Elisha et al. *EPJ Data Science* (2024) 13:70
https://doi.org/10.1140/epjds/s13688-024-00505-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

RESEARCH Open Access

Consumption-based approaches in proactive detection for content moderation

Shahar Elisha[1†], John N. Pougué-Biyong[2†] and Mariano Beguerisse-Díaz[1,2*]

*Correspondence:
marianob@spotify.com
[1]Spotify Ltd, London, UK
[2]Mathematical Institute, The University of Oxford, Oxford, UK
[†]Equal contributors

Abstract

Implementing effective content moderation systems at scale is an unavoidable and complex challenge facing technology platforms. Developing systems that automate detection and removal of violative content is fraught with performance, safety and fairness considerations that make their implementation challenging. In particular, content-based systems require large amounts of data to train, cannot be easily transferred between contexts, and are susceptible to data drift. For these reasons, platforms employ a wide range of content classification models and rely heavily on human moderation, which can be prohibitively expensive to implement at scale. To address some of these challenges, we developed a framework that relies on consumption patterns to find high-quality leads for human reviewers to assess. This framework leverages consumption networks, and ranks candidate items for review using two techniques: Mean Percentile Ranking (MPR), which we have developed, and an adaptation of Label Propagation (LP). We demonstrate the effectiveness of this approach to find violative material in production settings using professional reviewers, and on a publicly available dataset from MovieLens. We compare our results with a popular collaborative filtering (CF) baseline, and we show that our approach outperforms CF in production settings. Then, we explore how performance can improve using Active Learning techniques. The key advantage of our approach is that it does not require any content-based data; it is able to find both low- and high-consumption items, and is easily scalable and cost effective to run.

Keywords: Content moderation; Proactive detection; Consumption networks; Label propagation; Node ranking; Collaborative filtering

1 Introduction

Proactively detecting content in violation of an online platform's policies (e.g., violence, hate-speech, disinformation, and adult content) is a challenge that can pose serious consequences for personal safety, degrade user experience, and lead to negative societal impact [1]. Due to the large amounts of content that many platforms routinely process, relying solely on humans to find and review content is not viable. It is therefore necessary to design methods to proactively detect and review content in a way such that transparency, cost efficiency, and scalability are central considerations.

Springer

Some platforms have developed automated detection systems to supplement human reviewers [2]. Such systems usually combine metadata and content to identify specific policy violations [3–7]. Though powerful, content-based systems are limited in that they are not always transferable across languages or content types. For example, methods that work for audio may not work for video. Similarly, systems trained on hate-speech may not work on disinformation, or even on different forms of hate-speech. Other methods (e.g., based on deep learning or Large Language Models) can achieve more flexibility, but come with substantial increases in cost that can make them inaccessible for many practitioners, which then have to depend on (often costly) third party providers [8–10]. These systems require a large amount of training data to be effective, which in some moderation contexts is not possible due to the scale required, the legality of the content, and the psychological safety of the people collecting and processing the data. Furthermore, content-based methods need to be adapted and retrained in response to external events and to account for data drift, group specificity, regional characteristics, and policy updates [11–14]. For these reasons, content-based systems are costly to develop and maintain, which is why human reviewers remain the most reliable way to moderate content [14].

A complementary approach to content-based systems is to rely on patterns of audience engagement with the content; that is, using *consumption-based* methods. In the context of content moderation, there are methods designed for social networks that leverage rich social interaction data [15, 16]; however, these data (e.g., likes, shares and follows) are not always available in every moderation context. An alternative approach based on consumption data is to rely on the collaborative filtering (CF) assumption that audiences who have consumed known violative items are likely to have consumed other similar ones. Popular CF recommendation systems have been deployed with success in many applications [17]. One challenge with the CF approach is that the large amount of data that they require, as well as their popularity bias [18–20], make them an unsuitable solution for some content moderation tasks. For example, when there are relatively few instances of a specific type of content (e.g., when a new type of violation appears), out-of-the-box CF systems can struggle to find them. In addition, the mechanism of some CF systems, especially those that rely on deep learning, can be too complex to understand or interpret, which can be problematic in sensitive applications that require interpretability, such as content moderation [21].

In this work, we propose a content-agnostic, consumption-based framework specifically designed to prioritise and focus the review of potentially violative content on streaming platforms. We approach the problem as an item ranking task: given a set of known violative items, can we find more of them by analysing their consumption patterns? This methodology ranks candidates using two graph-based techniques: Mean Percentile Ranking (MPR), which we have developed, and an adaptation of Label Propagation (LP) [22]. By leveraging the wealth of user-item interactions available in online platforms, this framework allows us to flag items for review without processing their content (e.g. imagery, transcripts, or audio). Our methods can be applied to any content type, and require a relatively small amount of initial data to get started, in contrast to content-based methods. Analysing consumption is thus an efficient way to find and rank violative content at scale, ensuring that human reviewers examine leads with the most contextual evidence. Due to their simplicity, interpretability and effectiveness, the techniques and experiments

outlined in this work can provide a helpful tool and valuable insights for researchers and practitioners in the moderation space.

This paper is organised as follows: we present a succinct review of the content detection and moderation literature in *Related Work* (Sec. 2). In *Methods* (Sec. 3) we present the MPR and LP techniques for ranking items. In *Results* (Sec. 4) we showcase the performance of the methods in finding violative content in an internal dataset of podcast episodes at Spotify, and compare them to a CF baseline. To ensure that our results are reproducible, we also conduct an experiment to find horror films in a publicly available dataset. We explore Active Learning strategies and show when they can improve performance. We systematically explore the performance of the methods on varying structures of the consumption graph. Finally, in Sec. 5 we discuss the approaches and their performance, and outline interesting directions for future work.

2 Related work

Traditionally, humans have been in charge of reviewing content [14]. Technology platforms operate at a scale that has created the need to develop automated systems that prioritise content to be reviewed by a human workforce, or flag, down-rank, or even remove potentially violative content [23]. These solutions often leverage two types of data: the items' content and metadata (content-based: text, image, video, audio, and speech), and the user-item interactions (consumption-based: streams, likes, follows, retweets, and so on).

2.1 Content-based approaches

Among the range of content-based approaches, there is one major distinction between systems that match items to specific instances of content, and those that classify or predict categories to which the items belong [2]:

- *Matching systems* compare a previously unseen item against a database of annotated items. Their goal is to identify duplicates or near-duplicates of monitored instances. This approach often involves hashing (transforming a piece of content into a string of data meant to uniquely identify it [24, 25]) or fingerprinting (obtaining a set of descriptors that jointly identify the content [26]). For example, matching uploaded images to a database or watch-list. See Ref. [27] for an overview of the challenges of creating repositories of sensitive or dangerous content.
- *Classification systems* predict whether a new item has violative content or not, often using supervised machine learning. These systems typically follow a two-step process: (i) a classifier is trained on annotated data, and (ii) the trained classifier is used to annotate previously unseen data. There are content classification systems for moderation that process text [3, 28, 29], images [4, 5, 30], video [6, 31], audio, and multi-modal data [7].

Although content-based methods can achieve impressive performance, they have some limitations:

- *Specificity*: Classifiers and hashing methods are usually not transferable across content types and policies. For example, a classifier trained to detect explicit language or images will not be able to detect content that promotes self-harm, or even different types of explicit content.

- *Lack of context*: Language and images afford a range of interpretations across social groups, cultures and regions. For example, when working with hate-speech, the meaning of an utterance can vary depending on the political, regional, and temporal context in which it is used. In addition, dog-whistling, intent, sarcasm, and irony are difficult to detect automatically [11, 12, 32], and require a sophisticated understanding of the context, which can be challenging, even for human experts.
- *Feasibility*: An obstacle to practical feasibility of using classifiers and hash/fingerprint databases is the amount of data they require to be effective [13, 14]. As new sub-areas of violations constantly appear and evolve (e.g., COVID-19 disinformation emerging in 2020 [33]), labelled data collections are scarce. Moreover, retraining or fine-tuning classifiers (e.g., due to policy changes) is challenging and expensive.
- *Cost:* Sophisticated content based models, such as those based on Large Language Models (LLMs), can be prohibitively expensive to run at scale. Beyond the vast computational and data resources required to train LLMs [34], operating an LLM-based approach for moderation requires processing large amounts of text, which can be costly to procure: in a podcast catalogue with millions of long-form audio files we would need to first transcribe all audio to text, which comes at a very high cost. For example, the cost of transcribing 10 million 1-hour episodes is almost 2.5 million USD.[1]

Due to these limitations, platforms often use content-based methods in tandem with other approaches [16, 35].

2.2 Human content review

Review workforces consist of human content reviewers trained to recognise violations of a platform's policies; these reviewers typically examine content flagged by automated systems or user reports. Perhaps the biggest strength of manual review is that human annotators are better at capturing contextual nuances, and understanding out of distribution data than content-based automated systems [35]. It is relatively straightforward to update reviewer guidance than to retrain automated systems as policies evolve, or when a new type of violation appears [16], but it is costly and time-consuming to pass a high volume of content for human review. Detecting certain violations, such as disinformation, hate-speech, or violent content, may require special training and the establishment of psychological guardrails for reviewers, which can limit the rate at which teams can expand. In addition, repeated exposure to violent or explicit content can result in long-lasting psychological and emotional damage [35, 36]. Therefore, it is important that human review efforts are targeted to where they have the most impact; this points to an opportunity to supplement content-based systems with other approaches that improve performance, reliability, transferability and flexibility. Below, we argue that examining the consumption patterns of the items is one such promising approach.

2.3 Consumption-based approaches

Consumption-based approaches exploit the wealth of user-item interactions gathered by streaming platforms. An expanding body of work is dedicated to leveraging the consumption patterns of users to mitigate disinformation, spam and violative content in social

[1] https://cloud.google.com/speech-to-text/pricing.

networks. Some efforts have modelled the spread of rumours and disinformation as a contagion process on a social graph, and found that they can be distinguished from real news through their spreading patterns [37, 38]. Other works use hand-crafted features that are specific to a particular social network for classification purposes: retweets, forwards and replies in studies of Twitter [39], and Facebook friend requests [40]. Hybrid methods combine content-based features (e.g., text, hashtag, sentiment) with user characteristics [39, 41–44].

So far, many such consumption-based methods have been tailored to specific social networks and exploit social interaction data, which is not always available in other settings (e.g. online streaming platforms) [15]. In the absence of social interaction data, one option is to analyse consumption patterns. One advantage of focusing on consumption is that there exists a body of mature collaborative-based recommendation approaches from which we can draw inspiration [17, 45]. However, their implementation in a moderation context is not straightforward. First, the more complex the approach, the more difficult it is to supplement it with content-based information [46]. Moreover, it is often not possible to interpret their results, which is an obstacle to auditing the system's fairness. Additionally, CF systems are known to be subject to popularity bias: while popular items get plenty of exposure, less popular ones can be under-represented in results [18–20]. Yet, violative items are often in the long-tail of the catalogue's consumption [47]. Ensuring the integrity of the platform and the safety of the users requires finding harmful items, even if they have minimal consumption.

3 Methods

We present a simple, yet powerful approach to find high quality leads for review by analysing the consumption patterns of known instances of violative content, which we call *seeds*. The method has two parts: first, we construct a series of graphs with user-item interactions; second, we rank the nodes based on their *graph proximity* to the seeds. We investigate the use of two simple, transparent and extendable approaches to rank items according to how well connected they are to the seeds. Each method relies on a different notion of *graph proximity* to find items relevant to the seeds, even with low consumption: Mean Percentile Ranking (MPR) exploits the graph's local structure while Label Propagation (LP) uses the graph's global structure.

3.1 The data is the model: constructing graphs from seeds

The starting point of our method is a set of *seeds*: identified examples of the type of content we want to detect (e.g., items that violate a specific policy). We analyse the consumption of the seeds to find more instances of the violation. An important assumption in this framework is that the type of content we want to detect is consumed by users who actively engage with it. Some types of content may be consumed inadvertently by users (e.g., disinformation disguised as legitimate news, or in 'autoplay'); however, here we assume that most of the consumption of this content is deliberate.

We begin by constructing graphs that connect items to users. Figure 1 gives a high-level overview of our method. Let $\mathcal{S}$ be the set of seeds (Fig. 1A); the set of *seed users* that interacted with at least one seed is $\mathcal{U}_s$ (Fig. 1B). Then we find the rest of the items that the seed users also interacted with $\mathcal{I}_r$. We call this the set of *items to review* (Fig. 1C). We construct a bipartite graph by connecting $\mathcal{U}_s$ to $\mathcal{S} \cup \mathcal{I}_r$. The adjacency matrix of the graph

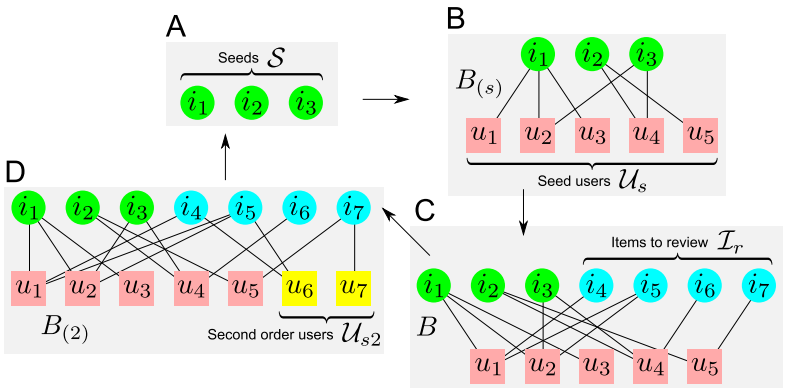

Figure 1 Overview of the consumption-based approach. **A**: The set of seed items is $\mathcal{S}$. **B**: We obtain the users that have interacted with any seed $\mathcal{U}_s$ to construct the bipartite graph $B_{(s)}$. **C**: The items to review $\mathcal{I}_r$ is the set of items that seed users interacted with that are not seeds; Together with the seeds and seed users they form the bipartite graph B. **D**: We can expand B into $B_{(2)}$ by adding all users $\mathcal{U}_{s2}$ that interacted with the items to review

is $B \in \mathbb{R}^{|\mathcal{U}_s| \times |\mathcal{S} \cup \mathcal{I}_r|}$ where $B_{ij} \neq 0$ if user u_i has interacted with item i_j, and zero otherwise. This graph may be unweighted (i.e., $B_{ij} \in \{0, 1\}$) or weighted (i.e., $B_{ij} \in \mathbb{R}^+$) if information exists about the level of affinity or extent of the interaction between the user and item such as the amount of time engaged or the rating given. We expand the graph by finding the set of users $\mathcal{U}_{s2}$ that interacted with the items in $\mathcal{I}_r$ but *not* with the seeds, that is, the *second order users* (Fig. 1D). We enlarge the bipartite graph by adding the second order users and their connections to the items to review so that the adjacency matrix is now $B_{(2)} \in \mathbb{R}^{|\mathcal{U}_s \cup \mathcal{U}_{s2}| \times |\mathcal{S} \cup \mathcal{I}_r|}$. Further expansions are possible, for example, by setting the seed set to be $\mathcal{S} \cup \mathcal{I}_r$, and expanding the set of items to review to construct larger graphs. In the remainder of this paper, we assume that the entries of $B_{(2)}$ are non-negative.

We approach the task of prioritising the items to review $\mathcal{I}_r$ as a node ranking problem. A substantial amount of work has been devoted to ranking nodes from a variety of points of view [48, 49]. Here, we propose two approaches: one using local information (i.e., using only a node's direct connections), and a global method using the whole structure of the graph.

In this framework, the combination of seeds *and* network analysis is the model. If we substitute one set of seeds for another (e.g., a set of hate-speech instances for a set of sexually explicit material), we effectively have a new model. This approach offers virtually unlimited models for any violation type, as long as we can compile a sufficient number of seeds, and the behaviour assumption holds.

3.2 Ranking methods

3.2.1 Mean percentile ranking

To rank item nodes using local node information, we develop a method based on the simple intuition that items *predominantly* consumed by seed users are more likely to be violative than items mostly consumed by second order users. Let $a_i^s = \sum_{u \in \mathcal{U}_s} B_{(2)_{ui}}$ be the accumulated interactions of seed users for a given item i (e.g., the sum of ratings), where $i \in \mathcal{I}_r$. Note that if the graph is unweighted, a_i^s is the number of seed users that interacted with i. Likewise, let $a_i^{s2} = \sum_{u \in \mathcal{U}_{s2}} B_{(2)_{ui}}$ be the accumulated interactions of second order users with $i \in \mathcal{I}_r$. In other words, a_i^s and a_i^{s2} are respectively the contributions of the seed

and second order users to the degree (or weighted degree) a_i of item i (i.e., $a_i = a_i^s + a_i^{s2}$). The accumulated interactions with seed users a_i^s and the proportion of the interactions with seed users $p_i = a_i^s/(a_i^s + a_i^{s2})$, contain complementary information about the nature of the consumption of i. By themselves, neither a_i^s nor p_i contain enough information to judge whether item i is likely to be a violation: a_i^s alone does not determine whether i's consumption is typical of users seeking content similar to the seeds and is skewed towards globally popular content. On the other hand, p_i is susceptible to noise because it is skewed towards low-consumption items. For example, if $a_i^s = 1$ and $a_i^{s2} = 0$, then $p_i = 1$ (its maximum value), even though there is scant information to conclude that i is violative. One solution is to aggregate a_i^s and p_i into a single score, but this must be done with care because they are not on the same scale.

To combine the information in a_i^s and p_i, we first consider their values for all items $i \in \mathcal{I}_r$. Given the sets $\{a_i^s : i \in \mathcal{I}_r\}$ and $\{p_i : i \in \mathcal{I}_r\}$, let $\hat{a}_i^s$ and $\hat{p}_i$ be the percentiles of a_i^s and p_i, respectively. The *mean percentile* (MP) of item i is

$$\mathrm{MP}(i;\gamma) = \gamma \hat{a}_i^s + (1-\gamma)\hat{p}_i, \tag{1}$$

where $\gamma \in [0,1]$ is a parameter that reflects how much weight we give to each quantity. The mean percentile combines how it ranks in both the distribution of seed users and proportion of seed users. If $\mathrm{MP}(i;\gamma)$ is high, then item i is consumed by several seed users who are also a large proportion of its consumers, so it is a good candidate for further investigation. Depending on the requirements of the moderation context, the value of γ can be tuned to give more weight to either $\hat{p}_i$, which emphasises the share of consumption by seed users, helping catch things early, or $\hat{a}_i^s$ which emphasises the actual consumption by seed users, to prioritise items that are getting more exposure. Finally, the *mean percentile ranking* (MPR) of $\mathcal{I}_r$ is:

$$\mathrm{MPR}(\mathcal{I}_r;\gamma) = \arg\mathrm{sort}\{\mathrm{MP}(i;\gamma) : i \in \mathcal{I}_r\}. \tag{2}$$

The MPR sorts the items in $\mathcal{I}_r$ by descending order of $\mathrm{MP}(i;\gamma)$. This is the order in which human content reviewers can evaluate the items for potential violations.

The main aspect of MPR is that it strikes a balance between the specificity of the consumption of an item by seed users (a_i^s) and the scale of its consumption (p_i), which, as we show below, tend to perform poorly by themselves. Another important aspect of MPR is that it uses percentiles instead of raw or normalised quantities, which removes scales and protects against extreme outliers.

3.2.2 Label propagation

Label propagation (LP) [22] is a semi-supervised learning method that propagates labels along the edges of a graph from labelled (in this case, the seeds) to unlabelled nodes. This method is based on a homophily premise: nodes that are tightly connected in a graph are likely to share a label. Although initially developed for unipartite graphs, LP has more recently been adapted to bipartite graphs [50]. Here we present a LP process on the graph B (Fig. 1C).

Intuitively, LP works by assigning unlabelled nodes a probability of having a label using a random walk on B. For this, we construct two row-stochastic matrices: $T^{I\to U}$ that encodes

the probability of transitioning from an item to a user, and $T^{\mathcal{U}\to I}$ for transitions from users to items:

$$T^{I\to\mathcal{U}}_{ui} = \frac{B_{ui}}{\sum_{j\in\mathcal{S}\cup\mathcal{I}_r} B_{uj}}, \tag{3}$$

$$T^{\mathcal{U}\to I}_{iu} = \frac{B_{ui}}{\sum_{v\in\mathcal{U}_s} B_{vi}}. \tag{4}$$

These matrices correspond to row-normalised versions of B and $B^\intercal$, respectively.

We assume that the first $|\mathcal{S}|$ columns of $T^{I\to\mathcal{U}}$ (rows of $T^{\mathcal{U}\to I}$) correspond to the seed items; in this case we can represent Eqs. (3)-(4) as

$$T^{I\to\mathcal{U}} = \begin{bmatrix} T^{\mathcal{S}\to\mathcal{U}_s} & T^{\mathcal{I}_r\to\mathcal{U}_s} \end{bmatrix}, \tag{5}$$

$$T^{\mathcal{U}\to I} = \begin{bmatrix} T^{\mathcal{U}_s\to\mathcal{S}} \\ T^{\mathcal{U}_s\to\mathcal{I}_r} \end{bmatrix}. \tag{6}$$

The blocks $T^{\mathcal{S}\to\mathcal{U}_s}$ and $T^{\mathcal{I}_r\to\mathcal{U}_s}$ contain the probabilities of transitioning from a labelled and unlabelled item to a user node, respectively. Likewise, the blocks $T^{\mathcal{U}_s\to\mathcal{S}}$ and $T^{\mathcal{U}_s\to\mathcal{I}_r}$ contain the probabilities of going from a user node to a labelled and unlabelled item, respectively. We define the label probability vector $C \in \mathbb{R}^{|\mathcal{S}|+|\mathcal{I}_r|+|\mathcal{U}_s|} = \begin{bmatrix} C^{\mathcal{S}} & C^{\mathcal{I}_r} & C^{\mathcal{U}_s} \end{bmatrix}^\intercal$ whose entries represent the probability that each node has the "violative" label. The first $|\mathcal{S}|$ entries correspond to labelled items (i.e., the seeds) whose values are clamped at 1. The rest of the items in C correspond to unlabelled items and users. Note that one can add more columns to C for multiple label scenarios. The interpretation for users is the amount of violative items that they interact with. The label information in C propagates from the seeds to the unlabelled items as follows:

$$C_{(k+1)} = \begin{bmatrix} C^{\mathcal{S}}_{(k+1)} \\ C^{\mathcal{I}_r}_{(k+1)} \\ C^{\mathcal{U}_s}_{(k+1)} \end{bmatrix} = \begin{bmatrix} 0 & 0 & T^{\mathcal{U}_s\to\mathcal{S}} \\ 0 & 0 & T^{\mathcal{U}_s\to\mathcal{I}_r} \\ T^{\mathcal{S}\to\mathcal{U}_s} & T^{\mathcal{I}_r\to\mathcal{U}_s} & 0 \end{bmatrix} \begin{bmatrix} C^{\mathcal{S}}_{(k)} \\ C^{\mathcal{I}_r}_{(k)} \\ C^{\mathcal{U}_s}_{(k)} \end{bmatrix}, \tag{7}$$

where $C_{(k)}$ is the state of the probability vector after k iterations. The blocks of zeros are there because nodes of the same type are not connected in the bipartite graph, so the label information only propagates through connections between items and users. Note that $C^{\mathcal{S}}_{(k)} = C^{\mathcal{S}} = 1 \in \mathbb{R}^{|\mathcal{S}|}\ \forall\, k$ because the seeds always have the "violative" label with probability 1. The evolution of the other two blocks is:

$$C^{\mathcal{U}_s}_{(k+1)} = T^{\mathcal{S}\to\mathcal{U}_s} C^{\mathcal{S}} + T^{\mathcal{I}_r\to\mathcal{U}_s} C^{\mathcal{I}_r}_{(k)}, \tag{8}$$

$$C^{\mathcal{I}_r}_{(k+1)} = T^{\mathcal{U}_s\to\mathcal{I}_r} C^{\mathcal{U}_s}_{(k+1)}. \tag{9}$$

The nth iteration of $C^{\mathcal{I}_r}_{(n)}$ can be written as:

$$C^{\mathcal{I}_r}_{(n)} = \left[\sum_{k=0}^{n-1} \left(T^{\mathcal{U}_s\to\mathcal{I}_r} T^{\mathcal{I}_r\to\mathcal{U}_s} \right)^k \right] T^{\mathcal{U}_s\to\mathcal{I}_r} T^{\mathcal{S}\to\mathcal{U}_s} C^{\mathcal{S}} + \left(T^{\mathcal{U}_s\to\mathcal{I}_r} T^{\mathcal{I}_r\to\mathcal{U}_s} \right)^n C^{\mathcal{I}_r}_{(0)}. \tag{10}$$

Because $T^{U\to I}$ and $T^{I\to U}$ are row-stochastic, one can show that (see Ref. [50]):

$$\lim_{n\to\infty}\left(T^{\mathcal{U}_s\to\mathcal{I}_r}T^{\mathcal{I}_r\to\mathcal{U}_s}\right)^n = 0. \tag{11}$$

As a consequence, the geometric series in the first term converges and the second term vanishes in Eq. (10). Therefore, the results of LP do not depend on the initial value of $C^{\mathcal{U}_s}$ and $C^{\mathcal{I}_r}$, and:

$$\lim_{n\to\infty} C^{\mathcal{I}_r}_{(n)} = \left(I - T^{\mathcal{U}_s\to\mathcal{I}_r}T^{\mathcal{I}_r\to\mathcal{U}_s}\right)^{-1} T^{\mathcal{U}_s\to\mathcal{I}_r}T^{\mathcal{S}\to\mathcal{U}_s}C^{\mathcal{S}}. \tag{12}$$

Finally, we rank the set of items to review $C^{\mathcal{I}_r}$ in decreasing order. It is possible to obtain an analogous expression for the seed user vector $C^{\mathcal{U}_s}$.

4 Results

We first evaluate our methods on Spotify's internal data using professional reviewers. While this is the intended use of the system, this data is not accessible to the wider research community. In consequence, to ensure that our results can be scrutinised and reproduced, we also perform evaluation on MovieLens [51], a publicly available dataset of film ratings, using genres as label annotations. As a baseline, we use the popular collaborative filtering method *Implicit* (IMP) [52] (see the Additional file 1 for a brief description). We run extensive experiments to explore the effects of parameter choices and different underlying data conditions, and we explore Active Learning (AL) strategies and in what setting they can help improve performance.

4.1 Detection of violative content in production settings

We study the consumption-based approach as a tool to help human reviewers prioritise review queues in an internal Spotify dataset in conditions that replicate a realistic production setting. In this setup, we start from a relatively small set of previously identified violative podcast episodes (i.e., the seeds) according to Spotify's Platform Rules.[2] We scan millions of items in Spotify's podcast catalogue to retrieve the seed users (i.e. users that streamed at least 5 minutes of a given seed episode) and the episodes to review (i.e. episodes streamed by the seed users for at least 5 minutes), construct the bipartite graphs as described in Fig. 1, and rank the episodes to review using MPR, LP, and IMP. Within the podcast catalogue, over 90% of the consumption comes from users selecting and clicking on an episode, or from shows that they may be subscribed to; therefore, we can assume that consumption is deliberate. Professional reviewers trained on the content policies evaluate the episodes on the ranked lists to decide which ones violate the content policies. Note that in this experiment, we only compute precision, as it is not feasible to compute recall because it would require reviewers to annotate a catalogue of millions of items.

The seed set $\mathcal{S}$ consists of 1000 podcast episodes violating one of our content policies. These seeds were consumed by 7700 users ($\mathcal{U}_s$) who also consumed 270,000 other podcast episodes ($\mathcal{I}_r$) over seven months. The bipartite graph B contains 271,000 episode nodes, 7700 user nodes, and approximately 1 million connections. We did not explicitly construct the graph $B_{(2)}$, which has over 2 billion edges, because we only require the degree of each

[2] https://www.spotify.com/us/safetyandprivacy/platform-rules.

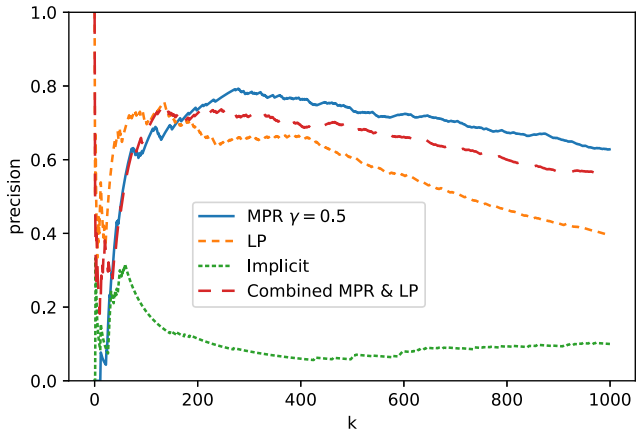

Figure 2 Performance in production settings. Detection of violative items in a production setting with professional reviewers. The curves show the precision at k of MPR with $\gamma = 0.5$, LP, and IMP, and a combination of MPR and LP on the top 1000 items from each method on the internal dataset

item node for MPR. We then rank the episodes in $\mathcal{I}_r$ using $\text{MPR}_{\gamma=0.5}$, LP, and IMP. Figure 2 shows the precision at k curves for the top-ranked 1000 episodes per method.

This experiment reveals the excellent performance of LP and MPR in realistic content moderation conditions, despite their simplicity. We assess performance using precision at $k \geq 10$ (when $k < 10$ there is a lot of volatility). The performance of MPR peaks at nearly 80% at $k \approx 300$; LP peaks at 75% at $k \approx 150$, while IMP peaks at only 30% for $k \approx 75$. The much lower performance by IMP suggests that general purpose collaborative filtering methods are not an appropriate approach for this setting. The higher performance of MPR, compared to LP, suggests that in the context of this data, local information bears more importance than longer paths in the consumption graph. Interestingly, the overlap between the candidates prioritised by MPR and LP is relatively low; only about 35% of the candidates in the top 1000 are shared between both methods, which suggests that the methods may be able to complement each other. We also test the combination of MPR and LP by re-ranking candidates using the minimum rank of the two methods, as seen in Fig. 2. The combined method does not outperform LP nor MPR.

4.2 Detecting horror movies from user ratings

We also evaluate our methods on a publicly-available collection of movie ratings from MovieLens [51]. Specifically, given a set of movie ratings (i.e., user-item interactions), we detect which films belong to the *Horror* genre using their consumption. For this example we focus on horror films because their consumption has some parallels with the consumption of violative content. Horror fans seek it because they find the experience of fright and intense emotions pleasurable, while people who are repelled by it will actively avoid consuming it [53]. In short, horror is not something people consume *repeatedly* by accident, it requires active engagement, which is the underpinning assumption of our consumption-based approach.

We use the MovieLens 25M dataset which contains 25 million ratings of 62,423 movies by 162,541 users between January 1995 and November 2019 [51]. In this data, each user is represented by an ID, an each movie is represented by its ID, title and genres. The data includes movies across 19 genres, a single movie may have one or more genres (see Fig. S1

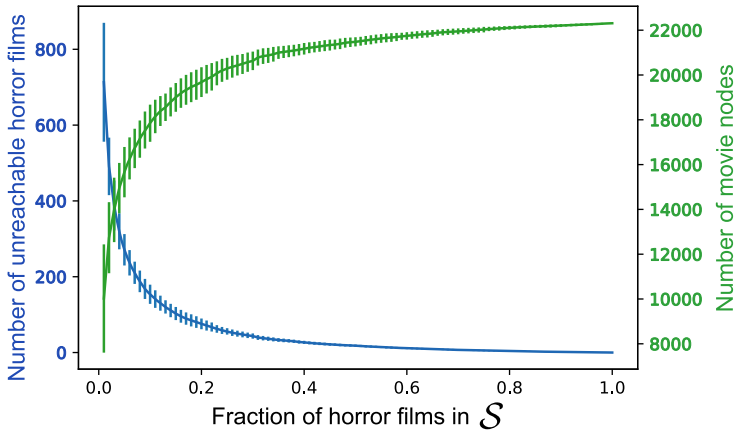

Figure 3 Bipartite graphs from MovieLens. Average number of movie nodes in the bipartite graph ± one standard deviation (green, right *y*-axis), and unreachable horror films (blue, left *y*-axis) after 30 random samples of the seed set $\mathcal{S}$

in Additional file 1). Rating entries contain of a score between 1 and 5, the user and movie IDs, as well as a timestamp. In our approach, we implicitly assume that a connection between a user and an item denotes a positive interaction (i.e., there are no negative edge weights); to ensure this assumption holds, we remove ratings below 4 stars. We also remove movies without a genre or with only one rating. We also remove ratings from prolific users who rated more than 200 movies with at least 4 stars. After this filtering, we remove all movies that were left with no ratings (i.e., movies that are not reachable in the consumption graph). The resulting dataset $\mathcal{D}$ contains 148,988 users, 23,211 movies, of which 2093 are horror movies, and 7,746,682 ratings.

The seed set of horror movies (i.e., $\mathcal{S}$) has a large impact on the graphs we construct, and, consequently, on the performance of the methods. To understand the effect of the size of $\mathcal{S}$, we vary its size from 1% to 99% of the 2093 horror films in $\mathcal{D}$. For each percentage, we sample 30 different seed sets uniformly at random, and compute the average performance of the methods. Figure 3 shows that as the percentage of horror films as seeds increases, the number of movie nodes in the bipartite graphs grows (in green, right *y*-axis), and the number of unreachable horror movies (i.e., horror films in $\mathcal{D}$ that are not in B nor $B_{(2)}$) decreases rapidly (in blue, left *y*-axis). For example, when the seed percentage is 20%, there are only about 100 unreachable horror movies.

We compute the performance of MPR, LP, and IMP for their top 1000 candidates, averaged over the 30 evaluations. Figure 4 shows that MPR's precision is close to 1.0 for up to $k = 300$ for seed sets with 20% of horror films or less. MPR's performance decreases rapidly as the seed set grows: as the seed fraction increases, the precision decreases for high values of k because there are fewer horror movies to find. This decrease in performance is also because as the seed set grows, a larger portion of the candidate films have a ratio of seed users p_i equal or close to 1. When p_i is high for too many candidates, it loses its discriminatory power (as does $\hat{p}_i$), so the accumulated interactions from seed users $\hat{a}_i^s$ (i.e. popularity) almost completely determines the order in which candidates are ranked. Figure 5 shows how $\hat{p}_i$ and $\hat{a}_i$ would perform on their own. Using $\hat{a}_i^s$ has a particularly poor performance. For this reason, recall is relatively stable: even though we find fewer horror films, there are fewer of them to be found.

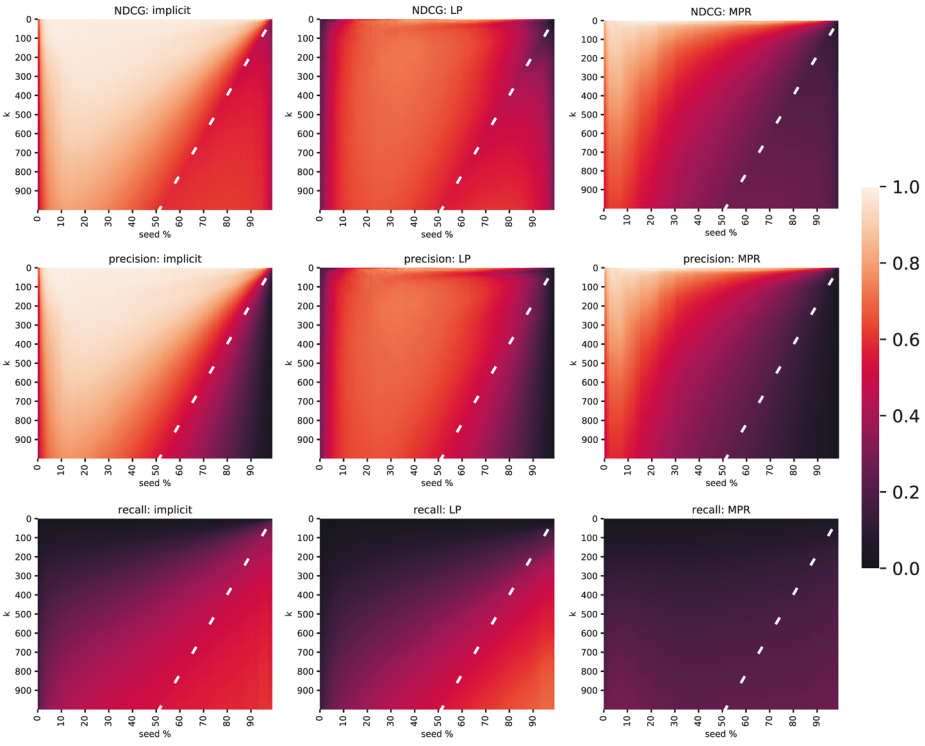

Figure 4 Performance on MovieLens dataset. Precision, recall and NDCG at k across all seed set sizes for IMP (left column), LP (middle column), and MPR (right column). The dashed lines indicate the number of horror films that remain to be found (i.e., not in the seed set). The white, dashed line in represents the amount of horror movies left to find with a given seed fraction (e.g., at 80% seed fraction, roughly 400 horror movies remain)

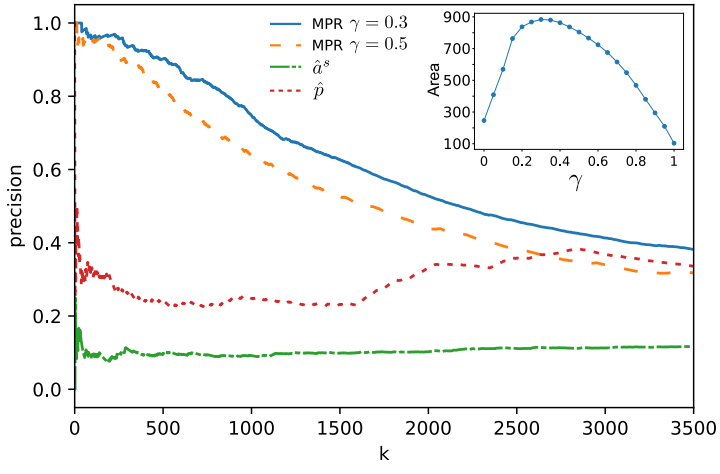

Figure 5 Performance on a 5% dataset. Precision of MPR on a 5% seed set with $\gamma = 0.5$ and 0.3, $\hat{a}^s$, and $\hat{p}$. Inset: Area under the precision curve for different γ

The precision in LP is stable around 0.6 for seed sets of up to 80% of horror films, near the area of the dashed line. This stability means that recall increases with the seed fraction (Fig. 4). IMP has a similar performance pattern to LP, but with a higher performance, ap-

proximately 20% more precision and Normalized Discounted Cumulative Gain (NDCG), and about 5% more recall. These results hint at the relative advantages of each method. For example, if we know (or believe) that the seed set contains a small fraction of the total number of violative items for a given content policy (i.e. if we know something about the prevalence of the violation), it may be better to employ MPR, but if the seed fraction is large, then LP or IMP may be a better choice.

In the rest of our analysis of the MovieLens data, we fix the seed set at 5% of horror films (105 films). From the 2093 horror movies in $\mathcal{D}$, we sample 105 uniformly at random to use as seeds, and we build the interaction graphs. We call this dataset $\mathcal{D}_{5\%}$; it contains 16,555 movies of which 1851 are horror, including the seeds. The movies have 7,731,468 ratings by 14,780 seed users and 148,987 second order users; of these ratings, 1,401,900 are from seed users alone. Because $\mathcal{D}$ contains 2093 horror movies and the bipartite graphs B and $B_{(2)}$ constructed from $\mathcal{D}_{5\%}$ only include 1851 of them, the highest possible recall with this seed set is $\frac{1851-105}{2093-105} = 0.88$.

We rank the 16,450 movies in $\mathcal{I}_r$ according to their likelihood of being a horror film using MPR, LP, and IMP. While MPR, LP, and IMP all place horror movies in the top ranks, MPR outperforms IMP and LP until they all converge around $k = 2500$ (see Fig. 6 and Table 1). We also compute the Normalized Discounted Cumulative Gain [54] (NDCG), which closely follows the precision trend before aligning with recall at the point where precision crosses recall (see Fig. 6). This inflection point in the NDCG curve also falls on the dashed line in Fig. 4. This behaviour is due to the normalisation of NDCG. MPR's NDCG score remains higher than LP and IMP's due to the more successful prioritisation

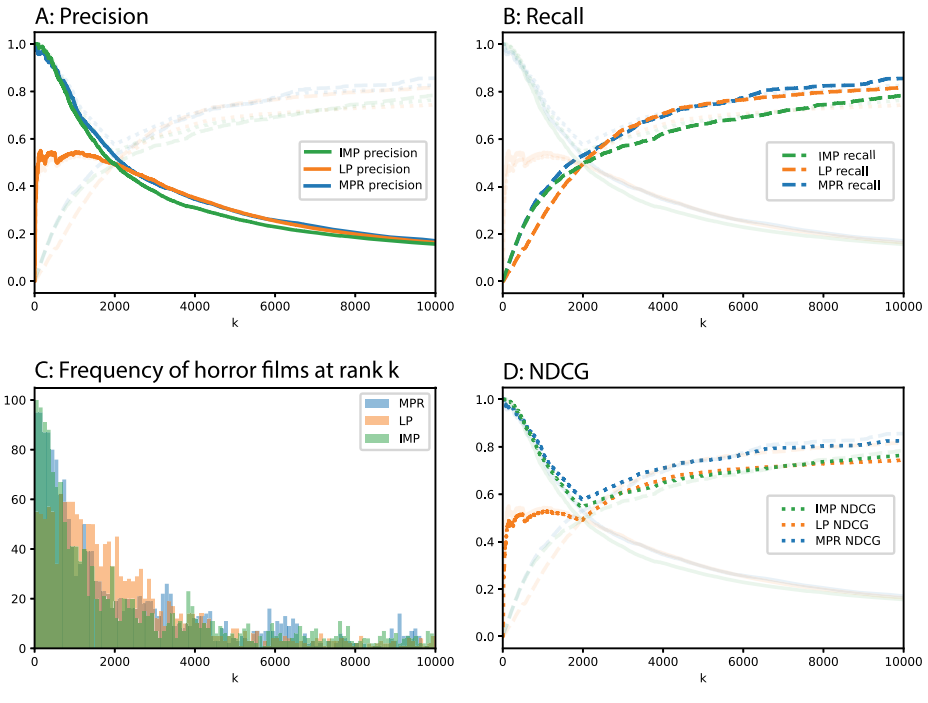

Figure 6 Performance of the methods on the MovieLens data. Line plots A: Precision (continuous line), B: Recall (dashed), D: NDCG (dotted) curves as a function of rank k for methods MPR (blue), LP (orange), and IMP (green). Bar plot C: Histogram of the number of true horror candidates in the rankings from the three methods. The size of each bin is 160

Table 1 Performance of $\text{MPR}_{\gamma=0.3}$, LP, and IMP on dataset $\mathcal{D}_{5\%}$

	Precision			*Recall*			*NDCG*		
	@100	@1000	@2500	@100	@1000	@2500	@100	@1000	@2500
MPR	96.0%	**74.8%**	**45.8%**	4.8%	**37.6%**	**57.6%**	96.8%	**77.9%**	**61.5%**
LP	52.0%	54.2%	45.6%	2.6%	27.3%	57.4%	44.7%	52.6%	55.8%
IMP	**100.0%**	71.4%	42.9%	**5.0%**	35.9%	53.9%	**100.0%**	75.2%	58.2%

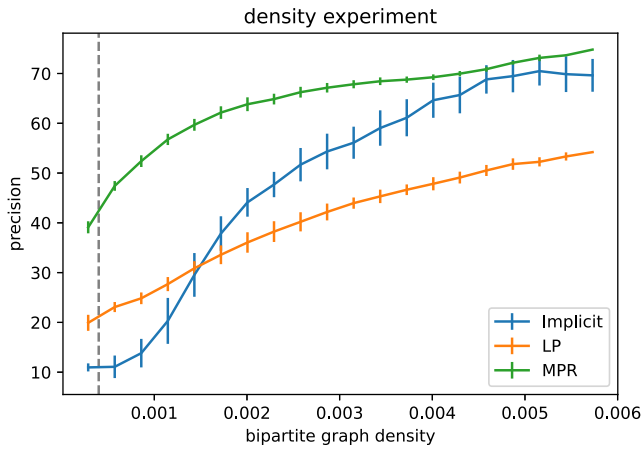

Figure 7 Performance with varying density. Precision@1000 of IMP, LP and MPR for different densities of the user-item bipartite network. For each density, we create 20 bipartite networks by randomly removing edges and evaluating the performance of the methods. The dashed vertical line indicates the density of the internal dataset from Fig. 2

in early ranks. All methods rank horror movies as low as 16,000, though these are movies that tend to have only one seed user and very few ratings overall.

While MPR and IMP have comparable performance on this dataset, IMP performed significantly worse on our internal experiment with realistic production setting data (see previous section and Fig. 2). This is because collaborative filtering methods usually require many instances of user-item interactions [55]. The MovieLens dataset is denser than our internal one, with densities of 6×10^{-3} and 4×10^{-4}, respectively. To examine the effect of density, we artificially decrease it in the MovieLens dataset by removing random edges from the bipartite graph. Figure 7 shows that as the density of the graph decreases, the performance for IMP rapidly deteriorates, and eventually performs worse than both MPR and LP. At the point at which the density is equivalent to the moderation dataset, IMP has a precision of around 10% at $k = 1000$, compared to MPR's 45%, and LP's 20%. This observation is consistent with our moderation experiments. While IMP has high performance on MovieLens, this performance depends on the underlying data conditions, which in this case means the density of the user-item network. This observation shows that it is vital to consider the settings and assumptions in each approach before choosing one.

Like in our results on our podcast data, both LP and MPR surface different candidates in the MovieLens dataset. In the top 1000 candidates, there are only 275 movies in common between LP and MPR, of which 267 are horror. There is an important difference in the popularity of the movies prioritised by each method. MPR initially retrieves movies with many more ratings than LP, as we show in Fig. S2 in the SOM. In the top 1000 candidates, MPR flags horror movies that have an average of 197 ratings per movie, compared to an

average of 35 ratings per horror movie flagged by LP. MPR tends to pick up popular movies with strong engagement from seed users (a_i^s) who dominate their consumption (high p_i). As we work our way down the ranking list, we find movies with a high seed user ratio but low popularity (e.g., low values at $k = 2500,\ 6000,$ and 9000 in Fig. S2), and more popular candidates (e.g. peaks at $k = 4000,\ 7500,$ and 10,000). In contrast, because LP uses edge weights expressed as probabilities for the next step in its random walk, the scale of their popularity has been factored out (see Eqs. (3) and (4)), which means the method does not have an automatic preference for popular items.

To better understand the contributions of the components of MPR in Eq. (1) and the effect of the parameter γ, we plot their precision in Fig. 5. Note that while individually $\hat{a}^{s_i}$ and $\hat{p}_i$ have poor performance finding horror films, their performance improves dramatically when we combine them in MPR. We also found that MPR works best when $\gamma = 0.3$ (i.e., when we give slightly more weight to $\hat{p}_i$ than to $\hat{a}_i^s$) on this dataset, with this number of seeds.

4.3 Active learning

Active learning (AL) is an approach in machine learning where a system receives a small amount of annotated data, produces results that are assessed by an agent, and then learns from these instances [56]. Our content moderation use case is a prime candidate to experiment with strategies in which human reviewers assess candidates, add new items to the seed set, and regenerate the candidate list. In principle, this cycle can be performed indefinitely; however, as our results below show, performance gains can eventually saturate. We evaluate two AL strategies to understand whether they can improve the performance of MPR, LP, and IMP:

- AL_{100}: Review the 100 candidates with the highest scores as normal. After reviewing the 100th candidate, add the newly identified items to $\mathcal{S}$ and regenerate the item ranking. Repeat these steps 10 times (excluding all reviewed items from the rankings each time). At the end, 1000 items have been reviewed.
- $AL_{50/50R}$: Review 100 items per iteration, but at each round we review the top 50 candidates and 50 candidates randomly sampled from the rest of top 1000 items. Add the positive items to $\mathcal{S}$ and regenerate rankings. Repeat 10 times excluding seen items from the rankings. At the end, 1000 items have been reviewed. This strategy can be useful when we are unsure about the coverage of our seed set, and some random exploration can help find items that are unreachable to the original seeds.

We compare these strategies to just seeding once and reviewing the top 1000 items (AL_{base}). Figure 8 shows the effect of the two AL strategies on MPR, LP, and IMP. We

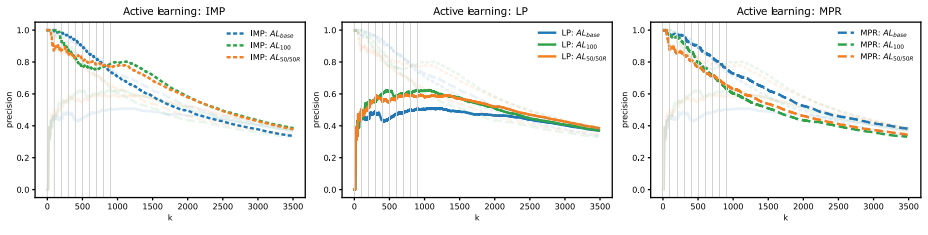

Figure 8 Active learning experiments. Active learning in MPR (dashed lines) and LP (smooth) using strategies AL_{100} and $AL_{50/50R}$ re-seeding every 100 items until $k = 900$, compared to the respective baselines AL_{base}

perform 10 seeding steps (i.e., the last re-seeding occurs at k = 900), and track the performance of the methods until k = 3500. Both AL strategies have a drastic negative effect on the performance of MPR. While this may seem surprising, such deterioration of performance explained by the decrease in precision as the seed set grows (Fig. 4). In contrast, AL noticeably improves LP's performance. The additional horror instances found in each iteration help enrich the structure of the graph that LP relies on. AL initially worsens IMP's performance, before precision increases again and outperforms the base from around k = 850 on. Due to IMP's stochastic nature, sometimes AL improves the performance across all values of k, but the average pattern across 30 runs is that AL performs worse at lower k before improving performance at higher values of k (see Fig. 8). A possible explanation for AL's initial decrease in performance is that the method explores the consumption network after finding all the candidates in a given region. The increase from AL then comes when the method finds a new region with more horror movies (see Fig. S3 in the SOM). There is little difference in the improvements from strategies AL_{100} and $AL_{50/50R}$. As k increases beyond 2500, all AL strategies converge to similar performance (this is 1600 items after the last seeding). MPR-AL strategies perform marginally worse than the LP and MPR baselines, while LP-AL perform marginally better.

To find whether re-seeding more frequently improves the performance of the method, we compare the precision curves of LP in the first 1000 items with 1 seeding (i.e., AL_{base}), 2 seedings (at k = 0 and 500), and so on until 10 seedings (i.e., AL_{100}). Figure 9 shows precision at k = 1000, which increases with every additional re-seeding. The gains in precision begin to saturate after each re-seeding cycle (see inset in Fig. 9). This result suggests that depending on the practicality of adding more AL iterations when using LP (e.g., regenerating the dataset, ranking, and so on), there is a clear benefit to re-seed.

4.4 Systematic evaluation on synthetic data

As we discuss in the preceding sections, the performance of MPR, LP and IMP depends on the statistical properties of the underlying consumption data. We investigate this depen-

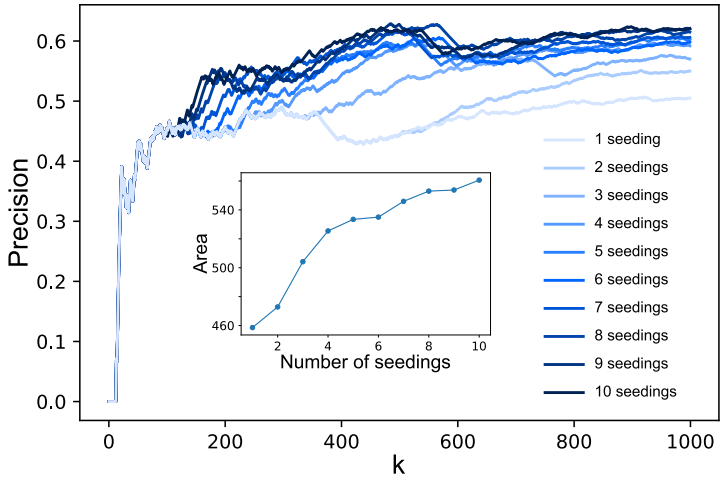

Figure 9 Active learning with multiple re-seedings. Effect of the number of re-seedings on the precision of LP on the first 1000 items. Inset: Area under the precision curve for each number of re-seedings

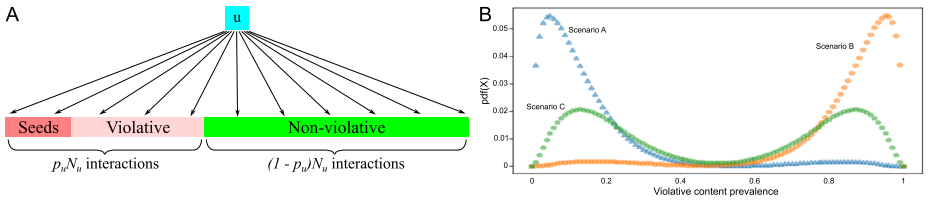

Figure 10 Sampling synthetic data. **A**: Sampling strategy of user-item interactions. The catalogue comprises violative and non-violative items. The *seeds* are a subset of the violative items that are known. A user u interacts with N_U items with the catalogue, $p_U N_U$ of which are violative, and the rest are non-violative. **B**: Different consumption and prevalence scenarios with 2-mode Beta distributions

dence by generating synthetic random consumption graphs $B_{(2)}$ sampling from a model that mimics a variety of user consumption patterns.

To generate random consumption graphs, suppose N^u is the number of items consumed by user u. This number can be an arbitrarily set parameter or computed from an existing dataset. The proportion of violative content consumed by user u is p_u, and the expected number of violative items consumed by each user is $p_u N^u$. The number of seeds is $|\mathcal{S}|$, and the number of items in the catalogue is $N = |\mathcal{I}_r + \mathcal{S}|$, which are all either known or can be arbitrarily determined. We also know the number of users a_i who have consumed an item i; however, the proportion p_u is typically unknown. We show here that these variables affect the performance of the three approaches.

We generate a randomised consumption graph $B_{(2)}$ as follows:

1. Generate a randomised graph B as we show on Fig. 10A, for each user u:
   a) Draw p_u from a 2-mode Beta distribution $f_{a,d}\beta(a, b = 10) + \beta(c = 10, d)$ (see Fig. 10B);
   b) Create edges between u and $\lfloor p_u N^u \rfloor$ violative items selected uniformly at random;
   c) Create edges between each user u and $\lfloor (1 - p_u)N^u \rfloor$ non-violative items selected uniformly at random.
2. We extend B to $B_{(2)}$ by using the known value a_i.

We explore three different consumption scenarios, as shown in Fig. 10B:

- Scenario A: most users consume a low proportion of violative items: $p_u \sim f_{a=.5,\, d=2}$.
- Scenario B: most users consume a high proportion of violative items $p_u \sim f_{a=2,\, d=.5}$.
- Scenario C: half of users consume a low proportion of violative items, and the other half a high proportion: $p_u \sim f_{a=1.5,\, d=1.5}$.

For each consumption scenario, we vary the number of violative items from almost none to N.

Figure 11 shows the performance of IMP, LP, $\text{MPR}_{\gamma=1}$, $\text{MPR}_{\gamma=.5}$ and $\text{MPR}_{\gamma=0}$, for each scenario. LP's performance has a positive correlation with the prevalence of violative items in each scenario. This method works best when most seed users do not interact with many violative items (Scenario A), and when there are many violative items to find. IMP is more suitable for settings in which most seed users interact with many violative items (Scenario B). The performance of MPR relies on the right choice of γ. When there are only a few violative items in the catalogue, the proportions of seed users ($\text{MPR}_{\gamma=0}$) works well when most users do not consume these violative items (Scenario A); the number of seed users ($\text{MPR}_{\gamma=1}$), in turn, is a reliable indicator when most users do consume violative items (Sce-

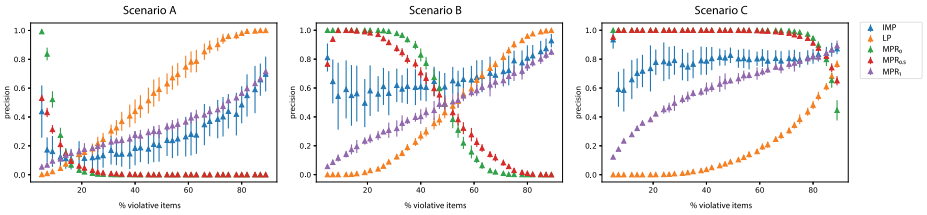

Figure 11 Synthetic experiment performance. Performance of IMP, LP, MPR$_{\gamma=1}$, MPR$_{\gamma=.5}$ and MPR$_{\gamma=0}$ in the three consumption scenarios as a function of the proportion of violative items in the catalogue. Vertical bars represent standard deviations across 30 runs

nario B). In the balanced Scenario C, MPR$_{\gamma=.5}$ and MPR$_{\gamma=0}$ are efficient methods if there are only a few violative items.

These results show, unsurprisingly, that there is no single method that works best in all scenarios; the performance of each varies depending on prevalence, and the consumption habits of the users. It is therefore important that practitioners invest in understanding their data and assumptions, so they can select the method that works best in their use-case.

5 Conclusions and future work

We introduced a consumption-based framework to detect violative content on technology platforms. The framework assumes that consumption from users who have engaged with known violative items can help us detect new ones. Based on this assumption, we build user-item graphs on which we apply Mean Percentile Ranking (MPR) and Label Propagation (LP), two effective node-ranking methods. These methods exploit complementary aspects of the data. On the one hand, MPR exploits the local structure of the consumption graph; on the other, LP uses the complete structure of the consumption graph to rank candidates.

We evaluated our method on an internal Spotify podcast dataset in a study aided by human reviewers. A comparison to a collaborative filtering-based baseline (*Implicit*), shows that the performance of MPR and LP methods is very strong (Fig. 2). In an evaluation to ensure the results are reproducible and enable scrutiny of our results, we applied the same methods to find horror films using the MovieLens dataset. The results of this evaluation are consistent in both studies: MPR achieves a precision of 0.75 at $k = 1000$ despite its simplicity, and is especially well-suited to find items with many interactions (Fig. 6). MPR's performance is the strongest in low-density scenarios (Fig. 7). LP is well suited to find items with few ratings (something that collaborative filtering recommendation systems are known to struggle with [19, 20]). We found that LP can benefit from AL strategies, which bring its precision at $k = 1000$ to 0.6 (Fig. 9). These two methods are complementary, and well suited to generate high quality leads for human reviewers to assess. Moreover, because their mechanisms are conceptually straightforward, these methods can be analysed and understood to help assess their fairness and performance, and easily maintained. One advantage of these simple methods is their cost effectiveness; whereas more sophisticated methods require vast quantities of data and complex infrastructure, and are expensive to train and run. Furthermore, consumption-based methods such as the ones we propose can generalise in a straightforward way to other settings where consumption graphs can be built and intentional consumption can be assumed. Our results and experiments with

synthetic consumption data show that the performance of the methods varies with the structure of the consumption graph, and the portion of violative items in the catalogue.

There are interesting future directions for this work. An enticing possibility is to combine this consumption-based framework with content-based techniques: for example, enriching our consumption graphs into multilayer networks [57] whose extra layers include content features (e.g., knowledge graphs that include descriptions, topics, speakers, language), or prompting LLMs with a combination of content and consumption representations. Moreover, given the complementarity of MPR and LP, understanding how to combine their results (e.g., using Dowdall-like rules [58]) and playing to their respective strengths is an appealing prospect. More work needs to be done to design optimal seed collection strategies and understand when these methods perform best (e.g., minimum data requirements, performance guarantees, conditions for detectability, finding optimal cutoff points in the candidate lists). Another interesting future direction is to expand the methods to detect multiple labels simultaneously (e.g., different types of violative content). For example, while LP can handle multiple labels separately, an interesting expansion would be to study multiple labels propagating in the presence of correlations among them.

Content moderation is a hugely challenging task that no single tool can solve. The complexity of the problem demands not only a broad palette of techniques, but also people from a variety of backgrounds and fields of expertise to work together to ensure that technology platforms are safe and enjoyable for everyone.

Abbreviations

AL, Active Learning; CF, Collaborative Filtering; IMP, Implicit; LLM, Large Language Model; LP, Label Propagation; MP, Mean Percentile; MPR, Mean Percentile Ranking; NDCG, Normalized Discounted Cumulative Gain; SOM, Supplementary Online Material.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1140/epjds/s13688-024-00505-x.

Additional file 1. (PDF 411 kB)

Acknowledgements

We thank Antonio Lima, Brian Regan, Dimitrios Korkinof, Henriette Cramer, Jean Garcia-Gathright, Linden Vongsathorn, Maria Dominguez, and Till Hoffmann for their useful feedback on several iterations of this manuscript. This work was done during John N. Pougué-Biyong's internship at Spotify.

Author contributions

MBD and SE conceived the study; SE, JNPB and MBD designed experiments; SE and JNPB performed analysis and computations; JNPB, SE and MBD interpreted results and wrote the manuscript. All authors read and approved the final manuscript.

Funding

Not applicable.

Data availability

The data analysed in *Detection of violative content in production settings* (Sec. 4.1) is proprietary and not available to the public. The data in *Detecting horror movies from user ratings* (Sec. 4.2) is publicly available and comes from the MovieLens 25M dataset [51] available from https://grouplens.org/datasets/movielens/25m/.

Declarations

Competing interests

MBD and SE are Spotify employees, and are named inventors on a patent application related to this research.

Received: 8 February 2024 Accepted: 23 October 2024 Published online: 15 November 2024

References

1. Grimmelmann J (2015) The virtues of moderation. Yale J Law Technol 17:42
2. Gorwa R, Binns R, Katzenbach C (2020) Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data Soc 7(1)
3. Chakrabarty T, Gupta K, Muresan S (2019) Pay "attention" to your context when classifying abusive language. In: Proceedings of the third workshop on abusive language online. ACM, Florence, pp 70–79
4. Ap-apid R (2005) An Algorithm for Nudity Detection
5. Zhelonkin D, Karpov N (2020) Training effective model for real-time detection of nsfw photos and drawings. In: Analysis of images, social networks and texts. Springer, Cham, pp 301–312
6. Hanson A, Pnvr K, Krishnagopal S, Davis L (2019) Bidirectional Convolutional LSTM for the Detection of Violence in Videos: Subvolume B, pp 280–295
7. Schroepfer M (2019) Facebook community standards report. Technical announcement, Facebook. https://ai.facebook.com/blog/community-standards-report/. Accessed 19 November 2020
8. Wang H, Hee MS, Awal MR, Choo KTW, Lee RK-W (2023) Evaluating gpt-3 generated explanations for hateful content moderation. In: International joint conference on artificial intelligence
9. Caselli T, Basile V, Mitrović J, Granitzer M (2021) HateBERT: retraining BERT for abusive language detection in English. In: Proceedings of the 5th workshop on online abuse and harms (WOAH 2021). Association for Computational Linguistics, Online, pp 17–25. https://doi.org/10.18653/v1/2021.woah-1.3. https://aclanthology.org/2021.woah-1.3
10. Tan F, Hu Y, Hu C, Li K, Yen K (2020) TNT: text normalization based pre-training of transformers for content moderation. In: Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Online, pp 4735–4741. https://doi.org/10.18653/v1/2020.emnlp-main.383. https://aclanthology.org/2020.emnlp-main. 383
11. Vidgen B, Harris A, Nguyen D, Tromble R, Hale S, Margetts H (2019) Challenges and frontiers in abusive content detection. In: Proceedings of the third workshop on abusive language online. ACM, Florence, pp 80–93
12. Duarte N, Llanso E, Loup A (2018) Mixed messages? The limits of automated social media content analysis. In: Proceedings of the 1st conference on fairness, accountability and transparency. Proceedings of Machine Learning Research, New York, pp 106–106
13. Singh S (2019) The limitations of automated tools in content moderation. Policy report. New America's Open Technology Institute. Accessed 15 October 2020
14. Gillespie T (2020) Content moderation, ai, and the question of scale. Big Data Soc 7(2)
15. Halevy A (2020) Preserving integrity in online social media. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. KDD '20. Association for Computing Machinery, New York, p 3601
16. Halevy A, Ferrer CC, Ma H, Ozertem U, Pantel P, Saeidi M, Silvestri F, Stoyanov V (2020) Preserving integrity in online social networks. arXiv:2009.10311
17. Lü L, Medo M, Yeung CH, Zhang Y-C, Zhang Z-K, Zhou T (2012) Recommender systems. Phys Rep 519(1):1–49
18. Brynjolfsson E, Hu Y, Smith M (2006) From niches to riches: the anatomy of the long tail. MIT Sloan Manag Rev 47
19. Abdollahpouri H (2019) Popularity bias in ranking and recommendation. In: Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society. AIES '19. Association for Computing Machinery, New York, pp 529–530
20. Abdollahpouri H, Mansoury M, Burke R, Mobasher B (2019) The unfairness of popularity bias in recommendation. CoRR. arXiv:1907.13286
21. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215
22. Zhu X, Ghahramani Z (2002) Learning from labeled and unlabeled data with label propagation
23. Seering J, Wang T, Yoon J, Kaufman G (2019) Moderator engagement and community development in the age of algorithms. New Media Soc 21(7)
24. Niu X-M, Jiao Y-H (2008) An overview of perceptual hashing. Acta Electron Sin 36(7)
25. Davis A, Rosen G (2019) Open-sourcing photo- and video-matching technology to make the Internet safer. Technical announcement, Facebook. https://about.fb.com/news/2019/08/open-source-photo-video-matching/. Accessed 15 October 2020
26. Wang AL-C (2003) An industrial-strength audio search algorithm. In: Proceedings of the 4th international conference on music information retrieval
27. Solorio T, Shafaei M, Smailis C, Diab M, Giannakopoulos T, Ji H, Liu Y, Mihalcea R, Muresan S, Kakadiaris I (2021) White paper: Challenges and considerations for the creation of a large labelled repository of online videos with questionable content. arXiv:2101.10894
28. Comandini G, Patti V (2019) An impossible dialogue! Nominal utterances and populist rhetoric in an Italian Twitter corpus of hate speech against immigrants. In: Proceedings of the third workshop on abusive language online. ACM, Florence, pp 163–171
29. Zhang X, Cao J, Li X, Sheng Q, Zhong L, Shu K (2020) Mining Dual Emotion for Fake News Detection. arXiv:1903.01728
30. Basilio JAM, Torres GA, Pérez GS, Medina LKT, Meana HMP (2011) Explicit image detection using ycbcr space color model as skin detection. In: Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on computer engineering and applications. American-MATH'11/CEA'11. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, pp 123–128
31. Papadamou K, Papasavva A, Zannettou S, Blackburn J, Kourtellis N, Leontiadis I, Stringhini G, Sirivianos M (2020) Disturbed youtube for kids: characterizing and detecting inappropriate videos targeting young children. In: Proceedings of the international AAAI conference on web and social media, vol 14
32. Pougué-Biyong J, Semenova V, Matton A, Han R, Kim A, Lambiotte R, Farmer D (2021) Debagreement: a comment-reply dataset for (dis) agreement detection in online debates. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2)
33. Sharma K, Ferrara E, Liu Y (2020) Identifying coordinated accounts in disinformation campaigns. arXiv:2008.11308
34. Hoffmann J, Borgeaud S, Mensch A, Buchatskaya E, Cai T, Rutherford E, de Las Casas D, Hendricks LA, Welbl J, Clark A, Hennigan T, Noland E, Millican K, van den Driessche G, Damoc B, Guy A, Osindero S, Simonyan K, Elsen E, Vinyals O, Rae J, Sifre L (2022) An empirical analysis of compute-optimal large language model training. In: Koyejo S, Mohamed

S, Agarwal A, Belgrave D, Cho K, Oh A (eds) Advances in neural information processing systems, vol 36. Curran Associates, Inc., Red Hook, pp 30016–30030. https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf
35. Das A, Dang B, Lease M (2020) Fast, accurate, and healthier: interactive blurring helps moderators reduce exposure to harmful content. In: Proceedings of the AAAI conference on human computation and crowdsourcing, vol 8, pp 33–42
36. Working with Traumatic Imagery. The Dart Center. https://dartcenter.org/content/working-with-traumatic-imagery
37. Friggeri A, Adamic L, Eckles D, Cheng J (2014) Rumor cascades
38. Jin F, Dougherty E, Saraf P, Cao Y, Ramakrishnan N (2014) Epidemiological modeling of news and rumors on Twitter. In: Proceedings of the 7th workshop on social network mining and analysis. ACM, New York
39. Yang S, Shu K, Wang S, Gu R, Wu F, Liu H (2019) Unsupervised fake news detection on social media: a generative approach. In: Proceedings of the AAAI conference on artificial intelligence 33(01), pp 5644–5651
40. Noorshams N, Verma S, Hofleitner A (2020) Ties: temporal interaction embeddings for enhancing social media integrity at Facebook. Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining
41. Shu K, Zhou X, Wang S, Zafarani R, Liu H (2019) The role of user profiles for fake news detection. In: Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining. Association for Computing Machinery, New York, pp 436–439
42. Ruchansky N, Seo S, Liu Y (2017) Csi: a hybrid deep model for fake news detection. In: Proceedings of the 2017 ACM on conference on information and knowledge management. Association for Computing Machinery, New York, pp 797–806
43. Gangireddy SCR, Deepak P Long C Chakraborty T (2020) Unsupervised fake news detection: a graph-based approach. In: Proceedings of the 31st ACM conference on hypertext and social media. HT '20. Association for Computing Machinery, New York, pp 75–83
44. Hanu L, Thewlis J, Asano YM, Rupprecht C (2022) Vtc: improving video-text retrieval with user comments. In: Computer vision–ECCV 2022: 17th European conference, Tel Aviv, Tel Aviv, Israel, October 23–27, 2022. Proceedings, part XXXV, Springer, Berlin, pp 616–633
45. Suganeshwari G, Syed Ibrahim SP (2016) A survey on collaborative filtering based recommendation system. In: Vijayakumar V, Neelanarayanan V (eds) Proceedings of the 3rd international symposium on big data and cloud computing challenges (ISBCC – 16'). Springer, Cham, pp 503–518
46. Zhang S, Yao L, Sun A, Tay Y (2019) Deep learning based recommender system: a survey and new perspectives. ACM Comput Surv 52(1)
47. Zhang Z, Luo L (2019) Hate speech detection: a solved problem? The challenging case of long tail on Twitter. Semant Web 10(5):925–945
48. Newman MEJ (2018) Networks: an introduction, 2nd edn. Oxford University Press, Oxford
49. Rodrigues FA (2019) In: Macau EEN (ed) Network centrality: an introduction. Springer, Cham, pp 177–196
50. Yu X, Chakraborty S, Brady E (2019) A co-training model with label propagation on a bipartite graph to identify online users with disabilities. In: Proceedings of the international AAAI conference on web and social media, vol 13, pp 667–670
51. Harper FM, Konstan JA (2015) The movielens datasets: history and context. ACM Trans Interact Intell Syst 5(4)
52. Hu Y, Koren Y, Volinsky C (2008) Collaborative filtering for implicit feedback datasets. In: 2008 eighth IEEE international conference on data mining. IEEE Press, New York, pp 263–272
53. Sparks GG, Spirek MM, Hodgson K (1993) Individual differences in arousability: implications for understanding immediate and lingering emotional reactions to frightening mass media. Commun Q 41(4):465–476
54. Schütze H, Manning CD, Raghavan P (2008) Introduction to information retrieval, vol 39. Cambridge University Press, Cambridge
55. Koren Y, Rendle S, Bell R (2022) Advances in collaborative filtering. Recommender systems handbook, 91–142
56. Settles B (2009) Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison
57. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer networks. J Complex Netw 2(3):203–271
58. Fraenkel J, Grofman B (2014) The Borda count and its real-world alternatives: comparing scoring rules in nauru and Slovenia. Aust J Polit Sci 49(2):186–205

Publisher's Note