Beyond Trial-and-Error: Predicting User Abandonment After a Moderation Intervention

Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, Stefano Cresci

***Abstract*—Current content moderation practices follow the *trial-and-error* approach, meaning that moderators apply sequences of interventions until they obtain the desired outcome. However, being able to preemptively estimate the effects of an intervention would allow moderators the unprecedented opportunity to plan their actions ahead of application. As a first step towards this goal, here we propose and tackle the novel task of predicting the effect of a moderation intervention. We study the reactions of 16,540 users to a massive ban of online communities on Reddit, training a set of binary classifiers to identify those users who would abandon the platform after the intervention—a problem of great practical relevance. We leverage a dataset of 13.8M posts to compute a large and diverse set of 142 features, which convey information about the activity, toxicity, relations, and writing style of the users. We obtain promising results, with the best-performing model achieving *micro F1* $= 0.800$ and *macro F1* $= 0.676$. Our model demonstrates robust generalizability when applied to users from previously unseen communities. Furthermore, we identify activity features as the most informative predictors, followed by relational and toxicity features, while writing style features exhibit limited utility. Our results demonstrate the feasibility of predicting the effects of a moderation intervention, paving the way for a new research direction in predictive content moderation aimed at empowering moderators with intelligent tools to plan ahead their actions.**

***Index Terms*—Content moderation, predictive moderation, user abandonment, machine learning.**

I. Introduction

Online social media has brought about the sharing of opinions and information of all sorts, locally and across the globe [1]. However, this ease of communication has also caused multiple problems, including the spread of toxic content, which is not only capable of causing harm to those users targeted by such misbehavior, but also to the overall health and inclusivity of the online environments [2]. Online platforms are thus required to enforce content moderation actions to mitigate this and other issues [3]. Moderation strategies can be broadly divided into two main categories: hard and soft. The former consists of permanently removing users, groups, or posts that are violating community guidelines, such as posts deemed toxic. This type of intervention —also called *deplatforming*— can have unintended consequences because certain users perceive content removals as a form of censorship, especially if unmotivated. As a consequence, they may feel less prone to posting again, or they could even decide to migrate to other platforms in which aggressive behavior is tolerated [4]–[6]. This led to the proposal of the second broad form of moderation —soft moderation— which warns users about the harmfulness of a content without removing it [7], [8]. An example of a soft moderation intervention is Reddit's quarantine, which affects the visibility of a whole subreddit (i.e., a community within Reddit) by limiting access to it and by warning new users about the potential harmful content therein [9].

The growing consensus among the thriving content moderation literature is that gradual approaches that initially apply less punitive interventions, and that subsequently escalate the severity whenever required, are generally more effective than other solutions [10]. Nevertheless, hard moderation is still the most widely used form of content moderation. A well-known example is the so-called *Great Ban*, a massive deplatforming intervention enforced by Reddit in response to the rise of toxic and hateful speech on the platform, which involved the permanent ban of around 2,000 subreddits in June 2020 [11]. Among the banned subreddits was `r/The_Donald`, a community of Trump supporters that was one of the most popular political subreddits at the time [12]. Before the ban the community had already faced two milder moderation interventions: a *quarantine* and a *restriction* that limited its visibility and the users that could serve as moderators of the community [13]. This sequence of interventions has been the subject of several studies that assessed changes in activity levels, use of language, political bias, factual news sharing, and toxicity after the interventions [13]–[17]. Interestingly, these studies revealed varying user reactions to the interventions. While some users exhibited a decrease in toxicity, others increased it, and in some instances very noticeably [14]. Many users were not much affected by the interventions, showing no significant change in behavior afterward. Others, however, seemed to be resentful of the platform and significantly decreased their activity, so much so that some completely abandoned the platform [11].

The previous results highlight the complexity of content moderation. First, in content moderation *one size does not fit all*, and effective moderation interventions should necessarily be tailored to the targets of the moderation [18]. Furthermore, the current moderation strategy follows a *trial-and-error* approach, where platforms apply sequences of interventions until they meet the desired outcome, or force the misbehaving users out of the platform. `r/The_Donald` was a prime example of this strategy, as it was subject to two unsuccessful interventions prior to the Great Ban. Given that the first of such interventions was applied in June 2019, while the last

B. Tessa, L. Cima, A. Trujillo, and S. Cresci are with the Institute of Informatics and Telematics (IIT) of the National Research Council (CNR), Pisa, Italy

L. Cima and M. Avvenuti are with the University of Pisa, Department of Information Engineering, Pisa, Italy

in June 2020, it is evident that it took several months for the platform to evaluate the effects of each intervention and to impose further restrictions. Importantly, during that time the misbehaving users continued to spread toxic content and to engage in harmful behaviors.

There is thus a pressing need for social media platforms to more efficiently evaluate the impact of a potential moderation intervention *before* its application. This would open up the possibility to plan interventions ahead of time, rather than to assess and correct afterwards. To date, however, no evidence exists as to whether the effects of a moderation intervention are predictable.

A. Contributions

Motivated by the need for a strategic planning and deployment of moderation actions, we tackle the new task of predicting the effect of a moderation intervention. While many studies carried out post-hoc descriptive analyses of the effects of moderation interventions [8], [15], [16], here for the first time we adopt a predictive approach. We leverage the results of previous studies on the Great Ban as our ground-truth for the effects of that intervention [11], [19]. Then, we analyze user behavioral traces before the Great Ban to predict possible changes in the behavior of those users after the ban. We specifically focus on predicting those users who abandon the platform after the ban, in a binary classification task. This choice is both theoretically and practically relevant. Changes in user activity are among the most widely studied effects of past moderation interventions [9], [13], [19]. As such, from the theoretical standpoint it is interesting to assess the extent to which those effects are predictable. Additionally, the task is also practically relevant since the loss of users negatively affects platform revenues, as user engagement and popularity are paramount in the economic success of social media [20], [21]. Moreover, the abandonment of influential users can also trigger chain reactions, as their followers may also choose to leave the platform, further worsening the issue. The main results of this work are summarized in the following:

- We evaluated different classification algorithms, with different sets of features and hyper-parameters, for the new task of predicting the effects of the Great Ban. Our best classifier achieved a promising *micro F1* $= 0.800$.
- We employed four different classes of features: those based on user (*i*) activity, (*ii*) toxicity, (*iii*) writing style, and (*iv*) relational characteristics. Activity features provided the most information to the classifiers, along with relational and toxicity ones. Instead, writing style features proved to be less informative, despite being widely used.
- We assessed the impact of user activity on the classification performance. We found that users exhibiting low activity are accurately classified, unlike users with high activity for which we obtained worse results. This finding suggests that more features should be designed to better characterize the online behavior of the most active users.

B. Significance

This work takes the first step towards providing strategic evidence and tools to support online moderators in planning their interventions. Our promising —yet preliminary— results validate the adoption of predictive models for assessing the outcomes of moderation interventions. This new approach marks a significant advancement in content moderation, offering moderators and platform administrators the ability to strategically deploy interventions to minimize adverse effects such as user attrition to less regulated platforms, while concurrently maximizing desired outcomes such as the reduction of toxic content. By enabling proactive decision-making based on predicted intervention effects, this approach stands to enhance the efficacy and efficiency of content moderation in fostering healthier online environments.

II. Related Works

While many works in the literature have already studied the effects of moderation interventions from a descriptive point of view, fewer have covered predictive modeling. In this section, we summarize and discuss previous works from both perspectives.

A. Descriptive moderation

In discussing the many works that took a descriptive approach to analyzing moderation interventions, we first survey studies on the Great Ban and other hard interventions, because they are the most similar to our present work. Among the works that studied the Great Ban is [19], which described the changes in activity and in-group vocabulary of those users who participated in the 15 most popular subreddits out of the 2,000 shut with the ban. The results highlighted the heterogeneity of the effects of the intervention and that top users tended to reduce their activity and their use of in-group vocabulary. The work in [11] carried out a similar analysis but focused on evaluating changes in toxicity rather than activity. Nonetheless the authors found that 15.6% of the affected users abandoned Reddit after the ban. Among those who stayed there was a general reduction in toxic comments, but a small subset of users drastically increased their toxicity. Other works that evaluated the effects of hard content moderation include [16], [22]. The former measured changes occurred in Google Trends and on Wikipedia, while the latter on Twitter. Both works revealed that banning toxic influencers reduced the attention toward them and the number of posts citing them. These studies concluded that deplatforming may be a valid moderation strategy, when used appropriately.

Due to its prominence on Reddit and the multiple interventions that it faced, `r/The_Donald` was the focus of many works that investigated the effectiveness of online content moderation. For example, some works showed how the quarantine was able to reduce the visibility of the subreddit. At the same time however, it was unsuccessful at reducing racist and misogynistic language [9], [15]. Others examined the interventions in terms of activity, toxicity, factual reporting, and political bias both at community and user levels [13], [14]. Community-level results found a general reduction in activity, a strong long-term increase in toxicity, a slight decrease in factual reporting, and no particular change in political ideology [13]. However, user-level reactions were more diversified

and sometimes even extreme, especially for the most active users [14]. These results highlight that community-level effects are not always representative of the underlying user-level effects, which once again reaffirms the limitations of one-size-fits-all moderation [18]. On the contrary, studying, and possibly even predicting user-level reactions to moderation interventions could be particularly beneficial to moderators.

B. Predictive moderation

Recent computational efforts in the study of moderation interventions increasingly adopt predictive rather than descriptive approaches. Among this body of work, the vast majority of studies tackled the task of predicting which pieces of content would be subject to moderation. For example, the authors of [23] predicted with high accuracy which YouTube videos would be later moderated. The predictions were accurate even at posting-time, reducing the disappointment users face upon deletion after publication. Similarly, [24] proposed LAMBRETTA, a learning to rank system developed to identify tweets likely to be removed from the platform because they convey false information. The system is intended to provide support to moderators in preventing the spread of fake news. Along the same line, CROSSMOD [25] automatically detects which Reddit posts are likely to be removed by the moderators. The system is also capable of taking actions based on certain conditions set by the moderators themselves. The aforementioned works serve a twofold goal. On the one hand, they contribute to reverse engineer the content moderation actions taken by large online platforms, which often lack transparency and consistency [3]. On the other hand, the automated systems developed as part of the studies could also support human moderators by reducing the burden of extensive manual analyses as well as by reducing the emotional toll that comes from being continuously exposed to toxic, harmful, or otherwise inappropriate content [26].

Other related works focused on predicting user behaviors, such as [27] that aimed to identify those users that would evade a Wikipedia ban by creating a brand new account. Additional tasks in the same work included early detection of such evasion and account matching. Instead, [28] discussed the feasibility of proactive moderation interventions on Reddit. Results showed that Reddit communities are constantly evolving, and that communities bound to become toxic can be preemptively identified. This early detection opens up the possibility to intervene before the problem escalates, thus possibly resulting in the application of less restrictive and punitive interventions.

This survey of the existing literature in predictive content moderation showed that all previous works focused on detecting which content, users, or communities would later face or evade moderation. Conversely, to the best of our knowledge, no one has ever focused on predicting the outcome of a moderation intervention. Our present work contributes to filling this gap.

III. PROBLEM DEFINITION

We introduce the new task of predicting the dichotomous effect on user abandonment (yes or no) of a moderation intervention on a social media community. We first define the general framing of the problem, then we specify our binary classification task in the context of Reddit's Great Ban.

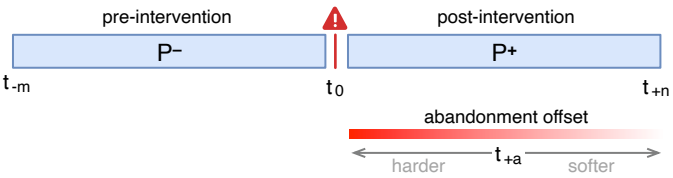


Fig. 1: Problem definition time frame. Given an intervention at t_0, we want to predict user abandonment in P^+, based on user data from P^- and inactivity during a time window from offset t_{+a} to t_{+n}.

A. Problem framing

Let i be a moderation intervention that occurred at time t_0, for which we delimit a surrounding time frame of interest that starts at t_{-m} and ends at t_{+n}. This time frame is thus composed of a pre-intervention period P^- that covers $[t_{-m}, t_0)$, and a post-intervention period P^+ that covers $(t_0, t_{+n}]$. We exclude t_0 at which i occurred because it is likely that user activity at this time is non-representative due to the application of the intervention itself (e.g., immediate and reactionary spikes in activity). Let P^* be the *abandonment time window* after the intervention that starts at offset t_{+a} and ends at t_{+n}, such that $P^* \subseteq P^+$. If there is no activity for a given user within the community of interest during P^*, we consider that user as *abandoning*. We also consider that the closer the abandonment offset t_{+a} to t_0, the *harder* is the abandonment, and the farther the *softer*, as illustrated in Fig. 1. For instance, in the case of $P^* = P^+$ with a user having non-zero activity immediately after the intervention and then no activity for the rest of P^*, the user would be considered as non-abandoning due to the rigidness of the $t_{+a} = t_0$ offset. However, it could be argued that users with no activity levels for several consecutive months some time after the intervention can hardly be considered active even if they had some activity immediately after the intervention. Hence, the selection of a larger offset $t_{+a} \gg t_0$ softens the abandonment time window P^*. At the same time, t_{+a} must not be too close to t_{+n}, otherwise P^* would be too short to be useful.

The classification task C is thus defined as predicting those users who were active in a given community before the intervention during P^-, but not afterward during the time window P^*, based on a dataset of user features D^- from P^-. This dataset is defined as:

$$D^- = \{(X_1, Y_1, \ldots, Z_1), \ldots, (X_N, Y_N, \ldots, Z_N)\}$$

where $X_u, Y_u, \ldots, Z_u$ are the values of the variables chosen as predictive features during P^- for a given user u.

In other words, let A^- be the activity level of user u during P^- and A^* the activity level for u during P^*. We then assign each user u to one of the following two classes:

$$C_{-1} = \{u | A^- > 0 \text{ and } A^* > 0\}$$
$$C_{+1} = \{u | A^- > 0 \text{ and } A^* = 0\}$$

subreddit	subscribers	users	comments
r/ChapoTrapHouse	159,185	9,295	1,368,874
r/The_Donald	792,050	4,262	619,434
r/DarkHumorAndMemes	421,506	1,632	35,561
r/ConsumeProduct	64,937	1,730	60,073
r/GenderCritical	64,772	1,091	94,735
r/TheNewRight	41,230	729	5,792
r/soyboys	17,578	596	5,102
r/ShitNeoconsSay	8,701	559	9,178
r/DebateAltRight	7,381	488	27,814
r/DarkJokeCentral	185,399	316	3,214
r/Wojak	26,816	244	1,666
r/HateCrimeHoaxes	20,111	189	775
r/CCJ2	11,834	150	9,785
r/imgoingtohellforthis2	47,363	93	376
r/OandAExclusiveForum	2,389	60	1,313

TABLE I: List of the 15 banned subreddits used for the analysis, sorted by number of active users. Data from these subreddits constitutes dataset $\mathbf{D}_{\text{B-B}}$.

with the negative class C_{-1} identifying non-abandoning users and the positive class C_{+1} identifying abandoning users. Then, the objective of the abandonment task is to learn the function f that, based on dataset D^{-}, assigns the correct class to each user u:

$$f(X_u, Y_u, \ldots, Z_u) \approx C_u \in \{C_{-1}, C_{+1}\}. \quad (1)$$

B. Task specification

In this work, we aim to predict those users of banned subreddits who abandoned the platform —that is, who ceased all activity on Reddit— after the Great Ban. Given the novelty of the problem and for the sake of simplicity, we perform two variants of the classification task C imposing $m = n$ so that P^{-} and P^{+} have the same duration, but with different abandonment time windows P^{*}. We call these tasks C^{H} for *hard abandonment* and C^{S} for *soft abandonment*, which are defined as:

$$C^{H} := \{C | P^{*} = P^{+}\}$$
$$C^{S} := \{C | P^{*} \subset P^{+} \text{ and } t_{+a} \gg t_0\}.$$

Based on previous research that analyzed the medium-term effect of moderation interventions on Reddit [11], [13], we determined the duration of both P^{-} and P^{+} to be 7 months (210 days), for a total time frame of 421 days (including the day t_0 of the Great Ban). Hence, for C^{H} the duration of P^{*} is also 7 months. On the other hand, for C^{S} we selected an offset t_{+a} of 4 months, with the duration of the corresponding P^{*} being 3 months (92 days), based on a precursory analysis of different offsets and class balances between non-abandoning and abandoning users.

IV. Data Collection and Preparation

For this study we use the dataset from [11], consisting of 16M Reddit comments made by 16,828 different users that were affected by the Great Ban.

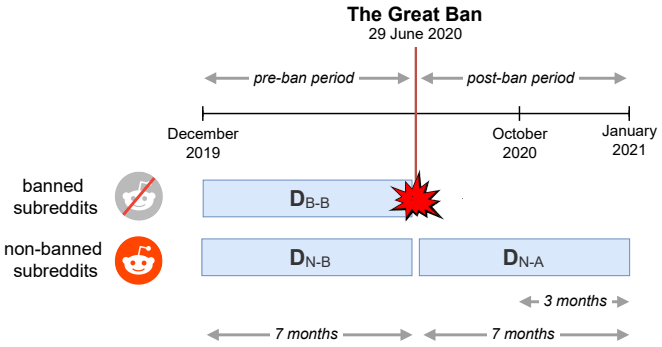


Fig. 2: Composition of our datasets and data collection periods. $\mathbf{D}_{\text{B-B}}$: data from within the banned subreddits, before the ban. Dataset $\mathbf{D}_{\text{B-B}}$ is used to select representative users from the banned subreddits. $\mathbf{D}_{\text{N-B}}$: data from non-banned subreddits, before the ban. Datasets $\mathbf{D}_{\text{B-B}}$ and $\mathbf{D}_{\text{N-B}}$ (i.e., both pre-ban datasets) are used to compute machine learning features. $\mathbf{D}_{\text{N-A}}$: data from non-banned subreddits, after the ban. Dataset $\mathbf{D}_{\text{N-A}}$ is used to provide ground-truth labels for the users based on their activity post-ban.

A. Banned subreddits selection

Each user in the dataset was an active participant before the ban in at least one of 15 selected subreddits. The selection of the 15 subreddits was done as follows. Out of the 2,000 subreddits shut during the Great Ban, Reddit only publicly disclosed the names of the 10 largest ones.[1] For all remaining banned subreddits, Reddit only provided a list of partially obfuscated names. The authors of [19] deciphered the list, removed all private subreddits as well as all those having less than 2,000 active users. Only 5 subreddits remained from the obfuscated list, which [19] and [11] studied in addition to the 10 publicly disclosed ones. Table I provides the list of the 15 subreddits on which our study is focused, together with some descriptive statistics.

B. Active users selection

For each of the 15 selected subreddits, the authors of [11] gathered all comments posted therein between December 2019 and June 2020, which resulted in a total of 8M comments posted by approximately 194K distinct users. As shown in Figure 2, the timeframe used for the data collection spans 7 months leading up to the Great Ban, providing a strong reference for the activity levels of the affected users before the moderation intervention [14]. The initial set of 194K users was later filtered so as to retain only those users who showed consistent activity in at least one of the 15 selected subreddits. In detail, only those users who posted at least one comment each month were retained. In [11], this filtering step was useful to obtain meaningful post-hoc estimations of the effect of the Great Ban on the activity of the users. Since here we also study changes in user activity, albeit in a predictive fashion, this filtering step is suitable for our analysis

[1] https://www.redditstatic.com/banned-subreddits-june-2020.txt (accessed: 03/15/2024)

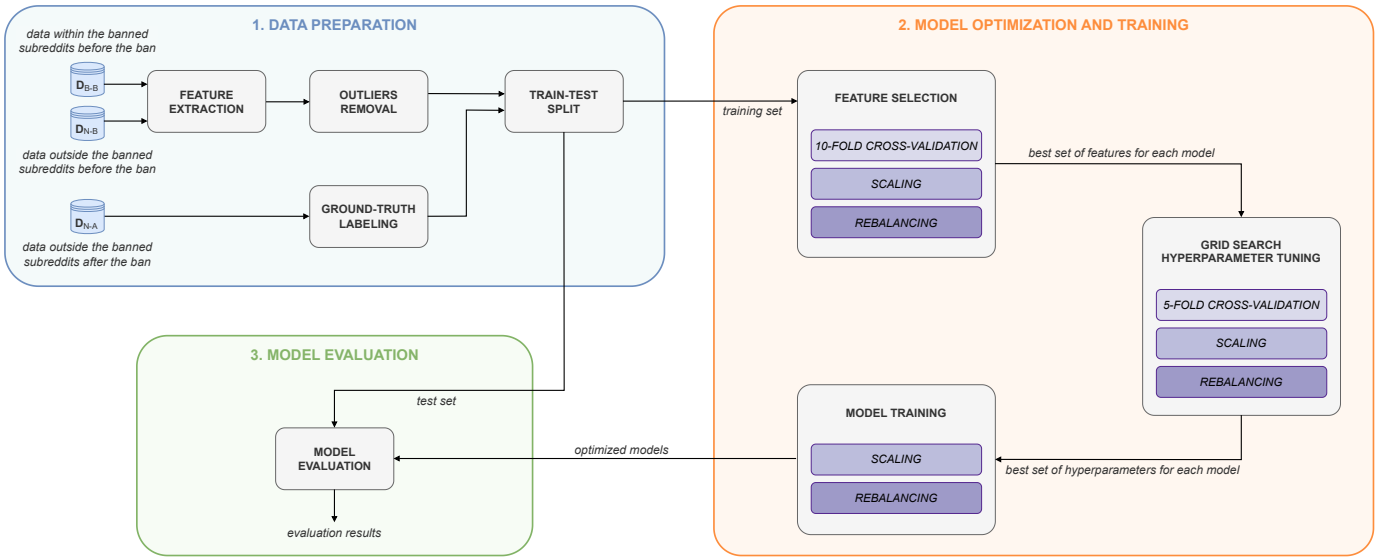


Fig. 3: Machine learning pipeline. Our data preparation steps involve feature extraction, outliers removal, ground-truth labeling, and splitting of the dataset into a training and a test set. The training set is used for model training and optimization. This involves feature scaling and selection, data rebalancing, and hyperparameters optimization. Finally, the optimized models are evaluated on the held-out test set.

as well. In addition, bots were also removed from the dataset. First, we followed reference practices in literature and we removed all accounts that posted two or more comments in less than a second [29]. Subsequently, we manually analyzed a random sample of 1,000 removed accounts to ensure that only bots were removed from the dataset. A similar analysis also revealed that among the remaining users in the dataset there were no clearly distinguishable bots. As a result of these filtering and validation steps, we obtained the dataset $\mathbf{D}_{\text{B-B}}$ composed of 2.2M comments posted within the 15 selected subreddits by 16,828 distinct users, as summarized in Table I.

C. Data from outside the banned subreddits

Evaluating (or predicting) the effect of a moderation intervention involves comparing data from before and after the intervention itself. This requirement also surfaces from the definitions that we gave in Section III. However, no activity exists post-intervention within the banned subreddits, since they were all permanently shut. Therefore, data about the behavior of the affected users *outside* of the banned subreddits must be used. For this reason, the authors in [11] collected all comments made by the 16,828 selected users outside of the 15 banned subreddits over a wide timeframe spanning 7 months before and 7 months after the Great Ban, as depicted in Figure 2. This additional data collection yielded approximately 13.8M comments, of which 8.2M (59%) were posted before the ban and constitute dataset $\mathbf{D}_{\text{N-B}}$ while 5.6M (41%) were posted afterwards and constitute dataset $\mathbf{D}_{\text{N-A}}$. We used the data from before the ban ($\mathbf{D}_{\text{N-B}}$) to compute our machine learning features, while the data posted after the ban ($\mathbf{D}_{\text{N-A}}$) allows assigning ground-truth labels to the users based on their post-ban activity. Notably, out of the 16,828 initially selected users, 288 (1.7%) did not have any activity outside of the 15 banned subreddits and were thus discarded.

V. Machine Learning Approach

The formulation of hard and soft abandonment prediction given in Section III-B cast both problems as binary classification tasks. Our approach to these tasks is sketched in Figure 3 and summarized in the remainder of this section, while the implementation details and the experimental settings are described in Section VII.

We begin by extracting user-level features from the portion of the dataset related to the activity of the selected users before the ban, both within ($\mathbf{D}_{\text{B-B}}$) and outside ($\mathbf{D}_{\text{N-B}}$) of the banned subreddits. Before further processing this data we perform outliers removal, as this helps enhancing the robustness and generalization of the trained models by mitigating the potential distortions and biases caused by extreme data points [30]. Data about user activities outside of the banned subreddits after the ban ($\mathbf{D}_{\text{N-A}}$) is used to assign ground-truth labels to the users. User-level features are then merged with the ground-truth labels and the resulting dataset is split into a training and a test set. The test set is only used for the final evaluation of the optimized models. Instead, the training set is used for feature selection and for optimizing and training the models. Given that we experiment with multiple classification models, each with its own characteristics and hyperparameters, the feature selection and optimization steps of our machine learning pipeline ensure that each model operates in optimal conditions. In other words, our approach ensures meaningful and fair comparisons between the different models. The first step in the model optimization process involves selecting an adequate number and set of features for each model. Specifically, we rank all features and we select the top-N ones to reduce the dimensionality of our models and to eliminate redundant and correlated features. For each model we experiment with multiple numbers N of selected features, as different models might benefit from a smaller or larger

number of features. To have accurate and robust estimates of the model performances with the different sets of features, we carry out a 10-fold cross-validation on the training set. At each iteration, we rescale features so that they all cover a similar range of values. Rescaling ensures that no feature dominates the learning process due to overly large magnitude, and also contributes to speeding up training times [31], [32]. At each iteration we also rebalance the training set. In fact, both the hard and soft abandonment tasks are heavily imbalanced, given that abandoning users constituted a small minority of all users (i.e., 14.9% for hard abandonment and 26.9% for soft abandonment). We note that this imbalance is common for tasks concerning the prediction of human behavior [33]. Therefore, by rebalancing the training set we reduce the bias that the trained models might exhibit towards those users who did not abandon the platform (i.e., the majority class). The averaged results of the 10-fold cross-validation process are used to select the optimal number and set of features for each model. Up to now, all models were trained and evaluated with default hyperparameters. In the last step of model optimization, for each model we perform a grid search with cross-validation over its hyperparameters so as to choose the best combination of them. Finally, we use the selected set of features and hyperparameters to train an optimized version of each model on the whole training set. Feature scaling and rebalancing are applied also for hyperparameter tuning and model training. Each optimized model is then evaluated on the held-out test set. Evaluation results and comparisons are presented in Section VIII.

VI. Feature Engineering

We compute a total of 142 features for each user in the dataset. Our features are organized into the following four main classes: 30 (21%) *activity* features; 40 (28%) *toxicity* features; 36 (25.5%) *writing style* features; and 36 (25.5%) *relational* features.

A. Activity features

Features in this class describe the overall level of engagement and participation of a user within Reddit. The motivation for including activity features is straightforward, given our goal of predicting user abandonment—that is, the interruption of posting activity on the platform. Moreover, previous descriptive studies on the effects of moderation interventions, such as those reviewed in Section II-A, showed that bans frequently cause a reduction in user activity [11], [13], [19], which supports the adoption of this type of features in our task. Some examples of the activity features that we implemented are the total number of comments posted by a user, the average time between the sharing of two subsequent comments, the slope of the trend of posted comments in either the banned and the non-banned subreddits, which captures whether user activity in certain subreddits was increasing or decreasing prior to the ban. Other features of this class capture possible peculiar characteristics of the comments, one of them being the number of *stickied* comments. This feature was included as it was among the strongest predictors of ban evasion in a previous study [28].

B. Toxicity features

The users considered in our study were active participants in subreddits banned due to the widespread presence of toxic content. As such, features gauging the degree of toxicity of the users pre-ban could be strong predictors in our task. In addition, integrating toxicity features into the classification models allows exploring the extent to which differences in toxic behavior are capable of explaining user reactions to the ban. We computed toxicity scores with Detoxify [34], a state-of-the-art [11] multilingual deep learning toxicity classifier that is widely used in literature [35], [36], including predictive studies for the detection of cyberbullying [37], [38]. Given a piece of text, Detoxify provides multiple indicators of toxicity, such as the *toxicity* and *severe toxicity* scores, as well as additional scores for *obscenity*, *insults*, *identity attacks*, and *threats*. In addition, we also computed sentiment scores given the strong relationship between toxicity and sentiment [39]. For obtaining sentiment scores we used Vader, a well-known rule-based sentiment analyzer that is specifically designed for social media content [40]. Vader outputs *positive*, *negative*, *neutral*, and *compound* sentiment scores, with the latter being the sum of the previous ones. To compute toxicity and sentiment user-level features we first classified each comment by each user with Detoxify and Vader. We then aggregated each score provided by the two tools over all comments by the same user, by computing the *mean*, *minimum*, *maximum*, and *standard deviation* of the scores, obtaining the $4 \times 10 = 40$ features in this class.

C. Writing style features

Several previous works demonstrated that certain users exhibit linguistic changes on a platform after suffering a moderation intervention. For example, [6] noted changes in the use of the first and third plural pronouns, as well as in word choices, after a ban. Similarly, [19] found that users who remained on Reddit after a ban exhibited reduced use of in-group language. Based on these preliminary descriptive results, here we assess the extent to which basic linguistic and writing style features provide predictive information about the activity of the users after a ban. Among the features that we computed are counts of the different *parts-of-speech* used and readability scores such as the well-known *Flesch-Kincaid Grade Level* and the *SMOG Index*. Each of these scores were computed for each user comment and subsequently aggregated at the user level by computing the *mean*, *minimum*, *maximum*, and *standard deviation*.

D. Relational features

The relationships between users and communities constitute the fundamental fabric upon which social media platforms are woven, shaping the dynamics of online engagement and interaction. Moreover, changes in user-user and user-subreddit relationships have already been observed as a consequence of moderation interventions [13], [28]. For this reason, considering relational features in our task allows us to capture the interaction dynamics that are inherent to Reddit communities.

As an example, we computed features that quantify the degree of influence that each user had in the banned subreddits, which could be predictive of future abandonment. This feature was computed by ranking users based on the average score (i.e., the difference between upvotes and downvotes) of their comments in the banned subreddits. In addition, the same feature was also computed for the non-banned subreddits. Similarly, we examined the relationship between users and specific subreddits, identifying those users who predominantly participated in the banned subreddits, which could provide an early sign of abandonment. In fact, users who almost exclusively participated in banned subreddits could lack motivation to stay on the platform after those subreddits were shut [6]. Finally, we leveraged results of recent works that demonstrated the usefulness of *initiative* and *adaptability* features in social media user classification tasks [41]. For example, we measured the number and ratio of threads started by each user as a proxy for their capacity to drive the conversation in a subreddit, rather than to follow what others say.

VII. Experiments and Settings

A. Comparisons

We experiment with multiple reference classification models to solve the hard and soft abandonment tasks, including Naive Bayes (`NB`), K-Nearest Neighbors (`KNN`), Decision Tree (`DT`), Random Forest (`RF`), Adaptive Boosting (`AB`), Gradient Boosting (`GB`), and Support Vector Machine (`SVM`). This selection spans the main families of classification algorithms, encompassing instance-based, probabilistic, ensemble-based, and discriminative methods of different complexities, thus providing an extensive exploration of many learning paradigms. Furthermore, similar choices of classification models were made in recent works on related tasks, such as the detection of different forms of online harms [42]–[45]. In addition to the previous classification models, we also implement three simple baselines for further comparison. The `Stratified` baseline generates predictions by following the ground-truth label distribution, without taking any feature into account. The `DT Ratio` baseline employs a decision tree that receives a single feature in input representing the ratio of comments made by a user in the non-banned subreddits over the comments made that user in the banned ones. The rationale for this baseline is that users who are more active on non-banned subreddits are less prone to abandon the platform. Finally, The `DT Trend` baseline employs a decision tree that receives a single feature in input representing the trend of the number of monthly comments by a user before the ban. The reason for including this baseline is that users with a decreasing posting trend are more likely to abandon the platform. For the sake of simplicity, we did not apply any preprocessing steps to the baselines. Finally, we remark that we resort to using these baselines instead of more powerful approaches because the novelty of our task makes it so that no previous work exists on predicting the effects of a moderation intervention.

B. Machine learning pipeline

In the following we report the experimental settings and implementation details of our machine learning pipeline, showed in Figure 3.

1) Outliers removal: To carry out outliers detection we apply *Isolation Forest*, a decision tree-based method that is particularly suitable for identifying anomalies in highly-dimensional data, such as ours [46]. Isolation Forest constructs decision trees by randomly selecting features and split points, efficiently isolating anomalies by assigning them shorter average path lengths in the trees compared to normal data instances, by leveraging the natural tendency of anomalies to require fewer splits for isolation. The application of Isolation Forest to our dataset led to the detection and removal of 297 outliers. A manual verification revealed that the removed users featured extremely low levels of activity, which supports their removal.

2) Train-test split: We use a standard 80/20 split of our dataset into the training and test sets, following the Pareto principle. Data in the two sets is stratified according to the class labels.

3) Cross-validation: During model optimization we perform two stratified cross-validations on the training set to obtain robust estimates of model performance. In detail, we perform a 10-fold cross-validation during feature selection and a 5-fold cross-validation during hyperparameter tuning. The stratification is necessary given our marked class imbalance, so that each class is represented proportionally in both the training and validation sets of the cross-validation.

4) Feature scaling and rebalancing: We perform feature scaling and data rebalancing during feature selection, hyperparameter tuning, and model training. Feature scaling is applied first, via *Z-score standardization*, so that each feature has mean $\mu = 0$ and standard deviation $\sigma = 1$. Rebalancing is applied last, as suggested in previous literature [47]. For rebalancing, we leverage both oversampling and undersampling techniques to mitigate the large class imbalance of our dataset. We perform oversampling with *SMOTE*, a state-of-the-art technique that creates new artificial samples of the minority class [47]. Instead, we use *Random Under Sampling* to randomly discard samples from the majority class [48]. For both tasks we use the two techniques in combination to reach a final fixed imbalance of 60/40 between the majority and minority class, down from 85/15 for hard abandonment and 73/27 for soft abandonment. This approach allows us to reduce the skewness of the class distribution without creating too many artificial samples that could introduce bias, and without discarding too much data that could result in degraded model performance.

5) Feature selection: We perform feature selection via *ANOVA F-value*, a technique that was shown to achieve good performance in recent works on the detection of malicious online user activities [49]. The ANOVA tests are used to obtain an F-value for each feature, expressing the likelihood for the feature to be predictive for the task at hand. Then, features are ranked based on their respective F-value and the top-k are selected. Out of 142 total features, for each trained model we vary k from 10 to 80 with a step of 10, and evaluate the performance of the resulting models as part of the 10-fold cross-validation. At the end of the 10-fold cross-validation process, for each model we select the number and set of features that maximizes its F1 score on the positive class, so

task	model	p	positive class			overall	
			precision	*recall*	*F1*	*macro F1*	*micro F1*
hard abandonment	*baselines*						
	Stratified	–	0.155	0.158	0.157	0.505	0.751
	DT Ratio	1	0.356	0.199	0.255	0.530	0.689
	DT Trend	1	0.325	0.219	0.262	0.547	0.726
	trained models						
	KNN	20	0.514	0.308	0.384	0.616	0.754
	NB	10	0.294	0.300	0.297	0.588	0.792
	RF	40	0.584	0.355	0.442	0.652	0.779
	AB	20	0.607	0.352	0.446	0.652	0.774
	DT	10	**0.617**	0.330	0.430	0.637	0.756
	GB	20	0.603	**0.392**	**0.474**	**0.676**	**0.800**
	SVM	50	0.520	0.373	0.434	0.656	0.797
soft abandonment	*baselines*						
	Stratified	–	0.263	0.266	0.264	0.497	0.605
	DT Ratio	1	0.374	0.328	0.349	0.543	0.624
	DT Trend	1	0.282	0.319	0.299	0.530	0.644
	trained models						
	KNN	10	0.513	0.438	0.473	0.627	0.691
	NB	10	0.400	0.485	0.438	0.627	0.723
	RF	20	0.454	0.501	0.476	0.647	0.730
	AB	10	**0.571**	0.470	0.515	0.654	0.710
	DT	20	0.558	0.443	0.494	0.636	0.690
	GB	20	0.538	0.496	**0.516**	**0.664**	0.728
	SVM	50	0.484	**0.513**	0.499	0.660	**0.737**

TABLE II: Classification results for the hard abandonment (top rows) and soft abandonment (bottom rows) tasks. Column *p* reports the number of features used by each model. For each task, the best result in each evaluation metric is shown in **bold** and the second-best is underlined.

that for each model, only its best set of features is used in the following steps.

6) Hyperparameter tuning: For each model, we perform hyperparameter tuning with a grid search 5-fold cross-validation on the training set.

7) Model training: Finally, we train each model on the whole training set by using the best set of features and hyperparameters that we determined.

C. Model evaluation

We evaluate the optimized models and the baselines on the test set by means of standard evaluation metrics for classification tasks. Specifically, we report averaged scores such as the *micro* and *macro F1 scores*. Additionally, we report the *precision*, *recall*, and *F1 score* of the positive class (i.e, abandoning users), given its importance for moderators and platform administrators in the context of planning moderation interventions.

VIII. Results

A. Prediction of hard and soft abandonment

Classification results for the hard abandonment task are reported in the top half of Table II, while those for the soft abandonment task are reported in the bottom half of the table. For each task we report the performance of the baselines and of the optimized version of each trained model. For each baseline and model, Table II shows the number of used features (*p*) and the evaluation metrics computed for the positive class and for both classes (**overall** columns). With respect to the hard abandonment task, Gradient Boosting (GB) achieved the best results all-round. It was the best performing model in each evaluation metric, with the exception of *precision* on the positive class, where it was surpassed by Decision Tree (DT) and Adaptive Boost (AB). The second-best model in this task is Support Vector Machine (SVM) that achieved competitive results in all metrics and second-best results in positive class *recall*, overall *macro F1*, and overall *micro F1*. Despite the relatively solid performance exhibited by some models (e.g., GB and SVM), the results from Table II demonstrate the difficulty of this task. First of all, in general the results obtained by the trained models were encouraging but not exceptionally good, with the best result being *micro F1* $= 0.800$ by GB. Second, the difficulty faced by the trained models is also testified by the small margin by which some of the models —such as K-Nearest Neighbors (KNN) and Naive Bayes (NB)— outperformed the baselines. Finally, the results in table highlight a clear trend where the more complex models consistently outperformed the simpler ones. This shows that the increased complexity of models such as GB, SVM, and AB was needed in order to obtain more accurate predictions. However, in spite of the difficulty of the task, the majority of optimized models used a relatively low number of features. Indeed, with the exception of SVM and RF, all other models used $p \leq 20$ features, out of the 142 total ones that we computed. This result implies that many features conveyed little or redundant information for the task. The scores reported in the bottom half of Table II reveal more nuanced results for the soft abandonment task. A first interesting finding is that the **overall** results are worse for the soft abandonment task than for hard abandonment, while those for the **positive class** are better. For example, the best *micro F1* $= 0.737$ in soft abandonment, versus best *micro F1* $= 0.800$ in hard abandonment. Similarly, the best *macro F1* $= 0.664$ in soft abandonment, versus best *macro F1* $= 0.676$ in hard abandonment. On the contrary, the best positive class *F1* $= 0.516$ in soft abandonment, which is higher than *F1* $= 0.474$ in hard abandonment. In other words, by moving from hard to soft abandonment we improved predictions on the positive class (i.e., the abandoning users) but we degraded those on the negative one (i.e., the users who remained active). There are two related reasons for this behavior: (*i*) the definition of abandoning users and (*ii*) class imbalance. For the former, we recall that the soft abandonment task was introduced precisely to address the issue with those users who maintain some activity in the aftermath of the ban, but that eventually leave the platform, which represent challenging instances to classify. Therefore, their labeling in the soft abandonment task as abandoning users (positive class) rather than active ones (negative class), led to a certain degree of improvement of the models on the positive class. The second related factor is about class imbalance given that, before rebalancing, the soft abandonment task is less imbalanced than hard abandonment: 73/27 versus 85/15, respectively. To this end, we recall that abandoning users are the positive and

task	level of activity†	users		evaluation metrics		
		pos.	*neg.*	*positive F1*	*macro F1*	*micro F1*
hard aban.	VH $731 < n$	353	2,896	0.400	0.656	**0.848**
	HI $334 < n \leq 731$	386	2,858	0.402	0.650	<u>0.826</u>
	ME $156 < n \leq 334$	432	2,801	0.446	<u>0.665</u>	0.808
	LO $55 < n \leq 156$	464	2,783	<u>0.460</u>	**0.666**	0.794
	VL $n \leq 55$	793	2,477	**0.517**	0.643	0.688
soft aban.	VH $731 < n$	764	2,485	0.487	0.666	0.736
	HI $334 < n \leq 731$	742	2,502	0.519	<u>0.691</u>	<u>0.788</u>
	ME $156 < n \leq 334$	759	2,474	<u>0.577</u>	**0.725**	**0.805**
	LO $55 < n \leq 156$	833	2,414	0.521	0.676	0.751
	VL $n \leq 55$	1,286	1,984	**0.600**	0.639	0.644

† VH: Very High; HI: High; ME: Medium; LO: Low; VL: Very Low

TABLE III: Classification results of Gradient Boosting for the hard abandonment (top rows) and soft abandonment (bottom rows) tasks at different levels of activity of the users. For each level of activity we report the corresponding range of comments *n*, the number of users in each class, and the evaluation metrics: *F1 score* of the *positive* class, overall *macro F1*, overall *micro F1*. For each task, the best result in each evaluation metric is shown in **bold** and the second-best is <u>underlined</u>.

the minority class in both tasks. As such, models trained for detecting soft abandonment train on a relatively larger number of positive instances than those for hard abandonment, hence the better performance on the positive class. Besides, results for the soft abandonment task are nuanced also in terms of the best performing models. In fact, there is no clear winner in this task, with SVM and GB —and to a lower extent, RF— achieving comparable performance. Nonetheless, the trend about the better performance of the more complex models is confirmed also in the soft abandonment task.

B. Predicting abandonment at different levels of user activity

We now delve deeper into the results by analyzing classification performance in relation to the level of user activity on the platform pre-ban. Evaluating classification performance based on user activity bears multiple implications, as it allows assessing whether classifiers exhibit varying levels of performance across different segments of users. For example, particularly active users —who exert the greatest influence on the platform— are pivotal in shaping community dynamics and behaviors [33], [50]. Therefore, correctly predicting their activity after an intervention provides moderators and platform administrators with valuable insights for their content management strategies. Conversely, accurately detecting the least active users —who are more likely to abandon the platform— enhances the utility of the prediction system by enabling proactive or alternative measures to retain users and mitigate churn [13].

To perform this analysis, we examine the distribution of user comments pre-ban and we assign each user to one of the following five activity levels: very low (VL), low (LO), medium (ME), high (HI), and very high (VH) activity. The five activity levels are obtained by binning the distribution of user comments at regular intervals of 20 quantiles each, so that $\texttt{VL} \leq Q_{20}$, $Q_{20} < \texttt{LO} \leq Q_{40}$, and so on up to $\texttt{VH} > Q_{80}$. We then perform a 80/20 train-test split of the dataset stratifying not only for class labels but also for the newly defined activity levels. Finally, we pick the best performing model based on the results in Table II —that is, Gradient Boosting (GB)— we train it on the training split, and we test it on the held-out data. The results of this analysis are reported in Table III for both the hard (top rows) and soft abandonment (bottom rows) task. In table are reported the number of comments *n* in each activity level, the corresponding number of abandoning (positive class) and remaining (negative class) users, and the evaluation metrics: the *F1 score* of the *positive* class, the overall *macro F1*, and the overall *micro F1*. The number of users in the positive and negative classes, reported in Table III for each activity level and for each task, shows that the imbalance in the dataset increases when considering increasingly active users. As we already observed in the results of Table II, class imbalance in the data has a detrimental effect on the classification performance on the minority class (i.e., the positive class, corresponding to the abandoning users). In Table III, this is reflected by the fact that the highest *positive F1* scores are obtained at the lowest levels of user activity, for both tasks. Indeed, the best *positive F1* $= 0.600$ is achieved on the soft abandonment task for users with very low (VL) activity, which is the experimental condition with the lowest class imbalance (61/39). In contrast, the best overall results are obtained for medium to high activity levels. The best all-round result is achieved in the hard abandonment task for users with very high (VH) activity, with *micro F1* $= 0.848$. Then, in the soft abandonment task the best results are obtained at the medium (ME) and high (HI) activity levels, with *micro F1* $= 0.805$ and 0.788, respectively.

C. Leave-one-out cross-validation

Until now, we trained our models on a dataset obtained by merging users from all 15 banned subreddits. However, users who participate in different subreddits could have different characteristics and exhibit different behaviors and reactions to a moderation intervention [14]. For this reason, we carried out an additional analysis aimed at assessing possible subreddit-specific factors influencing user abandonment. Importantly, this analysis also allows evaluating the extent to which a model trained on a certain set of subreddits is capable of generalizing to other —unseen— subreddits. To assess the generalizability of our best model we performed a leave-one-out cross-validation (LOOCV) by leveraging the multiplicity of banned subreddits in our dataset. This procedure represents the state-of-the-art in evaluating the generalizability of a classifier against multiple groups or classes of data instances [51], and has been recently used for related tasks such as the detection of different groups of bots [52], [53]. Based on the 15 banned subreddits in our dataset, we implemented the LOOCV by iteratively selecting 14 subreddits to train a Gradient Boosting (GB) hard and soft abandonment detection classifier, and by testing the trained model on users from the remaining (held-out) subreddit. For this analysis, we considered a user to participate in a subreddit if it posted more than 10 comments in that subreddit, so as to reduce noise caused by sporadic participation.

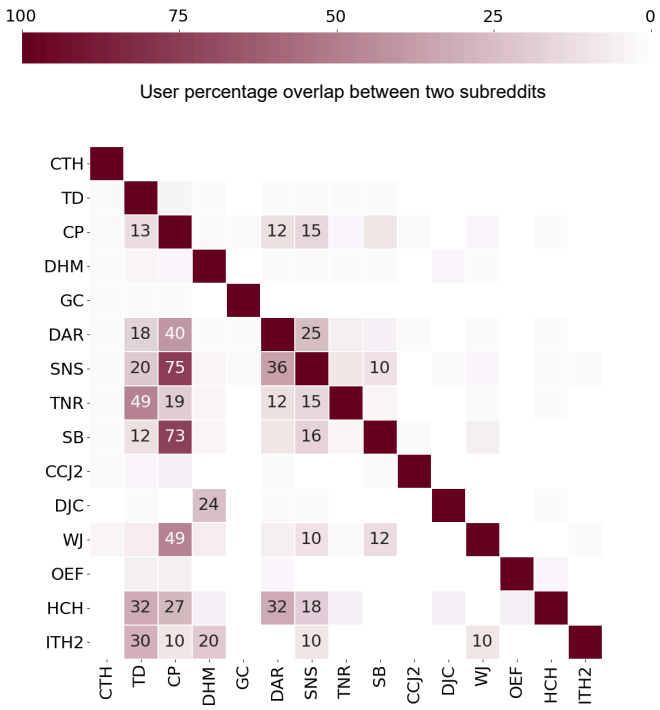


Fig. 4: User overlap matrix O between the banned subreddits. Each cell o_{ij} in the matrix reports the percentage of users from the i-th subreddit that also participated in the j-th subreddit. White-colored cells represent 0% overlap while dark-colored ones represent 100% overlap. The exact amount of overlap is shown for cells with $o_{ij} > 10\%$. Subreddits are ordered by decreasing number of participating users.

The validity of the LOOCV hinges on the assumption of independence between the training and testing datasets. Specifically, to ensure unbiased evaluation, it is required that the held-out subreddit contains users who are largely distinct from those in the training subreddits. While the independence assumption is easily verified in some domains of application, such as that of bot detection where each bot belongs to only one botnet [52], it does not hold in our context where any user can participate in multiple subreddits. When a considerable overlap exists between the sets of users who participate in two subreddits, utilizing either of these subreddits as the held-out test dataset may leak information to the classifier from the training datasets, thereby compromising the validity of the evaluation [54]. Before running the LOOCV we thus assessed the extent of overlap between the sets of users who participated in the 15 banned subreddits. The results of this analysis are presented in Figure 4 as a heatmap of the user overlap matrix between the subreddits. In figure, each cell o_{ij} in the user overlap matrix O reports the percentage of users from the i-th subreddit that also participated in the j-th subreddit. By definition, cells in the matrix diagonal correspond to 100% overlap and are dark-colored, while those that correspond to no overlap are white colored. Apart from `r/ShitNeoconsSay` (SNS) and `r/soyboys` (SB) that share up to 75% of their users with `r/ConsumeProducts` (CP), the matrix is overall very sparse, as demonstrated by the limited reported overlap. The results from this analysis show that the sets of users who participated in the banned subreddits are largely disjoint, in line with the literature on echo chambers [55], which supports the application of the LOOCV. Table IV shows the results of the LOOCV for the hard (top rows) and soft abandonment (bottom rows) tasks. Each row in the table reports testing results on a specific subreddit, when training on all others. In addition to the evaluation metrics, for each subreddit we also report the number of participating users and its maximum overlap with the other subreddits. Finally, for each task, the bottom rows report evaluation results aggregated across all subreddits, in terms of mean and standard deviation of each metric, as well the scores obtained by the `GB` model trained and tested on data from all subreddits —that is, without the LOOCV— as reported in Table II. The latter scores serve as reference values with which to compare the LOOCV results to quantify the performance decrease to expect when classifying users from an unseen subreddit. The averaged results reported in Table IV for both tasks indicate a high variability in the evaluation metrics of the positive class, as demonstrated by the large standard deviation with respect to the means. For instance, in the hard abandonment task, the *F1 score* ranges from a minimum of 0.235 for `r/Wojak` (WJ) to a maximum of 0.569 for `r/ConsumeProduct` (CP). Similarly, the *recall* ranges from 0.160 for `r/GenderCritical` (GC) to 1.000 for `r/OandAExclusiveForum` (OAE). The same can be observed for the soft abandonment task as well, where the *precision* ranges from 0.236 for `r/ChapoTrapHouse` (CTH) to 1.000 for `r/HateCrimeOaxes` (HCH). These results are indicative of substantial differences among the banned subreddits. In fact, in addition to the overall limited user overlap observed in Figure 4, the results from Table IV highlight notable differences in the behavior exhibited by the participants of such subreddits. Interestingly enough, the maximum overlap that each subreddit has with others, has little influence on the performance of the trained model. For each task, we computed the Pearson correlation coefficient r to estimate the strength of the linear relationship between the max overlap and three evaluation metrics —namely, the *F1 score* of the positive class, the *macro F1*, and the *micro F1*. The largest value measured for the hard abandonment task is $r = 0.237$ ($p = 0.394$), while the largest measured for the soft abandonment task is $r = 0.222$ ($p = 0.493$), both of which indicate weak and non-significant correlations. In turn, this strengthens the results reported in Table IV, including those about subreddit diversity, as the variable results in the table cannot be simply explained by the fact that some users participated in multiple subreddits, but rather to their inherent differences. The comparison between the results obtained without LOOCV and the LOOCV aggregate results sheds light on the capacity of our model to generalize to unseen data. To this end, results in Table IV show a moderate loss in performance in the LOOCV experiment, which is expected given the aforementioned differences among the subreddits. However, the percentage loss is generally contained. For example, in the hard abandonment task the average *F1 score* on the positive class obtained with the LOOCV is 71.9% of

task	test subreddit		users	max overlap	positive class			overall	
					precision	*recall*	*F1*	*macro F1*	*micro F1*
hard abandonment	CTH	r/ChapoTrapHouse	8,471	0.2%	0.206	0.465	0.286	0.581	0.790
	TD	r/The_Donald	3,539	4.1%	0.481	0.508	0.494	0.666	0.755
	CP	r/ConsumeProduct	1,098	15.0%	0.506	0.650	0.569	0.686	0.729
	DHM	r/DarkHumorAndMemes	929	3.8%	0.202	0.527	0.292	0.544	0.683
	GC	r/GenderCritical	918	1.4%	0.570	0.160	0.250	0.524	0.681
	DAR	r/DebateAltRight	325	40.0%	0.536	0.300	0.385	0.588	0.687
	SNS	r/ShitNeoconsSay	226	75.0%	0.653	0.368	0.471	0.618	0.676
	TNR	r/TheNewRight	144	49.0%	0.350	0.500	0.412	0.613	0.718
	SB	r/soyboys	142	73.0%	0.559	0.452	0.500	0.659	0.732
	CCJ2	r/CCJ2	110	4.5%	0.556	0.294	0.385	0.650	0.852
	DJC	r/DarkJokeCentral	94	24.0%	0.176	0.600	0.273	0.517	0.640
	WJ	r/Wojak	57	49.0%	0.500	0.154	0.235	0.548	0.764
	OEF	r/OandAExclusiveForum	37	5.4%	0.143	1.000	0.250	0.522	0.676
	HCH	r/HateCrimeHoaxes	22	32.0%	0.667	0.200	0.308	0.509	0.591
	ITH2	r/imgoingtohellforthis2	10	30.0%	0.000	0.000	0.000	0.438	0.778
	averaged LOOCV results				0.407 ± 0.202	0.412 ± 0.237	0.341 ± 0.137	0.577 ± 0.069	0.717 ± 0.063
	training/testing on all subreddits (no LOOCV)				0.603	0.392	0.474	0.676	0.800
soft abandonment	CTH	r/ChapoTrapHouse	8,471	0.2%	0.236	0.809	0.365	0.464	0.481
	TD	r/The_Donald	3,539	4.1%	0.658	0.541	0.594	0.684	0.709
	CP	r/ConsumeProduct	1,098	15.0%	0.701	0.473	0.565	0.651	0.672
	DHM	r/DarkHumorAndMemes	929	3.8%	0.444	0.446	0.445	0.626	0.713
	GC	r/GenderCritical	918	1.4%	0.739	0.213	0.331	0.520	0.594
	DAR	r/DebateAltRight	325	40.0%	0.697	0.317	0.436	0.533	0.554
	SNS	r/ShitNeoconsSay	226	75.0%	0.738	0.477	0.579	0.594	0.595
	TNR	r/TheNewRight	144	49.0%	0.571	0.483	0.523	0.618	0.641
	SB	r/soyboys	142	73.0%	0.711	0.403	0.514	0.615	0.641
	CCJ2	r/CCJ2	110	4.5%	0.438	0.212	0.286	0.538	0.676
	DJC	r/DarkJokeCentral	94	24.0%	0.297	0.733	0.423	0.592	0.663
	WJ	r/Wojak	57	49.0%	0.778	0.280	0.412	0.574	0.636
	OEF	r/OandAExclusiveForum	37	5.4%	0.455	0.714	0.556	0.706	0.784
	HCH	r/HateCrimeHoaxes	22	32.0%	1.000	0.250	0.400	0.545	0.591
	ITH2	r/imgoingtohellforthis2	10	30.0%	0.000	0.000	0.000	0.357	0.556
	averaged LOOCV results				0.564 ± 0.245	0.423 ± 0.213	0.429 ± 0.146	0.574 ± 0.085	0.633 ± 0.073
	training/testing on all subreddits (no LOOCV)				0.538	0.496	0.516	0.664	0.728

TABLE IV: Results of the leave-one-out cross-validation (LOOCV) analysis. The top half of the table reports results for the hard abandonment task, while the bottom half is related to the soft abandonment. For both tasks, each row shows the classification results obtained with the reported subreddit used as test set and the remaining ones as training set. For each task, the bottom rows report the LOOCV results aggregated over all subreddits (mean ± standard deviation) and the reference results obtained by the same model without the LOOCV.

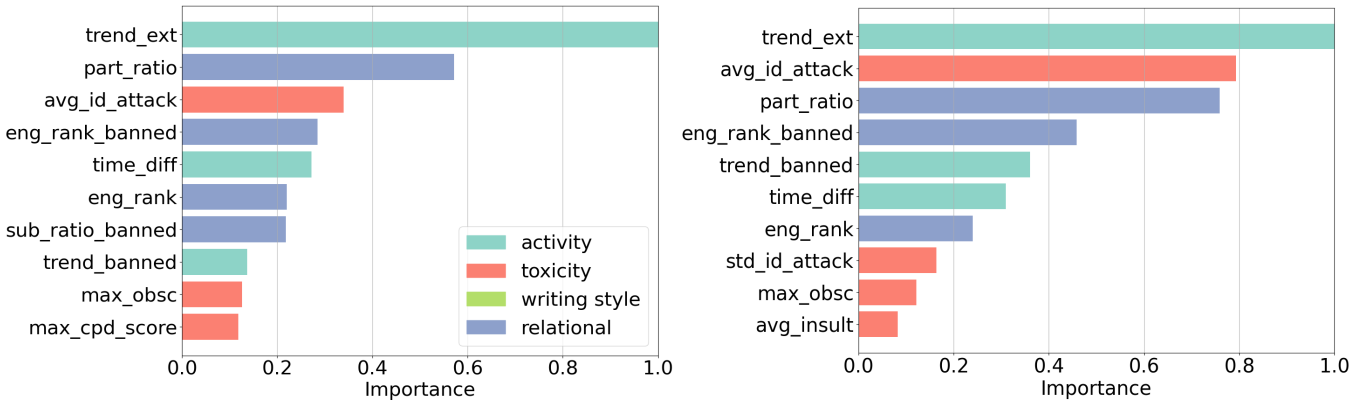
the original one. The loss for the overall metrics is even lower, with the average *macro F1* and *micro F1* being respectively 85.3% and 89.6% of the original respective metrics. Similar percentage losses are obtained for the soft abandonment task in the overall metrics, with the LOOCV average *macro F1* and *micro F1* maintaining respectively 86.5% and 87.0% of the original performance. Overall, these figures suggest that the model maintains encouraging performance even when applied to unseen data.

D. Feature importance

Our last analysis involves estimating the importance of the individual features as well as of the classes of features that we implemented, for both the hard and soft abandonment tasks. This analysis complements our previous results by providing additional information for interpreting model predictions and by highlighting which features are most predictive of user abandonment following a moderation intervention. We rely on SHapley Additive exPlanations (SHAP) as the building block of our feature importance analysis. SHAP is a method based on cooperative game theory that is widely used for interpreting the output of machine learning models and for estimating the contribution of individual features to a model's predictions [41], [56]. Here, we apply SHAP in conjunction with an information-fusion based sensitivity analysis technique to compute local and global feature importance scores [44]. Specifically, for each feature f_i and for each trained model $M_j \in \{\texttt{NB}, \texttt{KNN}, \texttt{DT}, \texttt{RF}, \texttt{AB}, \texttt{GB}, \texttt{SVM}\}$, we use SHAP to compute the contribution $C(i,j,d)$ of f_i to the prediction given by M_j for a specific data point d. We then aggregate the contributions over all data points N in the test set to compute a local score:

$$\text{score}_L(i,j) = \frac{1}{N}\sum_{d=1}^{N}\big|C(i,j,d)\big|. \tag{2}$$

The $\text{score}_L(i,j)$ expresses the *local* contribution, or importance, of the feature f_i for the model M_j. Next, we obtain



(a) Hard abandonment.

(b) Soft abandonment.

Fig. 5: Top-10 most important features for the hard and soft abandonment tasks. The ranking of the most important features is similar between the two tasks, with features such as *trend_ext*, *part_ratio*, and *avg_id_attack* dominating the ranking. All classes of features are represented in the top-10, with the exception of *writing style* features that do not seem to provide individually strong contributions to models predictions.

a *global* score for the feature f_i as the weighted mean of its local importance across all trained models:

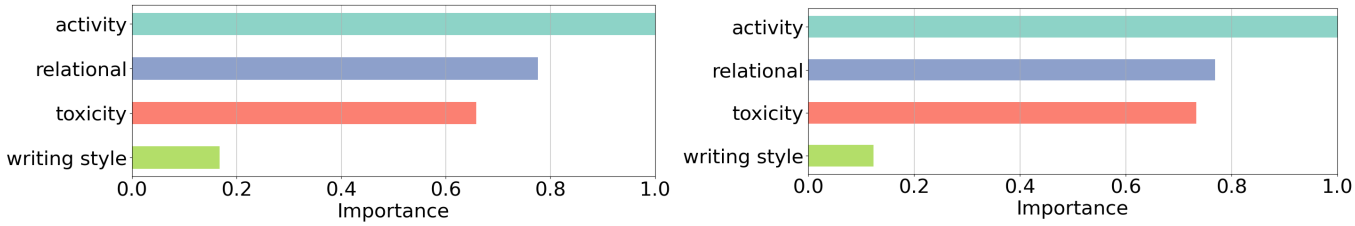
$$\text{score}_G(i) = \frac{\sum_j w_j \cdot \text{score}_L(i,j)}{\sum_j w_j}. \tag{3}$$

We use the *F1 score* on the positive class as the weighting factor w_j of each model M_j, so that models that are better at detecting abandoning users are weighted more in Eq. (3). The $\text{score}_G(i)$ expresses the weighted contribution of the feature f_i throughout all models that we trained and tested —hence the *global* label— thus representing a robust indicator of the actual importance of that feature for the task. As a final presentation step, we rank all features and assess their relative importance by normalizing their global scores so that the best performing feature has a normalized score of 1 and the others are rescaled proportionally.

Figures 5a and 5b show the top-10 most important features according to their normalized global scores, respectively for the hard and soft abandonment tasks. The set of the most important features for both tasks is very similar, although the ranking is different and the features exhibit different relative importance. In both tasks, among the most informative features are the trend of comments in non-banned subreddits (*trend_ext*), the ratio between the number of non-banned and banned subreddits to which a user participated (*part_ratio*), and the average score of "identity attacks" in the user's comments (*avg_id_attack*). For the hard abandonment task, in particular, the contribution of the *trend_ext* feature is almost double that of the second-best one. To this end we recall that the `DT Trend` baseline defined in Section VII-A is based on this exact feature, which explains the good performance of that baseline in the results of Table II. By observing the relative importance between the top and bottom features in Figures 5a and 5b, we also derive further insights into the results of Table II. Specifically, we note that in both tasks the first few features provide a much larger relative contribution than the last ones in the top-10. This means that, apart from the first few features, all others, including those highly ranked according to our feature importance analysis, provide relatively few information to the models. This result resonates with that presented in Table II about the small number of features used by the majority of models. Finally, we note that three out of four classes of features are represented among the top-10 features shown in Figure 5. The only class of features that did not make it to the top-10 is the one conveying information the *writing style* of the users, which does not seem to provide much relevant information.

To better assess the contribution of the different classes of features, rather than that of the individual features, we aggregate the contributions of all features based on their class. In detail, we sum the global importance scores of all features separately for each class, and we divide it by the number of features in the class, to obtain a class feature importance score. Lastly, we normalize the class scores so that the best class has a normalized score of 1 and the others are rescaled proportionally.

Figure 6 shows the relative importance of the different classes of features for the two tasks. The ranking is the same in both tasks, with *activity* features providing the largest contribution. *Relational* features are the second-most informative class, providing about 80% of the contribution of activity features. *Toxicity* features obtain a similar score. Instead, *writing style* features are ranked last in both tasks and provide small contributions with respect to the other classes. This result confirms that shown in Figure 5, where no *writing style* feature was ranked among the top contributing features in either task. Overall, the results obtained for *activity*, *relational*, and *toxicity* features highlight the importance of user engagement, social dynamics, and toxic speech in predicting user abandonment after the investigated moderation intervention. Conversely, information about the *writing style* of the users does not appear to be as relevant.



(a) Hard abandonment.

(b) Soft abandonment.

Fig. 6: Feature importance of each class of features, obtained by summing the contributions of the individual features in each class, normalized by the number of features in the class. The contribution of the different classes of features is largely the same in both tasks, with *activity* features providing the largest contribution, followed by *relational* and *toxicity* features. *Writing style* features provide overall marginal contributions.

IX. Discussion

A. Detecting abandoning users

In this study, we defined and tackled the novel task of predicting user abandonment following a massive deplatforming moderation intervention on Reddit [11]. In consideration of the novelty of the task —tackled here for the first time— our results are promising. For example, we achieved *micro F1* $= 0.800$ (*macro F1* $= 0.676$) when detecting users who immediately halted activity after the moderation intervention (hard abandonment), and *micro F1* $= 0.737$ (*macro F1* $= 0.660$) when detecting users who initially maintained some activity but who eventually left Reddit (soft abandonment). Albeit preliminary, these results demonstrate that the task of preemptively estimating the effects of future moderation interventions is indeed feasible. Adding to the encouraging findings are the results of our leave-one-out cross-validation (LOOCV) analysis, which underscore the absence of substantial performance degradation when applying our model to unseen data and shows that the model learned generalizable behavioral patterns. At the same time however, our study highlighted some of the challenges of this new task. Among them are the difficulties at forecasting complex and heterogeneous user reactions [18], a dire class imbalance [33], a partial and unreliable view of platform moderation actions [3], and the growing difficulties at accessing platform data [57], which culminate in the lack of extensive labeled datasets. The inherent challenges of predicting the effects of moderation interventions —exemplified by our promising yet preliminary results— imply that much work remains to be done before moderators and platform administrators will be able to leverage powerful and dependable tools for accurately estimating the effects of their actions. To this end, this study paves the way for future endeavors in this emerging area of content moderation, which proposes a new pathway for empowering moderators to optimize interventions based on desired outcomes, such as reducing toxicity while minimizing user attrition and churn. This is currently an as-yet unexplored frontier of moderation, which will open up once it achieves high performance in tasks such as the one proposed and tackled in this work.

B. Characterizing moderated users

Although the performances of the models developed in this study are promising, there is much room for improvement. Model performance strongly depends on the informativeness of the provided features. In this initial study, we computed and tested an extensive and diverse set of 142 features, many of which were borrowed from recent literature on related tasks [28], [41], [44], [58]. Our set of features included information about the activity of the users before the moderation intervention, their participation in multiple communities, their relationships with other users, their use of toxic or otherwise aggressive speech, and their writing style. While some of the extracted features turned out to be powerful predictors in our models, a large share of the features that we evaluated was not particularly informative, which hindered model performance. This suggests that our task, proposed here for the first time, is substantially different from others previously addressed in related literature and for which those features provided robust performance. This finding underscores the importance of tailoring feature selection to the specific characteristics of the task, rather than relying on already-proposed features. Future efforts towards the prediction of moderation intervention effects should thus consider implementing and experimenting with a broader and even more diversified set of features. To this end, promising directions for future work involve the development of socio-psychological features. In fact, predicting the effects of moderation interventions essentially revolves around predicting user behavior. Moreover, it is known that socio-psychological characteristics influence online behaviors, including toxic and aggressive ones [18]. A recent demonstration of the latter can be found in those works who measured a significant improvement in hate speech detection tasks, when incorporating socio-psychological features [59]. By the same token, those features could provide valuable information also in the related task of predicting behavioral changes following a moderation intervention, and particularly for those interventions targeting hateful or toxic users.

C. Predicting effects: classification, quantification, regression

In this work we tackled the problem of predicting the users that would abandon Reddit following a moderation intervention as a binary classification task. Currently, there is a dire lack of tools and systems to predict or estimate the effects of moderation interventions. Within this context, our work paves the way to the development of decision support tools to assist moderators and platform administrators in carrying out

effective and efficient content moderation, thus overcoming the limitations of the current trial-and-error approach [13]. In fact, in light of the promising results that we achieved, we envision the possibility to develop additional tools in the near future, to further support moderation endeavors by leveraging established machine learning paradigms. Among them are quantification and regression approaches. For instance, regarding the former, a platform may not be interested in knowing which specific users are likely to abandon or remain on the platform following a moderation intervention. Such a need could arise, among other reasons, due to privacy concerns [4]. In that case, effective moderation could still be guaranteed by computing aggregated estimates of how many users would likely leave compared to those who would remain. From the methodological standpoint, such a problem would be better addressed as a *quantification* task rather than as a classification one [60]. As a matter of fact, recent results on quantification have demonstrated the superiority of this approach over *classify-and-count* strategies [61]. Quantification tasks for obtaining aggregated estimates of the effects of moderation interventions thus represent a favorable avenue for future research. Another limitation of classification approaches is the narrow view it offers on post-moderation user behavior. As an example, in order to tackle the binary classification task, we had to come up with a definition for abandoning users. Given the inherent complexity and multifaceted nature of user behavior, we experimented with two different definitions of abandonment —*hard* and *soft*, each with its own strengths and weaknesses— so as to account for a broader spectrum of heterogeneous user behavior. Nonetheless, a binary classification task inevitably conceals important details about actual user reactions [14]. For this reason, we envision the possibility to cast the prediction of the effects of moderation interventions as a *regression* task [44]. This nuanced approach would avoid the need for limiting binary labels in favor of fine-grained estimates of the expected behavioral changes. By moving beyond simplistic labels, such an approach would empower platform moderators with deeper insights into user reactions and patterns, facilitating targeted and effective intervention strategies to foster user retention and community growth. However, solving a regression task with sufficient accuracy poses even more challenges than those addressed here. To this end, additional work on feature engineering is essential to provide the models with more informative features. Moreover, further work should also be directed towards model development, for example by employing sophisticated models capable of fully exploiting the provided features.

D. Limitations

This study focuses on a set of Reddit users who experienced a specific form of moderation intervention (i.e., multiple community bans), which may limit the generalizability of our findings. Each platform operates within a unique ecosystem with its own user demographics, interaction norms, and moderation policies. Therefore, in spite of our efforts at estimating the generalizability of our results, findings derived from a single platform and for a single intervention may not be fully applicable to other platforms or interventions. Moreover, we focused on a subset of all moderated users: those who were particularly active in the banned communities. As such, our models might exhibit decreased performance if applied to users with markedly different characteristics. Additionally, the specific definitions of hard and soft abandonment that we proposed here might represent a further limitation, given that alternative definitions may yield different interpretations of user behavior and corresponding ground truths for the machine learning tasks. Another limitation of our work stems from the set of features with which we experimented. Despite implementing a relatively large number of features that convey diverse information, many important dimensions of user behavior remain unexplored for this task. Encoding those dimensions in effective machine learning features would boost the performance of the trained models, allowing better results. By the same token, in this first study on predicting the effect of a moderation intervention, we focused on detecting abandoning users. However, effects of a moderation intervention can be defined and investigated in multiple other ways that do not necessarily involve user abandonment, but rather the degree of toxic speech on the platform after the intervention, the extent of polarization, the reliance on factual news, and more. The predictability of these and other moderation effects remains to be assessed.

X. Conclusions

We proposed and tackled —for the first time— the task of predicting the effects of a moderation intervention. Specifically, we detected those users who abandoned Reddit following the ban of a large number of communities on the platform. To solve this task we investigated the behavior of 16,540 users by leveraging 16M comments they posted in the banned communities and 13.8M comments they posted in non-banned ones. Starting from this extensive dataset, we extracted 142 features conveying information about the activity, toxicity, relations, and writing style of the analyzed users. Our results are promising, albeit preliminary as one would expect from a new task, with the best model achieving *micro F1* $= 0.800$ and *macro F1* $= 0.676$. Our model proved to be sufficiently generalizable when applied to users from unseen communities. Furthermore, we found that activity features are the most informative for this task, followed by relational and toxicity features. Conversely, writing style features seem to provide limited information. Given the novelty of the task, promising directions for future work are multifold. Among them are the inclusion of additional features that could provide relevant information on user behavior, and particularly, on their possible reactions to a moderation intervention, such as socio-psychological features. Additionally, one could cast the problem of predicting the effects of a moderation intervention as a regression task, rather than as a classification task as done in the present work. To this regard, being able to precisely estimate the behavioral changes of some users would provide even more fine-grained information to platform administrators for planning their moderation actions.

ACKNOWLEDGMENTS

This work is partially supported by the European Union, Next Generation EU, within the PRIN 2022 framework project PIANO (Personalized Interventions Against Online Toxicity); by the PNRR-M4C2 (PE00000013) "FAIR-Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded under the NextGeneration EU program; and by the Italian Ministry of Education and Research (MUR) in the framework of the FoReLab project (Departments of Excellence).

REFERENCES

[1] Y. K. Dwivedi, G. Kelly, M. Janssen, N. P. Rana, E. L. Slade, and M. Clement, "Social media: The good, the bad, and the ugly," *Information Systems Frontiers*, vol. 20, pp. 419–423, 2018.

[2] R. Kiddle, P. Törnberg, and D. Trilling, "Network toxicity analysis: An information-theoretic approach to studying the social dynamics of online toxicity," *Journal of Computational Social Science*, pp. 1–26, 2024.

[3] A. Trujillo, T. Fagni, and S. Cresci, "The DSA Transparency Database: Auditing self-reported moderation actions by social media," *arXiv:2312.10269*, 2023.

[4] R. Gorwa, R. Binns, and C. Katzenbach, "Algorithmic content moderation: Technical and political challenges in the automation of platform governance," *Big Data & Society*, vol. 7, no. 1, 2020.

[5] S. Jhaver, A. Bruckman, and E. Gilbert, "Does transparency in moderation really matter? User behavior after content removal explanations on Reddit," in *ACM CSCW*, 2019, pp. 1–27.

[6] M. Horta Ribeiro, S. Jhaver, S. Zannettou, J. Blackburn, G. Stringhini, E. De Cristofaro, and R. West, "Do platform migrations compromise content moderation? Evidence from r/The_Donald and r/Incels," in *ACM CSCW*, 2021, pp. 1–24.

[7] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "SoK: Content moderation in social media, from guidelines to enforcement, and research to practice," in *IEEE EuroS&P*, 2023, pp. 868–895.

[8] S. Zannettou, "'I Won the Election!': An empirical analysis of soft moderation interventions on Twitter," in *AAAI ICWSM*, 2021, pp. 865–876.

[9] E. Chandrasekharan, S. Jhaver, A. Bruckman, and E. Gilbert, "Quarantined! Examining the effects of a community-wide moderation intervention on Reddit," *ACM TOCHI*, vol. 29, no. 4, pp. 1–26, 2022.

[10] S. Kiesler, R. Kraut, P. Resnick, and A. Kittur, "Regulating behavior in online communities," *Building successful online communities: Evidence-based social design*, vol. 1, pp. 4–2, 2012.

[11] L. Cima, A. T. Larios, M. Avvenuti, and S. Cresci, "The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit," *arXiv:2401.11254*, 2024.

[12] J. Massachs, C. Monti, G. D. F. Morales, and F. Bonchi, "Roots of Trumpism: Homophily and social feedback in Donald Trump support on Reddit," in *ACM WebSci*, 2020, pp. 49–58.

[13] A. Trujillo and S. Cresci, "Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald," in *ACM CSCW*, 2022, pp. 1–28.

[14] ——, "One of many: Assessing user-level effects of moderation interventions on r/The_Donald," in *ACM WebSci*, 2023, p. 55–64.

[15] Q. Shen and C. P. Rosé, "A tale of two subreddits: Measuring the impacts of quarantines on political engagement on Reddit," in *AAAI ICWSM*, 2022, pp. 932–943.

[16] S. Jhaver, C. Boylston, D. Yang, and A. Bruckman, "Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter," in *ACM CSCW*, 2021, pp. 1–30.

[17] G. Etta, M. Cinelli, A. Galeazzi, C. M. Valensise, W. Quattrociocchi, and M. Conti, "Comparing the impact of social media regulations on news consumption," *IEEE Transactions on Computational Social Systems*, 2022.

[18] S. Cresci, A. Trujillo, and T. Fagni, "Personalized interventions for online moderation," in *ACM HT*, 2022, pp. 248–251.

[19] M. Trujillo, S. Rosenblatt, G. de Anda Jáuregui, E. Moog, B. P. V. Samson, L. Hébert-Dufresne, and A. M. Roth, "When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban," in *ACL WOAH*, 2021, pp. 164–178.

[20] C. R. Carlson and L. S. Cousineau, "Are you sure you want to view this community? Exploring the ethics of Reddit's quarantine practice," *Journal of Media Ethics*, vol. 35, no. 4, pp. 202–213, 2020.

[21] H. Habib and R. Nithyanand, "Exploring the magnitude and effects of media influence on Reddit moderation," in *AAAI ICWSM*, 2022, pp. 275–286.

[22] M. H. Ribeiro, S. Jhaver, M. Reignier-Tayar, R. West *et al.*, "Deplatforming norm-violating influencers on social media reduces overall online attention toward them," *arXiv:2401.01253*, 2024.

[23] M. Kurdi, N. Albadi, and S. Mishra, ""Video Unavailable": Analysis and prediction of deleted and moderated YouTube videos," in *IEEE/ACM ASONAM*, 2020, pp. 166–173.

[24] P. Paudel, J. Blackburn, E. De Cristofaro, S. Zannettou, and G. Stringhini, "Lambretta: Learning to rank for Twitter soft moderation," in *IEEE S&P*, 2023, pp. 311–326.

[25] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert, "Crossmod: A cross-community learning-based system to assist Reddit moderators," in *ACM CSCW*, 2019, pp. 1–30.

[26] M. Steiger, T. J. Bharucha, S. Venkatagiri, M. J. Riedl, and M. Lease, "The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support," in *ACM CHI*, 2021, pp. 1–14.

[27] M. Niverthi, G. Verma, and S. Kumar, "Characterizing, detecting, and predicting online ban evasion," in *ACM WWW*, 2022, pp. 2614–2623.

[28] H. Habib, M. B. Musa, F. Zaffar, and R. Nithyanand, "To act or react: Investigating proactive strategies for online community moderation," *arXiv:1906.11932*, 2019.

[29] S. Hurtado, P. Ray, and R. Marculescu, "Bot detection in Reddit political discussion," in *ACM SocialSense*, 2019, pp. 30–35.

[30] A. Smiti, "A critical overview of outlier detection methods," *Computer Science Review*, vol. 38, 2020.

[31] D. U. Ozsahin, M. Taiwo Mustapha, A. S. Mubarak, Z. Said Ameen, and B. Uzun, "Impact of feature scaling on machine learning models for the diagnosis of diabetes," in *IEEE AIE*, 2022, pp. 87–94.

[32] X. Wan, "Influence of feature scaling on convergence of gradient iterative algorithm," in *Journal of physics: Conference series*, vol. 1213, no. 3, 2019.

[33] R. E. Robertson, "Uncommon yet consequential online harms," *Journal of Online Trust and Safety*, vol. 1, no. 3, 2022.

[34] L. Hanu and Unitary team, "Detoxify," 2020.

[35] E. Bagdasaryan and V. Shmatikov, "Spinning language models: Risks of propaganda-as-a-service and countermeasures," in *IEEE S&P*, 2022, pp. 769–786.

[36] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi *et al.*, "OpenAssistant conversations – Democratizing large language model alignment," *NeurIPS*, 2024.

[37] T. H. Teng and K. D. Varathan, "Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches," *IEEE Access*, 2023.

[38] N. Ejaz, F. Razi, and S. Choudhury, "Towards comprehensive cyberbullying detection: A dataset incorporating aggressive texts, repetition, peerness, and intent to harm," *Computers in Human Behavior*, vol. 153, 2024.

[39] S. Gilda, L. Giovanini, M. Silva, and D. Oliveira, "Predicting different types of subtle toxicity in unhealthy online conversations," *Procedia Computer Science*, vol. 198, pp. 360–366, 2022.

[40] C. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *AAAI ICWSM*, 2014, pp. 216–225.

[41] M. Mazza, M. Avvenuti, S. Cresci, and M. Tesconi, "Investigating the difference between trolls, social bots, and humans on Twitter," *Computer Communications*, vol. 196, pp. 23–36, 2022.

[42] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake Twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[43] Y. Liu and S. Xu, "Detecting rumors through modeling information propagation networks in a social media environment," *IEEE Transactions on Computational Social Systems*, vol. 3, no. 2, pp. 46–62, 2016.

[44] S. Tardelli, M. Avvenuti, M. Tesconi, and S. Cresci, "Detecting inorganic financial campaigns on Twitter," *Information Systems*, vol. 103, 2022.

[45] A. Amira, A. Derhab, S. Hadjar, M. Merazka, M. G. R. Alam, and M. M. Hassan, "Detection and analysis of fake news users' communities in social media," *IEEE Transactions on Computational Social Systems*, 2023.

[46] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *IEEE ICDM*, 2008, pp. 413–422.

[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[48] T. Hasanin and T. Khoshgoftaar, "The effects of random undersampling with simulated class imbalance for big data," in *IEEE IRI*, 2018, pp. 70–79.

[49] L. Ilias and I. Roussaki, "Detecting malicious activity in Twitter using deep learning techniques," *Applied Soft Computing*, vol. 107, 2021.

[50] M. Zignani, C. Quadri, A. Galdeman, S. Gaito, and G. P. Rossi, "Mastodon content warnings: Inappropriate contents in a microblogging platform," in *AAAI ICWSM*, 2019.

[51] T.-T. Wong, "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," *Pattern Recognition*, vol. 48, no. 9, pp. 2839–2846, 2015.

[52] J. Echeverría, E. De Cristofaro, N. Kourtellis, I. Leontiadis, G. Stringhini, and S. Zhou, "LOBO: Evaluation of generalization deficiencies in Twitter bot classifiers," in *ACM ACSAC*, 2018, p. 137–146.

[53] L. Mannocci, S. Cresci, A. Monreale, A. Vakali, and M. Tesconi, "MulBot: Unsupervised bot detection based on multivariate time series," in *IEEE BigData*, 2022, pp. 1485–1494.

[54] S. Cresci, K.-C. Yang, A. Spognardi, R. Di Pietro, F. Menczer, and M. Petrocchi, "Demystifying misconceptions in social bots research," *arXiv:2303.17251*, 2024.

[55] M. Cinelli, G. De Francisci Morales, A. Galeazzi, W. Quattrociocchi, and M. Starnini, "The echo chamber effect on social media," *Proceedings of the National Academy of Sciences*, vol. 118, no. 9, 2021.

[56] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *NeurIPS*, 2017.

[57] J. Jaursch, J. Ohme, and U. Klinger, "Enabling research with publicly accessible platform data: Early DSA compliance issues and suggestions for improvement," Weizenbaum Institute, Tech. Rep., 2024.

[58] O. Varol, C. A. Davis, F. Menczer, and A. Flammini, "Feature engineering for social bot detection," in *Feature Engineering for Machine Learning and Data Analytics*, 2018, pp. 311–334.

[59] R. Raut and F. Spezzano, "Enhancing hate speech detection with user characteristics," *International Journal of Data Science and Analytics*, pp. 1–11, 2023.

[60] P. González, A. Castaño, N. V. Chawla, and J. J. D. Coz, "A review on quantification learning," *ACM CSUR*, vol. 50, no. 5, pp. 1–40, 2017.

[61] L. Milli, A. Monreale, G. Rossetti, F. Giannotti, D. Pedreschi, and F. Sebastiani, "Quantification trees," in *IEEE ICDM*, 2013, pp. 528–536.

BENEDETTA TESSA received the Master's degree in Artificial Intelligence and Data Engineering from the University of Pisa. She is currently a research fellow at the Institute of Informatics and Telematitcs (IIT) of CNR. Her research interests include social media analysis and content moderation.

LORENZO CIMA received the master's degree in computer engineering from the University of Pisa. He is a Ph.D. student in information engineering at the University of Pisa. He is also associated with the Institute of Informatics and Telematics (IIT) of CNR. His research interests include social media analysis with a focus on coordinated online behaviour and content moderation.

AMAURY TRUJILLO received his Master of Engineering degree in Information Systems from EFREI Paris. He is a Researcher at IIT-CNR, Italy, working in the areas of Social Computing and Human-Computer Interaction. His current research interests are moderation of online content on social media and emergent social phenomena around online activities.

MARCO AVVENUTI is a Full Professor of computer systems with the Department of Information Engineering, University of Pisa. He received the Ph.D. degree in information engineering from the University of Padua. He is chair of the master's degree in Artificial Intelligence and Data Engineering. His research interests include human-centric sensing and social network analysis.

STEFANO CRESCI received the Ph.D. degree in information engineering from the University of Pisa. He is a Researcher at IIT-CNR, Italy. His interests lay at the intersection of web science and data science, with a focus on content moderation and coordinated online behavior. For his achievements he received multiple awards, including an ERC grant to develop data-driven, user-centered, and personalized content moderation (DEDUCE), the ERCIM Cor Baayen Young Researcher Award, and the IEEE Next-Generation Data Scientist Award.