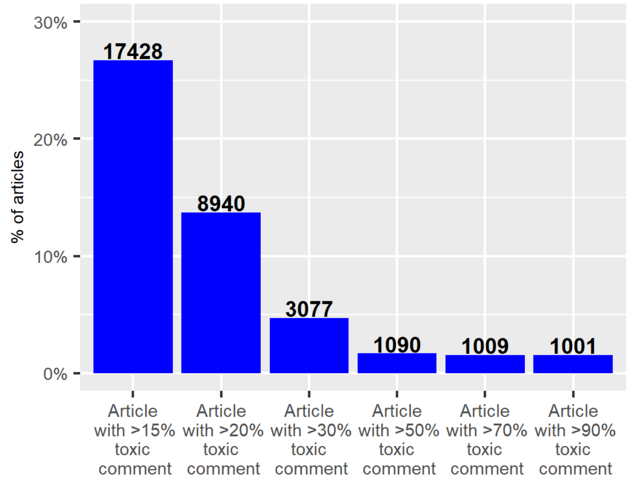

Figure 2: Number of article that contain at least a certain amount of toxic comments.

### George Floyd's Death

On May 25, 2020, a 46-year-old Black man, George Floyd, was murdered by a police officer in Minneapolis, Minnesota during an arrest on suspicion of using a counterfeit $20 bill. From Figure 3, we noticed a considerable increase in the average maxTOX score of comments on Breitbart in the days following his death until a local maximum on June 2nd, 2020.

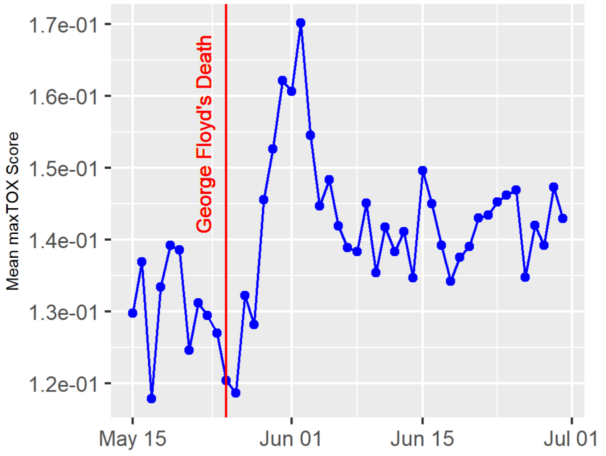

Figure 3: Daily average MaxTOX score - George Floyd's Death

Our model deemed 16.6% of comments made on June 2nd, 2020 to be toxic. Table 5 lists the five articles that generated the most toxic comments on June 2nd, 2020; all five articles have mentions of George Floyd's murder or the nationwide protests that followed. These five articles — and their corresponding headlines — feature public figures being critical of Donald Trump.

| Article Headline | Count |
|---|---|
| Sen. Ed Markey Calls Donald Trump 'Scum' for 'Fueling' Violence | 33 |
| CNN's Jim Acosta Shouts at Trump: 'Is This Still a Democracy?' | 31 |
| Pink: Trump Can Watch His Election Defeat From 'His Baby Bunker' | 31 |
| Kamala Harris: Trump 'Has Combined the Worst of George Wallace with Richard Nixon' | 28 |
| George Clooney: America Has a Racism 'Pandemic' That 'Infects All of Us' | 28 |

Table 5: Articles that generated most toxic comments on June 2nd, 2020

Our analysis of the toxic comments revealed vehement condemnation of the high-profile figures mentioned in the aforementioned article headlines, with strong support for President Trump amidst the protests following George Floyd's death. Some representative comments from this sample include:

> "Wait til Geo Clooney hears about Geo Floyd being intoxicated with fentanyl and meth on top of his health conditions."
>
> "Nothing says plastic banana like celebutards pretending to give 2 shits about Mr. Floyd."
>
> "Irony that they have concern about Mr. Floyd but kill black babies by the millions."
>
> "and Pink can watch President Trump win re-election like the stupid ugly manly looking hoebag she's always been."

## Announcement of Kamala Harris as Vice Presidential Nominee

On August 11th, 2020, Joe Biden officially selected California senator Kamala Harris as his running mate for president. From Figure 4, we observed in the days following Kamala's VP nomination a gradual increase in the average maxTOX score of comments on Breitbart up until a local maximum on August 14th, 2020.

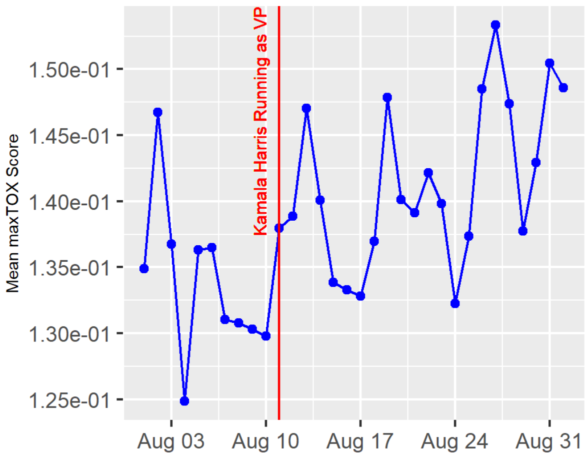

Figure 4: Daily average MaxTOX score - Kamala Harris VP Nomination

This time, our model deemed 13.5% of comments made on August 14th, 2020 to be toxic. Table 6 lists the five articles that generated the most toxic comments on August 14th, 2020. While none of the five articles are focused on Harris's nomination itself, we observed some similarities to the headlines of the most toxic comment-attracting articles following George Floyd's death- specifically the mentions of high-profile Democrats or celebrities expressing a critical view towards Donald Trump or one of his policies.

| Article Headline | Count |
|---|---|
| Lance Armstrong's Bike Shop Cancels Police Contract, Still Expect Cops to Protect from Threats | 29 |
| Rashida Tlaib 'Won't Celebrate' Israel-UAE Deal: No Credit to Bibi 'for Not Stealing Land' | 29 |
| Pelosi on Coronavirus Relief: 'Everything I Do Is About the Children' — 'I Have Advice for Them Whether They Want It or Not' | 29 |
| Schiff: 'No Racist Appeal Too Much, No Political Dirty Trick Beyond the Pale' for Trump | 26 |
| Susan Rice: Trump 'Sends Troops into the Streets of Our Cities to Attack Peaceful Protesters' | 24 |

Table 6: Articles that generated most toxic comments on August 14th 2020

Once again, our team found that the toxic comments contained obscene insults aimed at the high-profile liberal figures mentioned in the headlines. Though the five articles that generated the most toxic comments on August 14th, 2020, did not call out Kamala Harris, our model identified multiple toxic comments regarding her and her nomination were identified. Approximately 3% of the roughly 1700 toxic comments on August 14th 2020 explicitly mention Kamala Harris, including the following remarks:

“She should be vp with biden ..she clueless like old corn pop..or kamalier”

“The whole basis of reparations is blaming descendants of slaveowners for the sins of their ancestors. So, Kamala doesn’t get a pass for her slaveowning ancestors. If she’s not dirtied by her family history, then why are they tearing down Washington’s statues and demanding reparations? Hoist with the leftists own petard.”

“Kamala is more white that black. Her grandfather was one of the biggest slave owner in Jamaica and is of Irish descent.She is a phoney and even her father has no use for her.”

“A black woman in a position of power is a VERY bad thing.”

“Kamala BLOWJOB Harris....”

## U.S. 2020 Presidential Election

The 2020 presidential election took place November 3rd, 2020. From Figure 5, the average maxTOX score of comments on Breitbart reach a local maximum on November 7th, 2021. This finding is to be expected as multiple news outlets began to project Biden’s win on the morning of the 7th.

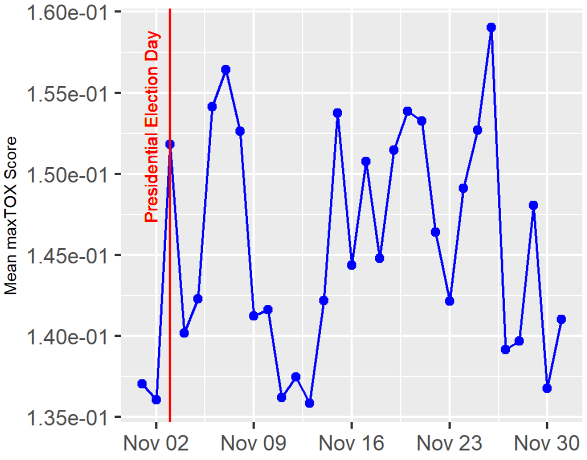

Figure 5: Daily comment MaxTOX score near the 2020 Presidential Election.

In this instance, the DistilRoBERTa model deemed 15.3% of comments made on November 7th, 2020 to be toxic. Table 7 lists the five articles that generated the most toxic comments on November 7th, 2020; all five articles pertain to Donald Trump and the election and have some mention of the 2020 election results. The research team identified various subtopics in these articles as potentially inciting toxic comments, including accusations of election fraud and relaying misinformation from Donald Trump. Headlines in Table 7 also highlighted celebrities and other high-profile figures expressing a critical view towards Donald Trump.

| Article Headline | Count |
|---|---|
| Actor Jon Cryer Explains to 69 Million Trump Voters How Trump Betrayed Them | 27 |
| Christie Rounds on Trump for Election Fraud Claims that ‘Inflame Without Informing’ | 25 |
| Facebook ‘Supreme Court’ Member Tawakkol Karman Says Trump Fed ‘Wave of Hate and Intolerance’ | 23 |
| Facebook Censorship: Platform Will ‘Temporarily Demote’ Posts that Share ‘Election Misinformation’ | 22 |
| Actress Natalie Morales: Cuban Americans Who Voted for Trump ‘Are 10,000 Percent Brainwashed’ | 21 |

Table 7: Articles that generated most toxic comments on November 7th, 2020

In addition to malevolent comments directed at high-profile figures referenced in article headlines, comments explicitly mentioned Joe Biden in at least 6.5% of toxic comments, while they referenced Donald Trump in at least 6.0% of toxic comments. Common themes among comments that explicitly mention Biden and/or Trump included their mental states, voter fraud, corruption, and critique of their respective supporters. Representative comments are listed below:

“To all you spineless communist, be patient, Trump isn’t done with you yet. Once the recounting is done you will be very disappointed. ”intellectualism” = moron obedient bums. No freedom, no voice, no life. STFU.”

“Alzheimers Biden and Kameltoe Harris will never be presidents.”

“Not from voting. Covid scare was the excuse to break the system and commit massive voter fraud. P.S. A virus is stupid just like you, they can’t tell a biden voter from a Trump voter.”

“Trump ”is the American people, you ignorant communist asshast..GFYS”

## U.S. Capitol Riots

On January 6th, 2021, right-wing rioters stormed the US Capitol to protest the 2020 presidential election results. From Figure 6, we noticed an upward trend in the average maxTOX score of comments made on January 7th 2021- achieving a local maximum in average comment toxicity the day after the Capitol riots.

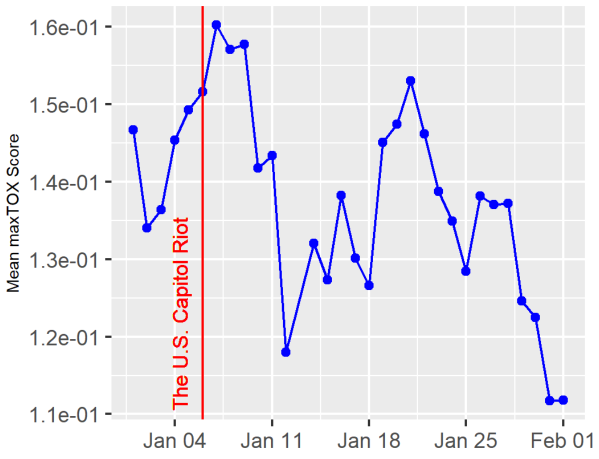

Figure 6: Daily average MaxTOX score - U.S. Capitol Riot

Finally, our model determined 15.7% of comments made on January 7th 2021 to be toxic. Table 8 lists the five articles that generated the most toxic comments on January 7th, 2021; four of the five articles center around the Capitol Riots with headlines highlight the political dissonance both within and across party lines.

| Article Headline | Count |
|---|---|
| Robert 'Beto' O'Rourke to Ted Cruz: 'Your Self-Serving Attempt at Sedition' Inspired 'Attempted Coup' | 25 |
| Hollywood Celebs Gush over Stacey Abrams Following Georgia Runoffs: Put Her On the $20 Bill | 24 |
| GOP Rep. Kinzinger: Trump 'Incited This Coup,' 'Did Little to Protect the Capitol' | 23 |
| Republican Vermont Governor Demands President Donald Trump's Resignation | 21 |
| GOP Rep. Pulled Wooden Leg from Table to Defend Self at Capitol | 21 |

Table 8: Articles that generated most toxic comments on January 7th 2021

We reviewed the toxic comments from January 7th 2021 and found various subtopics of interest regarding the Capitol Riots; multiple comments from right-wing supporters directed blame towards Antifa, Black Lives Matter, Mike Pence, and liberals for the protests, while comments from left-wing supporters accused Trump. Some representative comments from this sample include:

"They shot an unarmed Karen you moron. They've let BLM get away with everything for 9 months. Or don't you watch the news?"

"Big defeat for you. You're on notice right now,idiot Leftist. Don't cheat at elections and cause any more such protests. They could be a lot more dangerous – for you and other Leftist no-nothings."

"Democrat backed Antifa and BLM terrorists attacked the police not MAGA. But you are too stupid too even check out the guys leading the way. All BLM and Antifa"

"I'll never accept Biden as president. To me he is just a sick deranged man who has nothing of interest to say to anyone."

"Pence is a useless coward moron. Chamberlain anyone?"

"Violence and Mob Rule Is Wrong and Un-American ??? unless your a Damnarat"

## 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 a distilRoBERTa model, and performed a robust semi-supervised learning process to predict the probability that our scraped comments could belong to any of six toxicity classes; the maximum value across these six probabilities constitutes the "maxTOX score."

Both qualitative and quantitive analyses of toxic comments from Breitbart's comment threads reveals a consistent pattern of toxicity on its platform near major polarizing political events from 2020-2021. Specifically, we noticed local maxima in average maxTOX scores in the days following these major events. Additionally, we observed that the articles that incited the most toxic comments after each major polarizing event discussed in this paper involved a high-profile figure (usually liberal and/or a celebrity) sharing a critical opinion of Donald Trump or one of his policies.

Our analysis suggests that readers appear to use the platform to air out political grievances, and Breitbart may be inciting these behaviors and catering to its audiences through partisan articles and emotionally charged headlines. 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. We did have some limitations in terms of computational power. Typically data augmentation would be done until class balance is archived but with limited computational power, we did not achieve full class balance. Furthermore, after collecting the full comments dataset on all Breitbart articles published since 2014, the team is interested in a more in-depth analysis of which subjects attract extremist comments in articles. Our findings encourage further research into these matters to better study, quantify, and anticipate the dangers of online toxic behaviors.

## References

Cascante-Bonilla, P.; Tan, F.; Qi, Y.; and Ordonez, V. 2020. Curriculum labeling: Self-paced pseudo-labeling for semi-supervised learning. *arXiv preprint arXiv:2001.06001* 8.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Dopierre, T.; Gravier, C.; Subercaze, J.; and Logerais, W. 2020. Few-shot pseudo-labeling for intent detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4993–5003.

D'Sa, A. G.; Illina, I.; and Fohr, D. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN.

Feldman, I.; and Coto-Solano, R. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, 3965–3976.

Gunasekara, I.; and Nejadgholi, I. 2018. A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, 21–25.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692. URL http://arxiv.org/abs/1907.11692.

Merayo-Alba, S.; Fidalgo, E.; González-Castro, V.; Alaiz-Rodríguez, R.; and Velasco-Mata, J. 2019. Use of Natural Language Processing to Identify Inappropriate Content in Text. In Pérez García, H.; Sánchez González, L.; Castejón Limas, M.; Quintián Pardo, H.; and Corchado Rodríguez, E., eds., *Hybrid Artificial Intelligent Systems*, 254–263. Cham: Springer International Publishing. ISBN 978-3-030-29859-3.

Posner, S. 2016. How Donald Trump's new campaign chief created an online haven for white nationalists. URL motherjones.com/politics/2016/08/stephen-bannon-donald-trump-alt-right-breitbart-news/.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv* abs/1910.01108.

Turc, I.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models.

Williams, E.; Rodrigues, P.; and Tran, S. 2021. Accenture at CheckThat! 2021: Interesting claim identification and ranking with contextually sensitive lexical training data augmentation.

Williams, E. M.; Rodrigues, P.; and Novak, V. 2020. Accenture at CheckThat! 2020: If you say so: Post-hoc fact-checking of Claims using Transformer-based Models. In Cappellato, L.; Eickhoff, C.; Ferro, N.; and Névéol, A., eds., *Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020*, volume 2696 of *CEUR Workshop Proceedings*. CEUR-WS.org. URL http://ceur-ws.org/Vol-2696/paper_226.pdf.

Wulczyn, E.; Thain, N.; and Dixon, L. 2016. Wikipedia Talk Labels: Personal Attacks. figshare. doi:https://doi.org/10.6084/m9.figshare.4054689.v6.

Wulczyn, E.; Thain, N.; and Dixon, L. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, 1391–1399.

Xie, Y.; Xing, L.; Peng, W.; and Hu, Y. 2021. IIE-NLP-Eyas at SemEval-2021 Task 4: Enhancing PLM for ReCAM with Special Tokens, Re-Ranking, Siamese Encoders and Back Translation. *arXiv preprint arXiv:2102.12777* .