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

Table 1: The category, abbreviation, number of users, boards, threads, posts, and collection period of the 12 forums.

| Category	Forum Name	Abbr.	Users	Boards	Threads	Posts (% Empty)	Data Period
WHITE	STORMFRONT	ST	137 292	152	782 347	10 703 957 (2.65%)	2001/08-2023/04
SUPREMACY	VANGUARD NEWS NETWORK	VA	8 504	108	206 300	1 712 924 (1.65%)	2001/10-2023/04
INCELDOM	INCELSIS	II	13 825	7	439 675	9 030 678 (11.34%)	2017/11-2023/04
	INCELSNET	IN	3 614	6	23 110	364 183 (3.87%)	2017/11-2022/11
LOOKISM	LOOKISM	LK	16 362	10	770 429	7 294 617 (8.65%)	2015/06-2021/12
	LOOKSMAX	LS	12 570	17	650 126	10 177 131 (7.21%)	2018/08-2023/04
PICKUP	ROOSH V	RV	13 389	33	36 849	1 656 740 (4.81%)	2008/08-2023/04
ARTISTRY	PICKUP ARTIST	PA	65 115	43	177 438	943 853 (0.22%)	2006/03-2023/04
MEN'S	MEN GOING THEIR OWN WAY	MG	3 870	18	55 093	853 286 (2.02%)	2014/07-2021/05
MOVEMENT	GOING YOUR OWN WAY	GY	1 974	21	15 578	189 862 (1.00%)	2014/02-2023/04
ONLINE	KIWI FARMS	KF	61 432	51	51 337	10 528 403 (3.51%)	2013/01-2023/04
HARASSMENT	LOLCOW	LC	543	19	2 042	56 990 (6.96%)	2021/04-2022/06
6 Categories	12 Forums		338 490	485	3 210 324	53 512 624 (5.96%)	2001/08-2023/04
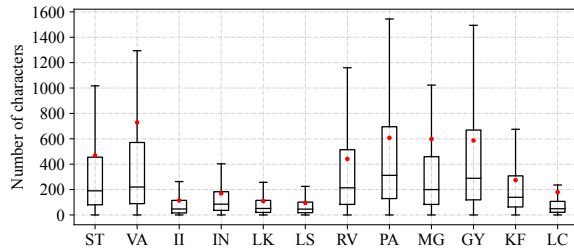
Figure 1: Post length distribution of each forum. Red dots are means. Abbreviations are shown in Table 1.
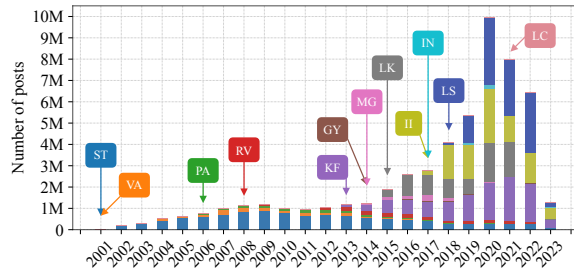

Figure 2: Number of forum posts; coloured labels mark forum inceptions. Abbreviations are shown in Table 1.

training to pick up, date, and have sex with women. The so-called 'coaches' on the forums share their experiences and discuss techniques for bullying and tricking women into sex and relationships; many of these techniques are seen by many women as sexual harassment (Bates, 2021). Forum users exhibit toxic, sexist ideas and violence against women (Banet-Weiser and Bratich, 2019).

MEN'S MOVEMENT. This includes two influential forums of movements whose goal is to completely separate men from women, protecting and preserving male sovereignty while exhibiting antifeminist, misogynistic and toxic-masculine ideals.

ONLINE HARASSMENT. This includes the two largest

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 (Zannettou 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).

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

Table 2: User overlaps across forums. $n(i, j)$ depicts the number of users on forum i who also register on forum j.

n	ST	VA	II	IN	LK	LS	RV	PA	MG	GY	KF	LC
ST	–	945	403	154	673	412	906	2 346	273	149	1 624	14
VA	945	–	56	27	86	48	109	203	35	20	166	0
II	403	56	–	205	549	769	144	297	44	31	365	4
IN	154	27	205	–	127	116	45	119	27	18	139	2
LK	673	86	549	127	–	993	268	596	79	52	546	8
LS	412	48	769	116	993	–	155	360	62	23	299	2
RV	906	109	144	45	268	155	–	815	117	79	477	5
PA	2 346	203	297	119	596	360	815	–	212	114	1 113	16
MG	273	35	44	27	79	62	117	212	–	120	158	1
GY	149	20	31	18	52	23	79	114	120	–	84	0
KF	1 624	166	365	139	546	299	477	1 113	158	84	–	97
LC	14	0	4	2	8	2	5	16	1	0	97	–

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.

essary. Our web scraper, based on CrimeBot (Pastrana 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).

References

Sarah A Aghazadeh, Alison Burns, Jun Chu, Hazel Feigenblatt, Elizabeth Laribee, Lucy Maynard, Amy LM Meyers, Jessica L O'Brien, and Leah Rufus. 2018. GamerGate: A Case Study in Online Harassment. *Online Harassment*.

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

Stephane J. Baele, Lewys Brace, and Travis G. Coan. 2019. From "Incel" to "Saint": Analyzing the Violent Worldview Behind the 2018 Toronto Attack. *Terrorism and Political Violence*.

Sarah Banet-Weiser and Jack Bratich. 2019. From Pick-up Artists to Incels: Con(fidence) Games, Networked Misogyny, and the Failure of Neoliberalism. *International Journal of Communication*.

Laura Bates. 2021. *Men Who Hate Women: the Extremism Nobody is Talking About*. Simon & Schuster.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*.

British Society of Criminology. 2015. Statement of Ethics.

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)*.

Thomas J. Holt, Joshua D. Freilich, and Steven M. Chermak. 2020. Examining the Online Expression of Ideology among Far-right Extremist Forum Users. *Terrorism and Political Violence*.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving Google's Perspective API Built for Detecting Toxic Comments. *Preprint arXiv:1702.08138*.

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

Sylvia Jaki, Tom De Smedt, Maja Gwóźdź, Rudresh Panchal, Alexander Rossa, and Guy De Pauw. 2019. Online Hatred of Women in the Incels.me Forum: Linguistic Analysis and Automatic Detection. *Journal of Language Aggression and Conflict*.

Bennett Kleinberg, Isabelle van der Vegt, and Paul Gill. 2021. The Temporal Evolution of a Far-right Forum. *Journal of Computational Social Science*.

Clare Lally and Rowena Bermingham. 2020. Online Extremism. *UK Parliament Research Briefing*.

Alice E. Marwick, Lindsay Blackwell, and Katherine Lo. 2016. Best Practices for Conducting Risky Research and Protecting Yourself from Online Harassment. *Data & Society*.

Debora Nozza and Dirk Hovy. 2023. The State of Profanity Obfuscation in Natural Language Processing Scientific Publications. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Catharina O'Donnell and Eran Shor. 2022. "This is a Political Movement, Friend": Why "Incels" Support Violence. *The British Journal of Sociology*.

Sergio Pastrana, Daniel R. Thomas, Alice Hutchings, and Richard Clayton. 2018. CrimeBB: Enabling Cybercrime Research on Underground Forums at Scale. In *Proceedings of the World Wide Web Conference (WWW)*.

Ildiko Pete, Jack Hughes, Andrew Caines, Anh V. Vu, Harshad Gupta, Alice Hutchings, Ross Anderson, and Paula Buttery. 2022. PostCog: A Tool for Interdisciplinary Research into Underground Forums at Scale. In *Proceedings of the IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Anna Talas. 2023. *Hacker's Paradise: Analysing Music in a Cybercrime Forum*. Master's Thesis, Department of Computer Science and Technology, University of Cambridge.

TechCrunch. 2022. Web Scraping is Legal, U.S. Appeals Court Reaffirms.

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*.

Anh V. Vu, Jack Hughes, Ildiko Pete, Ben Collier, Yi Ting Chua, Ilia Shumailov, and Alice Hutchings. 2020. Turning Up the Dial: the Evolution of a Cybercrime Market Through Set-up, Stable, and Covid-19 Eras. In *Proceedings of the ACM Internet Measurement Conference (IMC)*.

Anh V. Vu, Alice Hutchings, and Ross Anderson. 2024. No Easy Way Out: the Effectiveness of Deplatforming an Extremist Forum to Suppress Online Hate and Harassment. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*.

Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and Characterizing Hate Speech on News Websites. In *Proceedings of the ACM Conference on Web Science (WebSci)*.