ExtremeBB: A Database for Large-Scale Research into Online Hate, Harassment, the Manosphere and Extremism

Anh V. Vu†, Lydia Wilson†, Yi Ting Chua⋆, Ilia Shumailov⋄, Ross Anderson†‡
†University of Cambridge, ⋆University of Alabama
⋄University of Oxford, ‡University of Edinburgh

Abstract

We introduce ExtremeBB, a textual database of over 53.5M posts made by 338.5k users on 12 extremist bulletin board forums promoting online hate, harassment, the manosphere and other forms of extremism. It enables large-scale analyses of qualitative and quantitative historical trends going back two decades: measuring hate speech and toxicity; tracing the evolution of different strands of extremist ideology; tracking the relationships between online subcultures, extremist behaviours, and real-world violence; and monitoring extremist communities in near real time. This can shed light not only on the spread of problematic ideologies but also the effectiveness of interventions. ExtremeBB comes with a robust ethical data-sharing regime that allows us to share data with academics worldwide. Since 2020, access has been granted to 49 licensees in 16 research groups from 12 institutions.

1 Online Extremist Forums

Online hate, harassment, the manosphere and extremism are growing and pernicious problems, associated with real-world threats and violent crime (Aghazadeh et al., 2018; Lally and Bermingham, 2020; Vu et al., 2024). Large-scale quantitative studies of these emergent problems and the links between them have been limited by the difficulties of collecting high-quality data, and of sharing these data ethically with other academic researchers. The recent withdrawal of free public access to Twitter is a real challenge (Twitter Developer, 2023). Extremist material is concentrated in a variety of other channels and places online, particularly on Reddit and in forums.

While Reddit data has been collected and shared widely (Baumgartner et al., 2020), the research community has lacked a comprehensive collection of other online forum data at scale. Forums are structured into themed boards, each a collection of threads with one or more posts. Posts have not just textual content but valuable metadata such as posting time, authors, their joining date and reputation. We collect online forum data and make it available to researchers. Our collection methodology is dynamic. We first compiled a list of known online communities using in-house and external experts, and shortlisted publicly accessible active extremist forums with at least 50k posts. As some forums are agile in the sense that they are short-lived or their domains rotate, a feedback loop was created with experts monitoring the list and alerting collection engineers of any changes. We also actively monitor extremist Telegram and Discord channels to spot new forums. At present, most ExtremeBB forums originate from the USA, grouped into the following six categories based on the nature of their content and their self-definition.

WHITE SUPREMACY. This contains the two largest and longest-running race-hate, antisemitic and white supremacist forums. Some forum users were reportedly involved in serious crime, and these forums are associated with terrorist atrocities such as the Wisconsin Sikh temple shooting, the Pittsburgh police officer shooting, and the Norway attacks. Our collection is significantly larger than those reported in previous work (Holt et al., 2020; Kleinberg et al., 2021).

INCELDOM. This is an extreme misogynistic ideology whose dominant theme is its members’ inability to have sexual relationships with women. Forum users envy attractive men and blame feminism and female nature for their plight, indulging in violent fantasies and hate speech. The two major forums we collect are larger than datasets of incel material used in existing work (Baele et al., 2019; Jaki et al., 2019; O’Donnell and Shor, 2022).

LOOKISM. This consists of the two largest forums on techniques to enhance male physical beauty, ranging from non-invasive methods such as styling and workouts to extreme ones involving cosmetic surgery, synthetic hormones, and skin-lightening products. This is strongly linked to the incel subcultures, with users often entangled with misogyny involving a stereotype of female superficiality that requires men to invest in a masculine appearance.

PICKUP ARTISTRY. This includes two popular men’s online forums offering instructions, advice and
online harassment forums associated with real-world attacks on transgender, feminist, autistic and disabled individuals, with members engaging in organised trolling, cyberstalking, doxxing and swatting of their targets, some of whom have committed suicide after being harassed (Ambreen, 2019).

## 2 The ExtremeBB Database

ExtremeBB has over 53.5M posts (Table 1); the post length distribution is shown in Figure 1. Just as with cybercrime forums (Vu et al., 2020), increasing activity was seen on extremist forums during the COVID-19 pandemic, with posting in 2020 being double that of 2019 (but decreasing 20% in 2021), presumably as users had more time online during lockdowns (Figure 2). Thousands of users join multiple forums,\(^1\) suggesting considerable overlaps between different types of online extremism (see Table 2). These forums are considered extreme, but Google’s Perspective API rates a large proportion of individual posts as non-toxic, when measured in January 2023 (Figure 3).\(^2\) The main themes of each category are shown in Figure 4, with stop words excluded and terms lemmatised then obfuscated (Nozza and Hovy, 2023).

### Data Collection

ExtremeBB is maintained and expanded continuously, and in real time when nec-

\(^1\) A lower-bound overlap is estimated based on identical rare usernames and correlated posting patterns. Cross-forum analysis may be limited due to the lack of ground-truth insight.

\(^2\) This is the state-of-the-art tool for toxicity analysis (Zanettou et al., 2020), but determining posts’ toxicity may heavily depend on context. The tool only looks at the first 501 characters of each post (Talas, 2023), and is limited in dealing with both adversarial samples (Hosseini et al., 2017) and racial bias (Sap et al., 2019).
Table 2: User overlaps across forums. $n(i, j)$ depicts the number of users on forum $i$ who also register on forum $j$.

<table>
<thead>
<tr>
<th></th>
<th>ST</th>
<th>VA</th>
<th>II</th>
<th>IN</th>
<th>LK</th>
<th>LS</th>
<th>RV</th>
<th>PA</th>
<th>MG</th>
<th>GY</th>
<th>KF</th>
<th>LC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>945</td>
<td>403</td>
<td>154</td>
<td>673</td>
<td>412</td>
<td>906</td>
<td>2346</td>
<td>273</td>
<td>149</td>
<td>1624</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>VA</td>
<td>945</td>
<td>56</td>
<td>205</td>
<td>549</td>
<td>769</td>
<td>144</td>
<td>297</td>
<td>44</td>
<td>31</td>
<td>365</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>403</td>
<td>56</td>
<td>205</td>
<td>127</td>
<td>116</td>
<td>45</td>
<td>119</td>
<td>27</td>
<td>18</td>
<td>139</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>IN</td>
<td>154</td>
<td>27</td>
<td>205</td>
<td>127</td>
<td>116</td>
<td>45</td>
<td>119</td>
<td>27</td>
<td>18</td>
<td>139</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>LK</td>
<td>673</td>
<td>86</td>
<td>549</td>
<td>127</td>
<td>993</td>
<td>268</td>
<td>596</td>
<td>79</td>
<td>52</td>
<td>546</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>412</td>
<td>48</td>
<td>769</td>
<td>116</td>
<td>993</td>
<td>155</td>
<td>360</td>
<td>62</td>
<td>23</td>
<td>299</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>RV</td>
<td>906</td>
<td>109</td>
<td>144</td>
<td>45</td>
<td>268</td>
<td>155</td>
<td>815</td>
<td>117</td>
<td>79</td>
<td>477</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>PA</td>
<td>2346</td>
<td>203</td>
<td>297</td>
<td>119</td>
<td>596</td>
<td>360</td>
<td>815</td>
<td>212</td>
<td>114</td>
<td>1113</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>MG</td>
<td>273</td>
<td>35</td>
<td>44</td>
<td>27</td>
<td>79</td>
<td>62</td>
<td>117</td>
<td>212</td>
<td>120</td>
<td>158</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GY</td>
<td>149</td>
<td>20</td>
<td>31</td>
<td>18</td>
<td>52</td>
<td>23</td>
<td>79</td>
<td>114</td>
<td>120</td>
<td>84</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>KF</td>
<td>1624</td>
<td>166</td>
<td>365</td>
<td>139</td>
<td>546</td>
<td>299</td>
<td>477</td>
<td>1113</td>
<td>138</td>
<td>84</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>LC</td>
<td>14</td>
<td>0</td>
<td>4</td>
<td>2</td>
<td>8</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>1</td>
<td>97</td>
<td>–</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: The toxicity level of posts. The purple areas: probability density; TX: toxicity, ST: severe toxicity, IA: identity attack, IS: insult, PF: profanity, TH: threat. Red dots are means. A large proportion of posts are not toxic.

Figure 4: Word clouds of the top 100 keywords on White Supremacy, Inceldom, Lookism, Pickup Artistry, Men’s Movement, and Online Harassment forums, respectively. The word sizes correspond to their popularity.

essay. Our web scraper, based on CrimeBot (Pastreña et al., 2018), can bypass various types of CAPTCHAs and other bot prevention techniques. It is designed to ensure data completeness by visiting each board regularly while persistently storing its progress for incremental crawling when disruptions occur. A new crawl starts when one finishes: some sites complete within hours, while some take days or weeks. Vanished valuable forums such as Lookism and LoLcow have been fully captured.

Data Licensing. Our regime was carefully crafted in consultation with legal academics, university lawyers and specialist external counsel, allowing the data to be shared in multiple jurisdictions. ExtremeBB aims to allow both computer and social scientists – especially those who may lack time or technical skills for data collection – to focus on their research instead of spending months or even years gathering the data. Although the forums are public, some content may be personally identifiable, or even illegal. Licensees are therefore required to sign a formal agreement with the Cambridge Cybercrime Centre to prevent misuse. ExtremeBB is provided as PostgreSQL snapshots; once registered, licensees will automatically get access to the most recent data as they are published. Shared data can only be used for research. Data leakage is a contract breach regardless of whether directly exposing the data or carelessly publishing a model from which the data could be extracted.

Limitations. Our database is unlikely to capture all available extremist forums, but we believe it includes the most prominent ones. Our ethical safeguards do not allow automatic downloads of images because of the risk of potentially illegal material, with the result that 5.96% of posts are empty. The
database thus includes text only, but media URLs are kept for further analyses if needed. Some forums have fewer posts than the statistics shown on their homepage as (1) our collection started in 2020 thus older content might have been removed or hidden, (2) content posted then removed quickly between crawls may be missed, and (3) access to some sub-forums may be restricted. ExtremeBB only consists of English forums, but we are also starting to expand it to other languages.

**Our Vision.** We aim to offer a high-quality, fine-grained dataset to enable researchers to better understand the behaviours and interactions of extremists within and across different subcultures, and to monitor incidents affecting them such as law-enforcement interventions, the pandemic (Vu, 2020) and platform disruption (Vu et al., 2024). Our data are mostly unlabelled and can best be used for statistical or unsupervised tasks, but we have started labelling hate speech for language-model training and evaluation. We have built an interactive web-based tool to assist less technical scientists to explore our cybercrime data (Pete et al., 2022); this will shortly be extended to ExtremeBB.

### 3 Ethics Statement

Our team has robust ethical procedures to deal with scraped data that may contain sensitive information. The collection of ExtremeBB and the license agreement were approved by our institutional ethics committee. We only collect publicly accessible posts, which solves many but not all problems of privacy (TechCrunch, 2022). Our scraper does not flood targeted servers with unnecessary traffic; we do our best to collect only what is needed. The scraper operates slowly and stores raw HTML files to avoid repetitive crawls in case offline parsing is required. The parsed data are kept encrypted to avoid data leakage. Carrying out research on sensitive and abusive data may introduce a wide range of risks for researchers (Doerfler et al., 2021), so we follow guidance to protect researchers from online harassment (Marwick et al., 2016). One licensing condition is that every project must be approved by the licensee’s institutional ethics committee or review board; another is that research results must be placed in the public domain, ensuring transparency. We only analyse aggregated behaviours of groups instead of studying individuals, which follows the British Society of Criminology Statement on Ethics (British Society of Criminology, 2015).

**Acknowledgments**

We are grateful to Alice Hutchings, Ben Collier, and Sergio Pastrana for their insightful feedback, as well as Richard Clayton and Tina Marjanov for maintaining the data scrapers’ infrastructure. This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/V026178/1] and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme [grant agreement No 949127].

**References**


Sam Ambreen. 2019. *Kiwi Farms Linked to At Least 2 Murders and 4 Suicides. Left at the Lights*.


Periwinkle Doerfler, Andrea Forte, Emiliano De Cristofaro, Gianluca Stringhini, Jeremy Blackburn, and Damon McCoy. 2021. *"I’m a Professor, Which isn’t Usually a Dangerous Job": Internet-facilitated Harassment and its Impact on Researchers*. In *Proceedings of the ACM on Human-Computer Interaction (HCI)*.


Anh V. Vu, Lydia Wilson, Yi Ting Chua, Ilia Shumailov, and Ross Anderson


Twitter Developer. 2023. No Longer Support Free Access to the Twitter API.

Anh V. Vu. 2020. The Pandemic as Incels see it. Cambridge Cybercrime Centre Covid Briefing Paper.
