

Decentralised Moderation for Interoperable Social Networks: A Conversation-based Approach for Pleroma and the Fediverse

Vibhor Agarwal¹, Aravindh Raman², Nishanth Sastry¹,
Ahmed M. Abdelmoniem³, Gareth Tyson⁴, Ignacio Castro³

¹University of Surrey, Guildford, United Kingdom

²Telefonica Research, Barcelona, Spain

³Queen Mary University of London, London, United Kingdom

⁴Hong Kong University of Science and Technology (GZ), China

{v.agarwal, n.sastry}@surrey.ac.uk, aravindh.raman@telefonica.com,
{ahmed.sayed, i.castro}@qmul.ac.uk, gtyson@ust.hk

Abstract

The recent development of decentralised and interoperable social networks (such as the “fediverse”) creates new challenges for content moderators. This is because millions of posts generated on one server can easily “spread” to another, even if the recipient server has very different moderation policies. An obvious solution would be to leverage moderation tools to automatically tag (and filter) posts that contravene moderation policies, *e.g.* related to toxic speech. Recent work has exploited the conversational context of a post to improve this automatic tagging, *e.g.* using the replies to a post to help classify if it contains toxic speech. This has shown particular potential in environments with large training sets that contain complete conversations. This, however, creates challenges in a decentralised context, as a single conversation may be fragmented across multiple servers. Thus, each server only has a partial view of an entire conversation because conversations are often federated across servers in a non-synchronized fashion. To address this, we propose a decentralised conversation-aware content moderation approach suitable for the fediverse. Our approach employs a graph deep learning model (GraphNLI) trained locally on each server. The model exploits local data to train a model that combines post and conversational information captured through random walks to detect toxicity. We evaluate our approach with data from Pleroma, a major decentralised and interoperable micro-blogging network containing 2 million conversations. Our model effectively detects toxicity on larger instances, exclusively trained using their local post information (0.8837 macro-F1). Yet, we show that this approach does not perform well on smaller instances that do not possess sufficient local training data. Thus, in cases where a server contains insufficient data, we strategically retrieve information (posts or model parameters) from other servers to reconstruct larger conversations and improve results. With this, we show that we can attain a macro-F1 of 0.8826. Our approach has considerable scope to improve moderation in decentralised and interoperable social networks such as Pleroma or Mastodon.

1 Introduction

Network effects and economies of scale favour concentration (Varian 2019). As a result, the Web has witnessed in-

creasing concentration into a few key players (Doan et al. 2022). For social networks, interoperability has been proposed as a remedy (Beck and Moore 2023) and the Digital Markets Act (DMA) will soon enforce it in the EU (EU 2023; Mahy 2023). We believe this opens new opportunities and challenges, particularly with regards to content moderation.

To understand these we look into the fediverse, an ecosystem of decentralised, interoperable platforms powered by a common underlying protocol, ActivityPub (W3C 2018). In contrast to X (previously, Twitter), Reddit and other centralised social networks, in the fediverse anyone can create an instance of a service by running open-source code on a server. Users can then sign up with such instances and use their services. Fediverse users can interact with each other regardless of the instance they have signed up with or even the service. For instance, users from an instance in the micro-blogging Pleroma can interact with users from another Pleroma instance, as well as with users from instances across other services such as Mastodon (micro-blogging) or Peertube (video streaming). The (coming) support for ActivityPub by new services such as Threads,¹ may expand this interoperable ecosystem even further.

Decentralisation and interoperability, however, pose new challenges. Decentralisation limits the human and data resources required for (semi) automating moderation. Interoperability limits the visibility and ability to affect content generated in a remote instance. In contrast to other volunteer-moderated communities (*e.g.*, Reddit) (Birman 2018; Iqbal et al. 2022), in the fediverse an administrator may need to handle content originating from a different instance or even service (*e.g.* Mastodon posts frequently appear in Pleroma instances). With toxic content being common in the fediverse (Hassan et al. 2021; Bin Zia et al. 2022), moderation can be challenging for administrators (Anaobi et al. 2023).

We believe that this is a challenge that further interoperability (*e.g.*, due to Threads supporting ActivityPub or thanks to DMA) will exacerbate. To study this problem and identify solutions, we use a comprehensive dataset comprising 713 instances and 16.5 million posts from a large micro-

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

¹<https://engineering.fb.com/2023/09/07/culture/threads-inside-story-metas-newest-social-app/>, last accessed 14 Sep 2023.

blogging fediverse service, Pleroma (Bin Zia et al. 2022). Our main assumption is the usage of ActivityPub and, as such, our solutions should be amenable to other fediverse services (*e.g.*, Mastodon). Our goal is to validate the viability of our approach rather than necessarily advocate for Pleroma or other fediverse platform to adopt it.

To cope with the limitations in terms of the data that an instance has available, we take a conversational approach². We capture the conversational context of each post using GraphNLI (Agarwal et al. 2022), a state-of-the-art graph learning-based framework. This is in stark contrast to previous works (Bin Zia et al. 2022; Kurita, Belova, and Anas-topoulos 2019; Risch and Krestel 2020) that look at each post in isolation. Because of fediverse’s interoperability, conversations are scattered across instances. We reconstruct all the conversations in Pleroma and identify 2 million conversations with more than two posts. This proved to be a complex endeavour: we discover that each instance participating in a conversation can have a different and partial view of the complete conversation. Indeed, interoperability poses an additional challenge: moderators may have different partial views of the conversations and therefore, may develop different perceptions about the content to be moderated. We find that up to 14 instances participate in the same conversation and that the level of fragmentation increases with the number of instances. We believe this challenge is inherent to interoperability.

Then we obtain toxicity scores of each of the posts in Pleroma conversations from Perspective API. We train GraphNLI model on our Pleroma conversations for predicting these toxicity scores. We find that large (>75 th percentile by the number of posts) and medium instances (between 25th and 75th percentile by the number of posts) are self-sufficient: each instance can develop a model trained just on its own data and achieve a performance similar to what would be obtained by a centralised platform with access to data from all instances. This is a significant improvement over earlier work, which did not exploit conversational context (Bin Zia et al. 2022).

This strategy fails with smaller instances though (*i.e.*, <25 th percentile). Smaller instances do not contain a sufficiently large fragment of the conversation to effectively capture the surrounding conversational context. We overcome this with two information-sharing strategies. We first evaluate a *data-borrowing approach*, where small instances obtain all toots from all conversations with at least 5 toots from the largest instance they federate with. This results in large improvements with an overall macro-F1 of 0.8258 for small instances. To limit the sharing of private information, we also evaluate a *model-borrowing* approach, where small instances adopt the model trained on large instances and fine-tune the model parameters. We achieve a further increase in performance ($\approx +3\%$ macro-F1).

In summary, we analyse the moderation challenges resulting from decentralisation and interoperability. In doing so, we perform the first-ever dissection of *conversations* in the

fediverse using data collected from Pleroma, a large fediverse micro-blogging platform. We observe that interoperability results in heavily decentralised conversations and examine its implications on content moderation. We address the following Research Questions (RQs):

- **RQ1.** *What do Pleroma conversations look like?* We reconstruct all conversations in 713 Pleroma instances in Section 5 by linking posts and reposts of these posts across instances. We find over 2 million conversations that have more than two posts. Interestingly, users see *fragmented conversations* because different instances have different fragments of the same conversation. The fragmentation of conversations and the likelihood of partial federation increases with the number of instances involved in a conversation. Unfortunately, we also find that some of these larger conversations are also more likely to contain toxic content. This leads us to develop strategies to identify such content (RQ2 and RQ3).
- **RQ2.** *Can instances locally detect toxic content by exploiting conversational context?* Using GraphNLI, a state-of-the-art graph-based framework (Agarwal et al. 2022, 2023), we find in Section 6.2 that the model trained on local content succeeds in identifying toxicity in Medium and Large instances (macro-F1 of 0.7675 and 0.8837, respectively). However, we find that this method *does not detect any toxic content at all for small instances*, resulting in an F1-score of 0.
- **RQ3.** *Can large instances help small ones in detecting toxic content?* We propose and evaluate strategies where small instances exchange information with larger ones in Section 6.3. First, we explore modifications to the current federation mechanism, to federate the *full conversation*. Then we propose a strategy that uses toots or posts from the largest instance that an instance federates with; we call this approach *toot exchange*. Finally, we develop a more advanced approach wherein a small instance adapts the model trained by the largest instance it federates with; we call this *model exchange*. While full conversation federation does not help the small instances, we find that both toot and model exchange are extremely effective, with model exchange improving the macro F1-score on small instances by up to 28.26%.

2 Background & Related Work

2.1 Pleroma Background

Pleroma is an open-source decentralised and interoperable micro-blogging server platform where users create their own social communities hosted on independent servers, aka **instances**. Each Pleroma instance works similarly to Twitter, allowing users to register new accounts and share **statuses** (aka **toots**) with their followers — equivalent to a tweet on Twitter. Users can also **repeat** others’ statuses — equivalent to retweeting on Twitter.

In contrast to a user on Twitter, a Pleroma user on one instance can follow users from other instances as well, and Pleroma federates content from other instances as required

²The code is publicly available at <https://github.com/vibhor98/decentralised-moderation-pleroma>, last accessed 23 Mar 2024.

(federation is described further below). This allows *decentralised* operation and administration — each instance can function independently of other instances.

Interoperability. Decentralisation (*e.g.*, fediverse, IPFS (Trautwein et al. 2022)) has emerged as a way to reduce the growing concentration in the Internet (Doan et al. 2022). To further increase competition and reduce concentration (Doan et al. 2022; Varian 2019), interoperability has been proposed (Beck and Moore 2023) and will soon be mandatory for large (“gateway”) social networks in the EU thanks to the DMA (EU 2023). While work is underway to implement interoperability for these large players (Mahy 2023), the W3C’s ActivityPub subscription protocol (W3C 2018) already supports interoperability.

Fediverse. The fediverse refers to the growing group of ActivityPub compatible, and therefore interconnected, platforms. Pleroma is thus interoperable with other services that support ActivityPub, including micro-blogging implementations such as Mastodon (He et al. 2023).

Federation. Fediverse instances are *interoperable* and can **federate**: users registered on one Pleroma instance can follow users registered on another instance from Pleroma (or any other ActivityPub supporting service). This results in the instance **subscribing** to posts performed on the remote instance, such that they can be pushed across and presented to local users. We refer to users registered on the same instance as **local**, and users registered on different instances as **remote**. Note that a user registered on their local instance does *not* need to register with the remote instance to follow the remote user. Instead, a user just creates a single account with their local instance; when the user wants to follow a user on a remote instance, the user’s local instance performs the subscription on the user’s behalf.

2.2 Related Work

Decentralised Social Networks. Powered by the ActivityPub subscription protocol (W3C 2018), multiple social networks have recently emerged and gained popularity, thanks, in part to the recent acquisition of Twitter by Elon Musk (He et al. 2023). Decentralisation is challenging and earlier work has found that strong re-centralisation forces exist in Mastodon (Raman et al. 2019). Recent work found that, paradoxically, while most users join the larger instances, smaller instances attract more active users (He et al. 2023). The general characterisation of fediverse services has focused on the content of the posts and found that user behaviour does vary across instances (La Cava, Greco, and Tagarelli 2022b,a). In contrast, we focus on conversations rather than posts: we reconstruct Pleroma’s conversations and characterise them. We find that instances frequently have a different and fragmented view of a conversation due to the partial federation of the toots that compose it.

Toxicity Detection. Many studies have examined toxic activities in social networks, including for Twitter (Ribeiro et al. 2018; Burnap and Williams 2015; Waseem and Hovy 2016), Reddit (Chandrasekharan et al. 2017; Mohan et al. 2017; Almerekhi, Jansen, and Kwak 2020; Wu et al. 2022;

Yin et al. 2023), and 4Chan (Bernstein et al. 2011; Pasavava et al. 2020). Deep learning has been commonly used to classify toxic posts (Badjatiya et al. 2017; Cao, Lee, and Hoang 2020). We also use a deep learning approach, however, we explicitly include conversational context. We use GraphNLI (Agarwal et al. 2022), a recent graph-based deep learning model. GraphNLI was originally used to infer polarities in the debating platform, Kialo. We repurpose GraphNLI for toxicity detection in a decentralised environment.

Moderation in the Fediverse. Several works have recently addressed the challenges of decentralised moderation. Zignani et al. (2019) explored content warnings from Mastodon, Hassan et al. (2021) analysed policies in Pleroma, and Anaobi et al. (2023) measured the heavy moderation load of instance administrators. Similar to fediverse services, centralised platforms such as Reddit also rely on volunteer moderation (Waseem and Hovy 2016). Differently from the fediverse, these centralised platforms have no interoperability with other platforms.

Closest to our work is Bin Zia et al. (2022), which measured the presence of toxicity in Pleroma and proposed a model exchange mechanism between instances to improve local moderation capabilities. Bin Zia et al. (2022) focuses exclusively on the 30 largest instances and individual posts. As a result, local models perform poorly (macro-F1 score less than or equal to 0.6 for the great majority of instances) and all instances need to exchange models. Our conversational approach improves those results for all 713 instances, rather than just the largest 30. The additional conversational context allows large and medium instances to be self-sufficient with similar performance to a global model. We also look into smaller instances. While smaller instances typically have, by nature, a smaller moderation load (Gehl and Zulli 2022), we believe that supporting semi-automated moderation for any player – large or small is useful. Semi-automated moderation can have potential benefits in the context of Pleroma and the fediverse: 1) federated conversations might become viral and overwhelm the administrator; 2) some administrators might be time-constrained, *e.g.*, because they admin multiple instances (Hassan et al. 2021), or because this is simply a side activity for them; 3) it might enable new use-cases or allow more people to become administrators thanks to the reduced moderation load.

3 Problem Statement

Interoperability and decentralisation of platforms such as Pleroma give communities the independence to exchange ideas and content without having to rely on the whims and gatekeeping of centralized platforms like Twitter. However, the interoperable nature of fediverse platforms leads to new questions about the *means by which* content can be moderated in a decentralised way, without the scale of platforms like Twitter, or when the full conversation is fragmented across instances.

Decentralised Moderation. The administration of fediverse instances is strictly decentralised, *i.e.* the administrator of an instance is responsible for formulating and implementing its

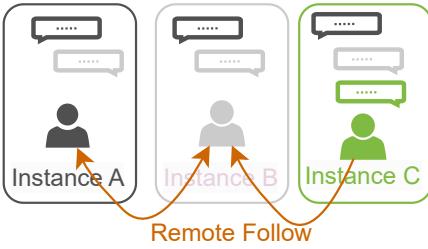


Figure 1: Example scenario of a fragmented conversation in the fediverse due to partial federation: Instance C has full visibility on the conversation (3 toots) but instances A and B can only see a fragment of the conversation (2 toots)

own moderation policy. However, administrators have only limited control of the content of the instance they administer, as content generated on one instance can easily spread to another one thanks to the content federation enabled by the ActivityPub protocol. With instance administrators typically operating in a pro bono voluntary fashion, moderation and other administrative aspects of running an instance can be rather challenging. While machine learning could in principle help to semi-automate this task, the limited amount of data on each instance constrains this possibility.

Fragmented Conversations. Toots are usually part of a longer conversation involving other toots from multiple users. This context is frequently necessary to understand the meaning and moderate content adequately. However, the interoperable nature of the fediverse can result in a partial federation where a conversation is fragmented across instances. Conversations can be fragmented partially due to moderation and blocking by other instances (Hassan et al. 2021). However, conversations can also be fragmented due to how federation works (even when no instance is blocked).

For example, suppose a user in instance A posts a toot. If a user in instance B follows instance A's user, they will see this post federated to instance B's. Now if this user takes part in a conversation, another user in instance C will see their conversation since the user in instance C follows instance B's user (*cf.* Figure 1). However, if the user in instance A does not follow the user in instance C, the toot of instance C's user will not be federated to instance A even though instance A's user originated the conversation.

Such *partial federation* can result in *fragmented conversations* where different administrators see different parts of a conversation and therefore different contexts of the content they have to moderate. For instance, in the above example, the administrator of instance A (or a machine learning model running on their behalf) might need the context from instance C in order to determine the toxicity of a post and whether a particular post is hate speech. Further interoperability is likely to reproduce this challenge.

4 Dataset and Methodology

In this section, we approach the problem of reconstruct-

ing a global, centralised view of conversations in the decentralised fediverse, and use the task of determining the toxicity of a post to investigate the impact of the fragmented conversation views on overall conversational understanding. Here, we give details of the dataset we use, and how we process it to construct conversation trees and to attach toxicity scores.

4.1 Pleroma Dataset

We use the Pleroma data crawled by Bin Zia et al. (2022)³. The dataset consists of 713 Pleroma instances with all the publicly available toots before January 22, 2021. In total, the dataset contains 16.5 million toots from Pleroma, one of the largest platforms in the fediverse, by both content and user counts in the fediverse (Bin Zia et al. 2022). The dataset represents a complete collection of all available instances at the time of data collection. Each toot in the dataset includes information about the author, text content, associated media, timestamp, number of likes, number of reblogs, and any self-tagged content warnings.

We enhance this raw dataset in two ways: First, we reconstruct the global conversations, by tracing the threads of posts and replies across instances. Second, using Google's Perspective API (Jigsaw 2023), we assign toxicity scores to each post in over 2 million conversations.

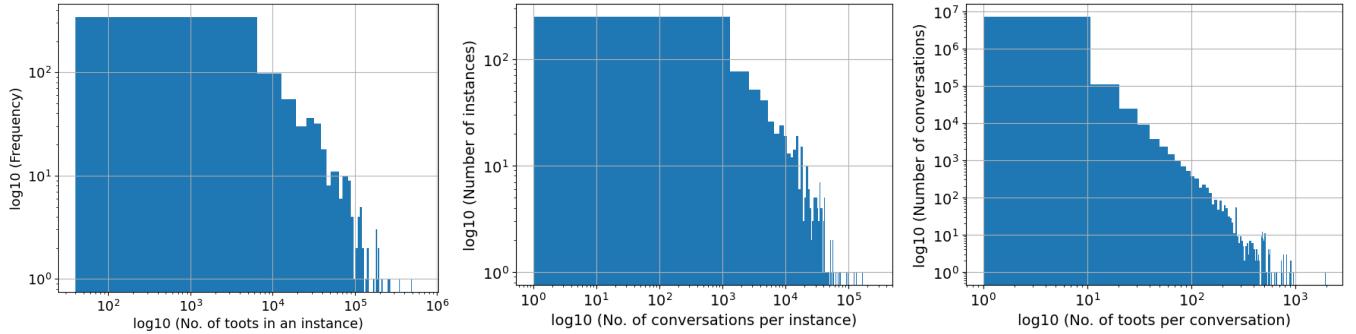
4.2 Constructing Conversation Trees

A conversation tree is a tree-structure where nodes represent the toots in conversations and a directed edge from a child to its parent node denotes that a child node replies to the parent node (Agarwal et al. 2022). A comment can have multiple replies, but it can reply to exactly one comment in Pleroma. Therefore, every node in a conversation can have only one outgoing edge (replied to) but several incoming edges (replies), thus forming a tree.

Constructing complete conversations in the fediverse is challenging as conversations can be federated across multiple instances when users from different instances participate in the same conversation. Partial federation can easily occur, as some toots are not federated across all the instances where a particular conversation is present. This leads to *fragmented conversations*. We regard a “complete conversation” as the conversation tree that includes all toots that span as replies from the first toot in the tree. An instance has a fragmented conversation whenever such instance can see less than 100% of the complete conversation.

An obvious question is how such conversations can be reconstructed. Unfortunately, we find that toot IDs, parent IDs and conversation IDs are unsuitable for reconstructing conversations as these IDs are local to the instance where we observe them. Thus, we use the toot URL as a unique identifier for toots in the dataset as it is unique across instances. By mapping local toot IDs and parent IDs with toot URLs, we construct complete conversation trees which contain toots from different instances. For each node (i.e., toot), we also store the information of which instances a toot is federated

³<https://github.com/harisbinzia/mastodoner>, last accessed 1 Apr 2024.



(a) Histogram of the number of toots in Pleroma instances.
(b) Histogram of number of conversations per instance.
(c) Histogram of the number of toots per conversation.

Figure 2: Histograms for Pleroma instances.

to in order to study the federation of conversations. Overall, we identify 7.2 million conversation trees in our dataset.

4.3 Toxicity Labels

To study toxicity and moderation in Pleroma, we use toxicity scores from Google Jigsaw’s Perspective API (Jigsaw 2023) as ground truth toxicity scores, following similar studies for Pleroma (Bin Zia et al. 2022), Mastodon (He et al. 2023), 4chan (Papasavva et al. 2020) and Voat (Papasavva et al. 2021). We set “doNotStore” option to true so that the Perspective API does not store the data after scoring them for toxicity. By extending the original dataset of Bin Zia et al. (2022) which used toxicity scores for the top 30 instances, we obtain toxicity scores for all posts in all conversations across all instances.

The Perspective API provides a toxicity score between 0 and 1, where 1 means the highest toxicity level. Following Bin Zia et al. (2022); Papasavva et al. (2020, 2021), we consider a toot to be toxic if its toxicity score is greater than 0.5 (moderately toxic) to capture a broad range of toxic content. For completeness, we also experimented with a stricter toxicity threshold of 0.8 and observed similar proportions of toxic toots in different Pleroma instances. To ensure the generality of our approach, we use the actual toxicity scores as the ground truth and train our models with them for the regression task of toxicity prediction. This ensures that our findings apply regardless of the actual threshold that an administrator might choose. More generally, we believe that Perspective labels are just one example and the approach proposed could be implemented with alternative human-based labels. Unfortunately, due to the cost-intensive nature of labelling, we do not have such an alternative.

5 Characterising Conversations (RQ1)

Having constructed a global view of all Pleroma conversations, we next proceed to characterize them. We first perform a basic characterization to understand the distribution and size of the conversations, then investigate the presence of toxic content, and finally develop an understanding of how individual instances or users within a conversation see the

conversation, finding that conversation can be fragmented due to the nature of federation across instances.

5.1 Basic Characterisation

Overall, we identify 7.2 million conversations in 713 Pleroma instances. Figure 2b shows the histogram of the number of conversations per instance. On average, the number of conversations per instance is 10,380.65, while the instance with the largest number of conversations has 267,717 conversations. The average number of users per instance who have at least one toot either posted or federated is 1609.60, while the maximum number of users per instance is 9,824. Next, we look at the number of toots per conversation. Figure 2c shows the histogram of the number of toots per conversation. We find that only 29% of the conversations have 2 or more toots. Note that even a single toot is itself a conversation, consisting of just the root of the conversation tree. The average (median) number of toots per conversation is 2.11 (1). If we only consider conversations with more than two toots, the average (median) conversation has 6.83 (4) toots, whereas the maximum is 1,952 toots.

5.2 Toxicity in Pleroma Conversations

Presence of toxic content. Figure 3 shows the top 20 Pleroma instances with the number of toxic toots, sorted in descending order, indicating that a few instances are responsible for a large proportion of toxic content. For example, *my.dirtyhobby.xyz* contributes about 23.33% of the toxic toots. Similarly, *plero.antopie.org* and *social.feathers.link* have 4.7% and 3.3% of the toxic content respectively. Unfortunately, due to the nature of the federation of content, such content is not constrained solely to a few instances but instead can potentially spread more widely to other instances whose users may take part in the toxic conversations. Figure 5a shows the Cumulative Distribution Function (CDF) of the number of toxic toots per conversation. On average, the number of toxic toots per conversation is 0.004, while the maximum number of toxic toots in a conversation can be as high as 77, which suggests that we should be taking a conversation-centric view of toxic toots. This motivates our

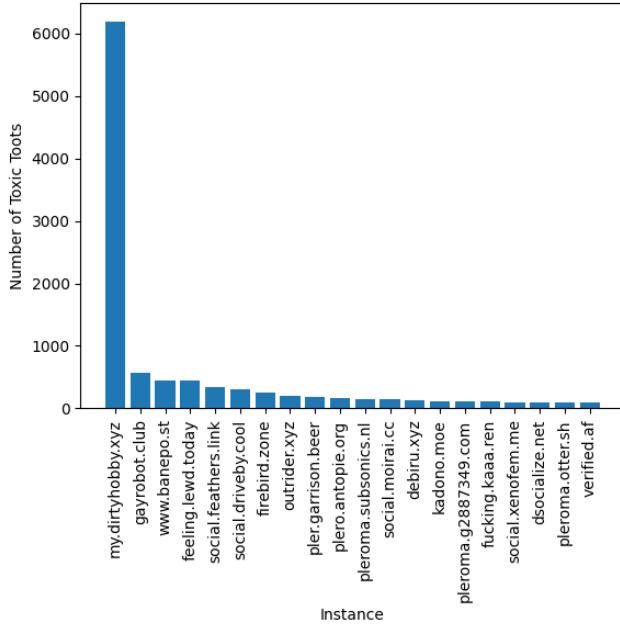


Figure 3: Top 20 instances with the number of toxic toots, sorted in descending order.

study in the next section, which focuses on the whole conversation context to predict toxic content.

Popularity of toxic content. We next examine the relative popularity of toxic versus non-toxic content from different angles. First, Figure 4a shows that toxic toots have a higher likelihood of “going viral” across multiple instances: Whereas most toots (both toxic and non-toxic) are federated to a handful of instances, the maximum number of instances a toxic toot is federated to is 13, while the maximum is 8 for non-toxic toots.

Next, we compare the engagement metrics of toxic versus non-toxic toots. Figure 4b shows the CDF of toot reblogs (reshares) for toxic and non-toxic toots. On average, the number of reblogs that toxic toots receive is 0.556, whereas non-toxic toots receive 0.271 reblogs. This suggests that toxic toots are reblogged (re-shared) 2x more than the non-toxic toots in Pleroma conversations. Figure 4c also shows the CDF of the number of direct replies to toxic versus non-toxic toots in Pleroma conversations. The average number of replies to toxic toots is 1.482, while the average number of replies to non-toxic toots is 1.354. Thus, toxic toots receive a higher number of replies than non-toxic toots.

Collectively the above results clearly show that toxic toots receive higher levels of engagement (reblogs and replies) in Pleroma. Toxic content may spread faster (i.e., have more views overall) because they receive a higher number of reblogs (2x) than the non-toxic toots.

Where can we find toxic content? Finally, we examine which kind of conversations are more likely to contain toxic content, and within such conversations, where are the toxic toots located.

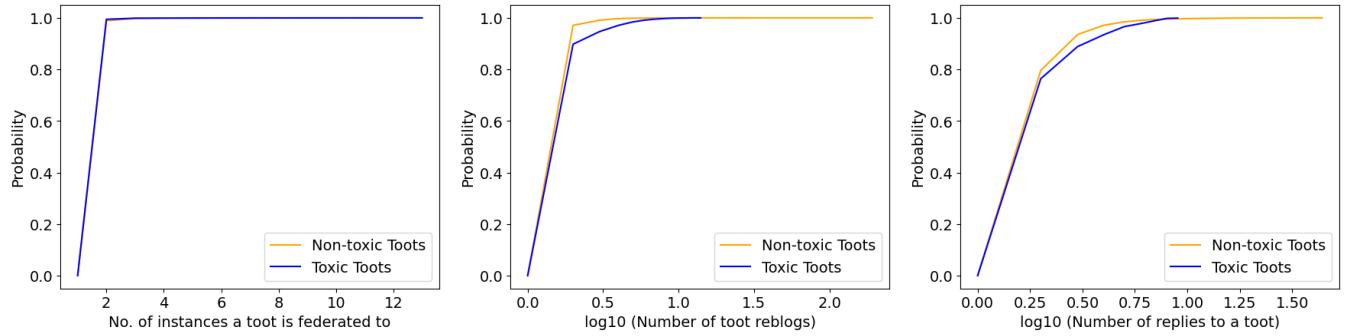
Figure 5b shows a bar plot of the percentage of conversations with at least 1 toxic toot versus the conversation length. As the conversation length increases, the likelihood of having toxic toots also increases. Thus, the longer a conversation, the more likely it is to have some toxic statements.

We look deeper and ask where, within those toxic conversations, are toxic toots to be found. Figure 5c shows the bar plot of the number of toxic toots per toot level in a conversation tree. We define the “level” of a toot as follows: The root node in a conversation tree is at level 0; its immediate children are at level 1. We proceed in this fashion with increasing levels until we reach the leaf nodes. From the figure, we see that it is more likely to have toxic toots at the beginning of the conversation near the root node and other ancestor nodes. This suggests that further down in a conversation (closer to the leaves), the immediate context or content of a toot may be insufficient to detect toxicity and it may require context from further up the conversation tree. This motivates our adoption in Section 6 of a graph-walk based approach (Agarwal et al. 2022) that can gather context from farther up the conversation tree to classify toxicity.

5.3 Federation and Fragmented Conversations

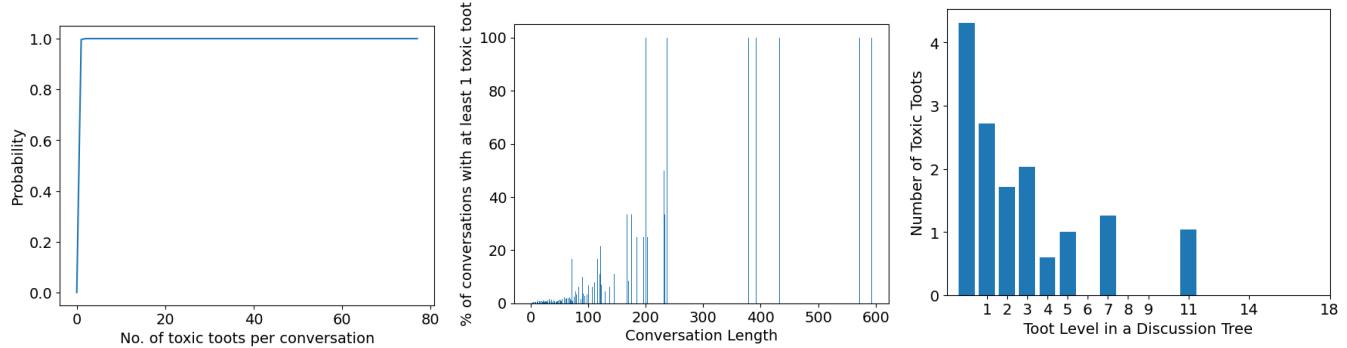
Content federation is at the very essence of the fediverse and allows users to interact with each other regardless of the instance they are signed up with. On average, a Pleroma instance federates with 17.73 instances, while the maximum number of instances an instance federates with is 118. Figure 6a shows the CDF of the number of local and federated toots per conversation in Pleroma instances. Note, local toots within an instance are the toots which originated in that instance, while federated toots are the ones which are federated from other instances. The average number of local toots per conversation is 0.018, while the average number of federated toots per conversation is 2.091. Therefore, on average 99% of the toots in a conversation are federated and the majority of the toots in a Pleroma instance are federated as well. Next, we look at how many instances a conversation is federated to, irrespective of where the conversation has originated. On average, the number of instances to which a conversation is federated is 1.009, whereas the maximum number of instances in a conversation goes up to 14.

Since conversations are composed of toots from different instances, partial federation is likely to result in fragmented conversations, as discussed in Section 3. Figure 6b shows how partial federation impacts the visibility of an instance on conversations depending on the number of instances involved in such conversation. The boxplot represents the variation in the percentage of federated toots per conversation for the instances involved in conversations in Pleroma. The x-axis shows the number of instances involved in a conversation, and the y-axis is the percentage of toots out of the complete conversation (100% toots) that are federated to a given instance. A conversation is fragmented whenever partial federation takes place, i.e., when the number of toots in a conversation is below 100%. It can be partially due to moderation and blocking of other instances. However, conversations can also be fragmented due to how federation works. It can occur when users from different instances take part in



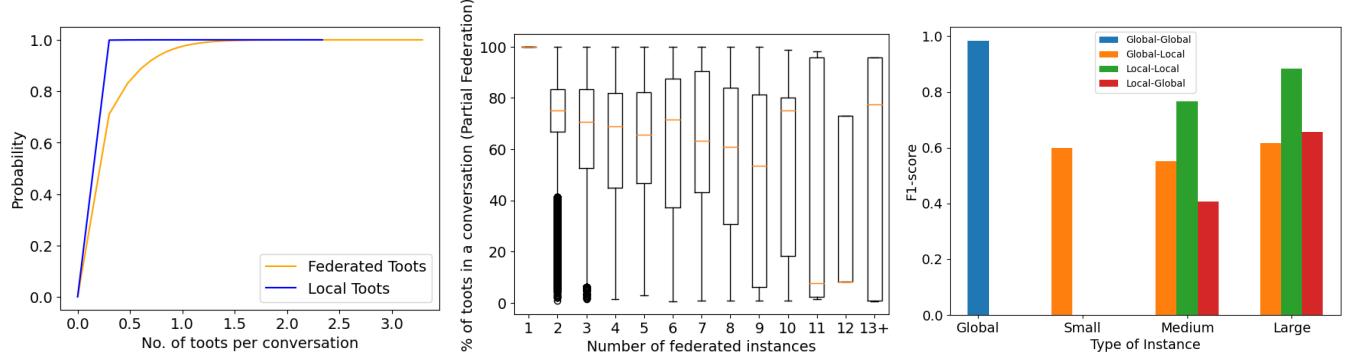
(a) Number of instances toxic and non-toxic toots are federated to. (b) Number of toot reblogs for toxic and non-toxic toots. (c) Number of direct replies to toxic and non-toxic toots.

Figure 4: Cumulative Distribution Functions (CDFs) for toxic toots in Pleroma instances.



(a) Number of toxic toots per conversation. (b) Percentage of conversations with at least one toxic toot versus conversation length. (c) Number of toxic toots per toot level in a discussion tree. Y-axis is on log-scale.

Figure 5: Characterisation of Toxic Toots in Pleroma conversations.



(a) Number of local and federated toots in Pleroma instances (log-scale on X-axis). (b) Box plot with the variation in the percentage of federated toots per conversation for instances involved in Pleroma conversations. (c) Bar plot with F1-scores for different types of Pleroma instances and GraphNLI models.

Figure 6: Conversations in Pleroma and F1-scores for toxicity detection.

a fediverse conversation, as discussed in Section 3. We observe from Figure 6b that the fragmentation of conversations and its likelihood increases with the increase in the number of instances involved in a conversation.

Such partial views of a conversation are likely to hamper

the ability of an individual instance to discover toxic content, especially since, as detailed above, context from farther up in a conversation tree is important.

6 Decentralised Toxicity Detection

In Section 5, we have shown that toxicity is common in Pleroma and that it spreads faster and has more engagement (in terms of reblogs and replies) than non-toxic content. It is a challenging task to not only monitor a given instance’s toots but also the federated toots coming from other instances. In this section, we explore the feasibility of individual instances being able to protect themselves from toxic content using automated machine learning-based approaches. Importantly, Section 5 also showed that context from farther up in a conversation can potentially be relevant when labelling a post as toxic or not.

6.1 Does Global Conversational Context Help in Detecting Toxic Content?

The characterisation in Section 5 suggests that, to infer the toxicity of a toot, understanding its conversational context is beneficial. To test whether this holds true in practice, we adapt GraphNLI (Agarwal et al. 2022), which uses graph walks to capture the conversational context.

GraphNLI (Agarwal et al. 2022) is a graph-based deep learning model which uses graph walks to capture the conversational context. In contrast to its original purpose to measure polarity, we adapt the model for regression tasks to infer toxicity scores. Using our constructed conversation trees, we first employ biased root-seeking random walks for each of the nodes in a tree to capture the relevant conversational context. A root-seeking random walk is a random walk which is biased towards the root as it selects the parent and other ancestor nodes with a higher probability than the sibling and child nodes in a conversation tree (Agarwal et al. 2022). As shown earlier, context from farther up in the conversation tree may be important for labelling toxic posts. We input the toots along with the conversational context into the GraphNLI model for predicting toxicity scores. The model is trained for 3 epochs with an Adam optimiser.

We compare GraphNLI with the popular transformer-based BERT (Devlin et al. 2018) that has no conversational context, i.e., the BERT model exclusively relies on the specific toot text for which it is inferring the toxicity. To assess which model performs better, we train both models with the complete Pleroma dataset. For generality, we model toxicity detection as a regression task that predicts the toxicity score that would be assigned by the perspective API for that post. Then, we convert the predicted scores into a classification (hate speech or not), based on whether the predicted toxicity score is greater than or less than the threshold of 0.5, as previously used in various studies and described in Section 4.3.

In all the models both here and subsequently in this paper, we split our data into 80:20 ratios for training and testing. We use a single 24 GB Nvidia Titan RTX GPU for training and each epoch took about an hour to train GraphNLI on the entire training set. We use macro-F1 scores (due to dataset imbalance) along with accuracy to measure the performance of the classification models, and the mean-squared error (MSE) to assess the regression task. Note that the lower the MSE loss, the better the performance, whereas in the case of accuracy and macro-F1, the higher the better.

Model	MSE	Accuracy	macro-F1
BERT	0.0686	0.8487	0.4591
GraphNLI	0.0056	0.9744	0.9846

Table 1: Performance of machine learning models for toxicity detection trained on the complete Pleroma dataset.

Table 1 shows the performance of both models. Our model with conversational context, GraphNLI, clearly outperforms the BERT baseline, with an overall MSE of 0.0056 and 0.9846 macro-F1 score.

6.2 Global versus Local Moderation (RQ2)

We term the above model, which uses a centralized global view of the entire data, as our **global** model. While the performance here is very robust, it is also infeasible in practice as individual instances will not have such a global view of data.

The decentralised moderation that characterises the federated implies that any form of administrative task or automation can only rely on local data. The partial federation and fragmented conversations discovered in the previous section may therefore present a barrier to effective automated models. To test the limitations posed by decentralization, we develop models that can work **locally** in each Pleroma instance using the data available just on that instance, and compare the results with our global model.

As we expect model performance to vary depending on the amount of training data available, we divide the training data of the Pleroma instances depending on the instance size (i.e., according to the number of toots). We classify instances as *Small*, *Medium*, and *Large*, depending on whether they are in the 25th quantile, between 25th and 75th quantile, or greater than 75th quantile, respectively. Figure 2a shows the histogram of the number of toots in Pleroma instances. In the distribution of toots, the 25th quantile is 1,290 toots, the 50th quantile is 7,020 toots, and the 75th quantile is 26,960 toots; thus large instances are an order of magnitude larger than the small instances.

Table 2 shows the toxicity detection performance of GraphNLI for *Global* and *Local* models for Pleroma instances of different sizes. We report average results for each instance type based on the size and find an overall standard deviation of 0.89. Global-Global means that the model is trained globally in a centralised manner and inferred globally as well; thus creating a situation akin to today’s centralised platforms such as Twitter, which have all the data in one location, allowing them to train large and accurate models. In the decentralised version, however, we can have other combinations: Global-Local refers to global training and local inference (i.e., the random walks of GraphNLI have access to conversation context from all instances during training, but during inference, the walks can only use the local instance, and thus have access to only the local context). This is more realistic than global training and global inference, as data exchange across instances only needs to happen once during training; rather than during inference, which

GraphNLI	Small			Medium			Large		
	MSE	Accuracy	macro-F1	MSE	Accuracy	macro-F1	MSE	Accuracy	macro-F1
Global-Global	0.0056	0.9744	0.9846	0.0056	0.9744	0.9846	0.0056	0.9744	0.9846
Global-Local	0.0506	0.8726	0.60	0.0461	0.9093	0.5513	0.0502	0.8637	0.6162
Local-Global	0.2781	0.3230	0.0	0.1758	0.4358	0.4081	0.0781	0.6565	0.6557
Local-Local	0.0091	0.9743	0.0	0.0076	0.9729	0.7675	0.0038	0.9731	0.8837

Table 2: Toxicity detection performance of GraphNLI for Global v/s Local models at Pleroma instances of different sizes.

happens for every toot. Similarly, the local-global model involves training using the local instance data and using the global data (i.e., the conversation across all instances involved) for inference. The local-global model needs global data for every inference made (i.e., for every new post/toot), and so is less feasible than the global-local model which uses global data only during training. However, this mechanism is shown for completeness. Finally, the local-local model is the most realistic, as it uses only the local instance for both training and inference.

Figure 6c shows the F1-scores obtained for all combinations. The global training-global inference model is a single model representing a centralised equivalent of the decentralised fediverse, only shown here as a benchmark. When there is local training or local inference, there is per-instance variation in performance. We show this as separate bar plots for instances of different sizes – small (<25th percentile in number of toots), medium (>25th percentile and <75th percentile) and large (>75th percentile). For the small instances, local training does not work at all; and we are unable to detect *any* hate speech (i.e., a toot with toxicity > 0.5). On the other hand, for medium and large instances, local training and local inference appear to perform better than variants involving either global training or global inference; suggesting that local context, when sufficiently rich, is preferable to global context from farther away on a different instance. Global training Local inference seems to be the only choice with macro-F1 of 0.60 for moderating Small instances. But the performance of the model on Small instances is not good enough and most importantly, Global training violates the principles of the decentralised web, which is based on each instance being as independent as possible. Hence, we further investigate how we can improve the local moderation of Small instances.

6.3 Federation Strategies to Improve Local Moderation (RQ3)

We demonstrated in the previous section that local moderation is effective for large and medium-sized instances. Unfortunately, this strategy fails for Small instances due to the limited amount of data (and in particular toxic data). While the focus is usually on the larger instances, ensuring that smaller ones are viable can be critical to reducing the barrier of entry for new players and promoting the long-term sustainability of the ecosystem.

We now devise strategies to solve this problem. We experiment with three federation strategies to improve the local moderation: *(i) Full Conversations:* we federates toots

from other instances to complete the fragmented conversations (due to partial federation), thereby increasing the training data and conversational context for our GraphNLI model; *(ii) Toot Federation:* we select the largest instance that federates with the target instance and then federates all of the toots in conversations with a length of more than 5. *(iii) Model Sharing:* we share the trained toxicity detection GraphNLI model from the largest federating instance with the target instance. The target instance then fine-tunes on top of the already trained model using its local conversations.

Table 3 shows the performance of the GraphNLI model with Local training and Local inference with different federation strategies for local content moderation. We observe that even with full conversations, macro-F1 is still 0 for Small instances since no toxic toots are predicted correctly because of small datasets available locally. Toot federation with the largest federating instance works well with an overall accuracy of 0.9678 and 0.8258 macro-F1 score for Small instances. The model sharing strategy performs the best with an accuracy of 0.9786 and 0.8826 macro-F1 score. This represents an improvement of 28.26% in macro-F1 score over our previous best-performing variant (Global-Local) for Small instances. Model sharing improves the local moderation of Small instances in a privacy-preserving way since no actual toots content needs to be shared. However, it requires a communication channel with high bandwidth to share the weights of deep learning models between Pleroma instances.

In contrast, the toot federation strategy is feasible through existing channels between instances, since these instances continuously federate content among themselves. For instance, an administrator could simply choose to follow a specific account from a remote instance it is already federating with to ensure additional toot federation. We also experiment with these federation strategies on Medium and Large instances as shown in Table 3. For Medium instances, the macro-F1 score goes up to 0.8169 for the model-sharing strategy, which is an improvement of about 5%. We find that there is only a slight improvement for Large instances because they already gain good performance on toxicity detection through local training only due to the large training dataset. Therefore, Medium and Large instances may not have strong incentives for federated moderation and may choose to deal with moderating toxic content locally. However, toot federation or model sharing is very beneficial for small instances. Since the model-sharing strategy performs the best for small instances, and only requires sharing the model weights (no text or toots), model-sharing can be the preferred strategy with better privacy-preserving properties.

Local-Local Federation Strategy	Small			Medium			Large		
	MSE	Accuracy	macro-F1	MSE	Accuracy	macro-F1	MSE	Accuracy	macro-F1
Full Conversations	0.0230	0.9481	0.0	0.0073	0.9654	0.7696	0.0041	0.9711	0.8826
Toot Federation	0.0063	0.9678	0.8258	0.0067	0.9747	0.8077	0.0040	0.9813	0.8841
Model Sharing	0.0075	0.9786	0.8826	0.0061	0.9773	0.8169	0.0038	0.9839	0.8872

Table 3: Performance of GraphNLI model with different federation strategies for local content moderation of instances of different sizes.

6.4 Limitations and Ethical Considerations

The data includes user information and public posts. In collecting it, we took an “algorithmic thinking in the public interest” approach (Luscombe, Dick, and Walby 2022). We believe that our study has large societal benefits that outweigh potential risks: content moderation remains a challenge. We believe that research into novel moderation scenarios and approaches has the potential for large societal benefits that expand from Pleroma and the fediverse to other platforms, particularly if they are interoperable. The benefits include: (i) scrutiny into current moderation practices; (ii) identification of problems; (iii) potential solutions to help address them; and (iv) timeliness, as this is a young ecosystem with very different moderation and interaction mechanisms that are currently understudied but have a large potential to become mainstream.

We exclusively use publicly accessible information, adhering to well-established ethical protocols for collecting social data. We obtained Ethics approval from both the home institution of one of the authors, and independently from REPHRAIN, UK’s National Research Center for Online Harms. The authors also declare that they do not have any competing interests.

We also strive to minimise potential harm (Fiesler et al. 2016) and anticipate minimal risk: (i) we were careful to not overload the servers; (ii) the data collected is publicly available; (iii) the data is collected through public APIs; (iv) we anonymise the data upon collection; (v) we do not plan to make data publicly available; (vi) we ensure that no data is stored on the Perspective API servers (we choose the “doNotStore” option⁴).

7 Conclusions

This paper proposed and evaluated novel moderation solutions for decentralised, interoperable social networks. We believe this is particularly needed because interoperability is likely to increase thanks to regulation (DMA (EU 2023)) and the growing support for ActivityPub. We leverage the conversational context of each post with GraphNLI, a graph-based deep learning framework that makes random walks on the posts of the conversation. We evaluated our approaches using a comprehensive dataset of Pleroma (consisting of 713 instances and 16.5 million posts), a large decentralised micro-blogging service of the fediverse. We first

⁴https://support.perspectiveapi.com/s/about-the-api-faqs?language=en_US, last accessed 14 Sep 2023.

reconstructed the 2 million conversations with more than two posts that exist in our data. We found this to be more complex than expected as instances participating in the same conversation might only have a partial view of the entire conversation. Our conversation-based approach, succeeded in detecting toxicity in large instances (0.8837 macro-F1) but failed for small ones. We solve this problem via federation: we evaluate different strategies for sharing information across instances (toots or models) and achieve the highest macro-F1 of 0.8826 for model sharing strategy in small instances.

Acknowledgements

This work was supported by AP4L (EP/W032473/1), EU Horizon Framework grant agreement 101093006 (TaRDIS) and REPHRAIN’s “DSNMod” and “Fediobservatory” (EP/V011189/1).

References

- Agarwal, V.; Joglekar, S.; Young, A. P.; and Sastry, N. 2022. GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. In *Proceedings of the ACM Web Conference 2022*, 2729–2737.
- Agarwal, V.; Young, A. P.; Joglekar, S.; and Sastry, N. 2023. A graph-based context-aware model to understand online conversations. *ACM Transactions on the Web*, 18(1): 1–27.
- Almerekhi, H.; Jansen, B. J.; and Kwak, H. 2020. Investigating toxicity across multiple Reddit communities, users, and moderators. In *Companion proceedings of the web conference 2020*, 294–298.
- Anaobi, I. H.; Raman, A.; Castro, I.; Zia, H. B.; Ibosiola, D.; and Tyson, G. 2023. Will Admins Cope? Decentralized Moderation in the Fediverse. In *Proceedings of the ACM Web Conference 2023*, 3109–3120.
- Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW ’17 Companion, 759–760. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee. ISBN 9781450349147.
- Beck, M. D.; and Moore, T. R. 2023. Breaking Up a Digital Monopoly. *Communications of the ACM*, 66(6): 38–41.
- Bernstein, M.; Monroy-Hernández, A.; Harry, D.; André, P.; Panovich, K.; and Vargas, G. 2011. 4chan and/b: An Analysis of Anonymity and Ephemeral in a Large Online Community. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

- Bin Zia, H.; Raman, A.; Castro, I.; Hassan Anaobi, I.; De Cristofaro, E.; Sastry, N.; and Tyson, G. 2022. Toxicity in the decentralized web and the potential for model sharing. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(2): 1–25.
- Birman, I. 2018. Moderation in different communities on Reddit—A qualitative analysis study.
- Burnap, P.; and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2): 223–242.
- Cao, R.; Lee, R. K.-W.; and Hoang, T.-A. 2020. DeepHate: Hate Speech Detection via Multi-Faceted Text Representations. In *12th ACM Conference on Web Science*, WebSci ’20, 11–20. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379892.
- Chandrasekharan, E.; Pavalanathan, U.; Srinivasan, A.; Glynn, A.; Eisenstein, J.; and Gilbert, E. 2017. You can’t stay here: The efficacy of reddit’s 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–22.
- 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*.
- Doan, T. V.; van Rijswijk-Deij, R.; Hohlfeld, O.; and Bajpai, V. 2022. An empirical view on consolidation of the web. *ACM Transactions on Internet Technology (TOIT)*, 22(3): 1–30.
- EU. 2023. Q&A: DMA: Ensuring fair and open digital markets.
- Fiesler, C.; Wisniewski, P.; Pater, J.; and Andalibi, N. 2016. Exploring ethics and obligations for studying digital communities. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*, 457–460.
- Gehl, R. W.; and Zulli, D. 2022. The digital covenant: non-centralized platform governance on the mastodon social network. *Information, Communication & Society*, 1–17.
- Hassan, A. I.; Raman, A.; Castro, I.; Zia, H. B.; De Cristofaro, E.; Sastry, N.; and Tyson, G. 2021. Exploring content moderation in the decentralised web: The pleroma case. In *Proceedings of the 17th International Conference on emerging Networking EXperiments and Technologies*, 328–335.
- He, J.; Zia, H. B.; Castro, I.; Raman, A.; Sastry, N.; and Tyson, G. 2023. Flocking to mastodon: Tracking the great twitter migration. In *Proceedings of the 2023 ACM on Internet Measurement Conference*, 111–123.
- Iqbal, W.; Arshad, M. H.; Tyson, G.; and Castro, I. 2022. Exploring crowdsourced content moderation through lens of reddit during covid-19. In *Proceedings of the 17th Asian Internet Engineering Conference*, 26–35.
- Jigsaw. G. 2023. Perspective API. <https://www.perspectiveapi.com/>. Accessed: 2024-04-08.
- Kurita, K.; Belova, A.; and Anastasopoulos, A. 2019. Towards robust toxic content classification. *arXiv preprint arXiv:1912.06872*.
- La Cava, L.; Greco, S.; and Tagarelli, A. 2022a. Information consumption and boundary spanning in decentralized online social networks: the case of mastodon users. *Online Social Networks and Media*, 30: 100220.
- La Cava, L.; Greco, S.; and Tagarelli, A. 2022b. Network analysis of the information consumption-production dichotomy in mastodon user behaviors. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1378–1382.
- Luscombe, A.; Dick, K.; and Walby, K. 2022. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. *Quality & Quantity*, 56(3): 1023–1044.
- Mahy, R. 2023. More Instant Messaging Interoperability (MIMI) message content. Internet-Draft draft-ietf-mimi-content-00, Internet Engineering Task Force. Work in Progress.
- Mohan, S.; Guha, A.; Harris, M.; Popowich, F.; Schuster, A.; and Priebe, C. 2017. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, 51–56. Springer.
- Papasavva, A.; Blackburn, J.; Stringhini, G.; Zannettou, S.; and Cristofaro, E. D. 2021. “Is it a qoincidence?”: An exploratory study of QAnon on Voat. In *Proceedings of the Web Conference 2021*, 460–471.
- Papasavva, A.; Zannettou, S.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Raiders of the lost kek: 3.5 years of augmented 4chan posts from the politically incorrect board. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 885–894.
- Raman, A.; Joglekar, S.; Cristofaro, E. D.; Sastry, N.; and Tyson, G. 2019. Challenges in the decentralised web: The mastodon case. In *Proceedings of the internet measurement conference*, 217–229.
- Ribeiro, M. H.; Calais, P. H.; Santos, Y. A.; Almeida, V. A.; and Meira Jr, W. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.
- Risch, J.; and Krestel, R. 2020. Toxic comment detection in online discussions. *Deep learning-based approaches for sentiment analysis*, 85–109.
- Trautwein, D.; Raman, A.; Tyson, G.; Castro, I.; Scott, W.; Schubotz, M.; Gipp, B.; and Psaras, Y. 2022. Design and evaluation of IPFS: a storage layer for the decentralized web. In *Proceedings of the ACM SIGCOMM 2022 Conference*, SIGCOMM ’22, 739–752. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394208.
- Varian, H. R. 2019. Recent Trends in Concentration, Competition, and Entry. *Antitrust Law Journal*, 82(3): 807–834.
- W3C. 2018. ActivityPub W3C Recommendation. <https://www.w3.org/TR/activitypub/>. Accessed: 2024-04-08.
- Waseem, Z.; and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, 88–93.

Wu, Q.; Williams, L. K.; Simpson, E.; and Semaan, B. 2022. Conversations About Crime: Re-Enforcing and Fighting Against Platformed Racism on Reddit. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–38.

Yin, W.; Agarwal, V.; Jiang, A.; Zubiaga, A.; and Sastry, N. 2023. AnnoBERT: effectively representing multiple annotators’ label choices to improve hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, 902–913.

Zignani, M.; Quadri, C.; Galdeman, A.; Gaito, S.; and Rossi, G. P. 2019. Mastodon content warnings: Inappropriate contents in a microblogging platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, 639–645.

8 Paper Checklist

1. For most authors...

- (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? **Yes, we have discussed about the ethical considerations in Sections 4 and 6.**
- (b) Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? **Yes, we have described the main findings and contributions of this paper in the abstract and introduction.**
- (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? **Yes, we have discussed it in in Section 4.**
- (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? **Yes, we have discussed it in Section 4.**
- (e) Did you describe the limitations of your work? **Yes, we have described in Sections 6.3 and 7.**
- (f) Did you discuss any potential negative societal impacts of your work? **Yes, we have discussed in Section 6.**
- (g) Did you discuss any potential misuse of your work? **Yes, please check Section 6.**
- (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymisation, responsible release, access control, and the reproducibility of findings? **Yes, we have discussed in Section 4 for data and others in Section 6.**
- (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? **Yes**

2. Additionally, if your study involves hypotheses testing...

- (a) Did you clearly state the assumptions underlying all theoretical results? **NA**
- (b) Have you provided justifications for all theoretical results? **NA**

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? **NA**

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? **NA**

(e) Did you address potential biases or limitations in your theoretical framework? **NA**

(f) Have you related your theoretical results to the existing literature in social science? **NA**

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? **NA**

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? **NA**

(b) Did you include complete proofs of all theoretical results? **NA**

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **Yes, we have added all the relevant details for reproducibility or further research. We will make our code publicly available upon the paper’s acceptance. We do not intend to make the data available due to privacy reasons.**

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **Yes, in Section 6.**

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **Yes, we report the average results and standard deviation in Section 6.2.**

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **Yes in Section 6.1.**

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? **Yes, we have discussed this in Section 4.**

(f) Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? **Yes, we have discussed this in results in Section 6.**

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? **Yes, we have cited the paper proposing the dataset.**

(b) Did you mention the license of the assets? **NA**

(c) Did you include any new assets in the supplemental material or as a URL? **NA**

(d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? **Yes, we have discussed it in Section 6.**

- (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, we have discussed it in Section 4.
 - (f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see ?)? NA
 - (g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see ?)? NA
6. Additionally, if you used crowdsourcing or conducted research with human subjects...
- (a) Did you include the full text of instructions given to participants and screenshots? NA
 - (b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA
 - (d) Did you discuss how data is stored, shared, and de-identified? NA