Who is at Risk? Analyzing the Risk of Radicalization Among Reddit Users

Ece Calikus[1] (✉), Gianmarco De Francisci Morales[2], and Aristides Gionis[3]

[1] Department of Information Technology, Uppsala University, Uppsala, Sweden
ece.calikus@it.uu.se
[2] CENTAI, Turin, Italy gdfm@acm.org
[3] Division of Theoretical Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden argioni@kth.se

Abstract. Online radicalization is a growing societal concern. Extremist groups actively exploit online media to reach wide audiences, spreading ideologies that incite hate and violence. The lack of transparency and conscious use of social media worsens this issue, as users often remain unaware of being targeted by disinformation or radical propaganda. This work analyzes the risk of online radicalization and provides insights for individuals, platforms, and policymakers to mitigate its harmful effects. We conduct a data-driven study to analyze Reddit users' radicalization risk. We build a temporal classification model using interpretable machine learning to predict the risk of radicalization with features based on RECRO, a recent social theory of Internet-mediated radicalization. Our findings reveal RECRO features are strong indicators, with features from later stages having greater influence. We also analyze risk distributions across communities, showing higher risk in controversial groups but also identifying extremists in generic and neutral communities. This result highlights the importance of critical thinking when engaging with online content.

Keywords: social media analysis · risk prediction · online safety

1 Introduction

The rise of social media and online forums has significantly changed societal discourse and inspired new forms of collective action. Despite initially being seen as a democratizing tool for information access and distribution, social media has unfortunately also fostered negative phenomena such as echo chambers, polarization, and disinformation [14,9,11]. As a result, discourse on critical topics often becomes toxic, with extremists influencing public discussions, connecting with like-minded individuals [10], and radicalizing sympathizers [29].

Individuals engage with social media from different starting points. Some already hold extreme beliefs, while others may gradually accept radical views over time through their online interactions [28]. An important concern with online platforms is the lack of transparency regarding the activities and interactions

of social media users. Many users are unaware of whether they are targeted by misinformation campaigns or trapped within echo chambers [7]. Therefore, they lack the necessary tools to employ prevention strategies before becoming deeply connected with controversial groups online.

In this paper, we conduct a data-driven analysis of radicalization risk among Reddit users and provide insights into their susceptibility to radicalization [39]. We develop a risk-prediction model that not only assesses radicalization risk but also elucidates the contributing factors.

Our approach begins by identifying relevant features associated with the risk of radicalization. To achieve this goal, we adopt the RECRO model, a social science framework for internet-mediated radicalization proposed by Neo [29]. RECRO is the acronym for the five stages of radicalization presented in this framework: *Reflection, Exploration, Connection, Resolution, and Operational*. Similar to other works [33,36], our study focuses on the first three phases of RECRO, which can be inferred from the users' online behaviors unlike the final two phases corresponding to real-world actions. The phases considered are: (*i*) *Reflection*: users' predisposition to radicalization; (*ii*) *Exploration*: the process of developing new viewpoints; and (*iii*) *Connection*: bonding with radicalized groups [29]. We hypothesize that an individual's radicalization risk evolves throughout these phases and is strongly correlated with the factors outlined within each phase. We empirically ground this theoretical framework via an exploratory analysis of Reddit data.

We consider the prediction of radicalization risk to be a high-stakes decision-making problem where inaccurate estimation can have important consequences. Misclassifying some groups of users or individuals as having a high risk of radicalization can potentially intensify societal biases, while failing to estimate radicalized groups may legitimize and reassure their behaviors. Therefore, our approach prioritizes developing a risk-prediction model that is highly intelligible and reliable.

To achieve our research goals, we use *generalized additive models* (GAMs) [18], which are effective for building inherently interpretable models [6]. A GAM is a linear combination of nonlinear submodels, each processing a single feature [46]. The contribution of each feature can be visualized, making them highly intelligible [5]. For our risk-prediction model, we use *explainable boosted machines* (EBM) [30], a GAM-based model offering accuracy comparable to more complex machine learning (ML) models, while remaining transparent.

To train our model, we annotate Reddit users as high and low risk based on their online activities. We obtain data from banned subreddits known to produce violent and hateful content and assume their users are at higher risk of radicalization. We also collect data from `/r/neutralnews` and `/r/NeutralPolitics` as proxies for low-risk individuals.

Our analysis shows that REC features are strong predictors of radicalization risk, thus supporting our hypothesis. Features from later RECRO phases have a stronger impact on predicted risk, and the average risk among high-risk users

increases over time. These findings are consistent with social theories which describe radicalization as a gradual, multi-stage process [42,29].

In addition to feature analysis, we study the distribution of risk among users in various Reddit communities. Specifically, we compare users in non-banned but controversial communities (e.g., `/r/MensRights`, `/r/AskTrumpSupporters`) similar to those in banned ones (e.g., `/r/incels`, `/r/The_Donald`), their 'antipodal' communities (e.g., `/r/AskFeminists`, `/r/hillaryclinton`), and broader topic subreddits (e.g., `/r/politics`, `/r/economics`). While users in controversial communities are at higher risk, radicalization risk is also present in more general communities.

Our contributions can be summarized as follows:

- We propose a data-driven approach inspired by a social theory to analyze the risk of radicalization of Reddit users. We open-source anonymized data and code to facilitate reproducibility, with the potential for generalizing to all Reddit users.[4]
- We develop a risk-prediction model that uses an interpretable ML method, enabling an in-depth analysis of features contributing to radicalization risk. We empirically validate the RECRO theory on large-scale real-world data by analyzing the impact of REC features on predicting radicalization risk.
- We conduct a comprehensive study of Reddit users across diverse communities, revealing that no online platform offers a foolproof "safe space," highlighting the need for vigilance in online content consumption.

2 Related Work

Radicalization theories model how individuals shift from conforming viewpoints to extreme ideologies [29]. Social sciences offer multiple theories [42,29]. For example, the 3N approach [42] models the physiological factors in radicalization using three stages: (*i*) needs (drives behind radicalization), (*ii*) narratives (legitimating extreme behaviors), and (*iii*) networks (connections with like-minded individuals). In this work, we use the RECRO model [29], a pathway-based framework comprising five phases, as discussed in the introduction, and we exclusively focus on the first three stages (REC).

Computational analysis of online Extremism. The study of radicalization on the internet has attracted research from the area of computational science. Hosseinmardi et al. [20] examine the radical content on YouTube and discover that consumption of political videos is affected by both user preferences and platform features. Papadamou et al. [32] also analyze YouTube videos to investigate whether recommendation algorithms guide users towards Incel-related content, a movement known for hateful and misogynistic views. Garimella et al. [11] study the role of partisan users and gatekeepers in political echo chambers on social media. Rollo et al. [35] develop an "attention-flow" graph to track the shifting interests of Reddit users across subreddits, identifying potential gateways to radicalized communities.

[4] https://github.com/reguluslus/RedditRiskPrediction

Relation to prior work. Some existing approaches also incorporate social theories in their models. Lerman et al. [24] use the 3N radicalization model [42] to study pro-anorexia communities on 𝕏. The most related works to ours by Phadke et al. [33] and Russo et al. [36] also incorporate the initial three stages of the RECRO model. Phadke et al. [33] study the impact of RECRO stages on conspiracy theory discussions by analyzing users in a single subreddit `/r/conspiracy`. Conversely, Russo et al. [36] operationalize RECRO parameters and analyze their impact on users migrating to a fringe platform after a community on Reddit is banned by studying two subreddits (`/r/The_Donald` and `/r/fatpeoplehate`).

Our study differs in several ways. First, previous work focuses on specific use cases (e.g., conspiracy communities and user migrations) and examines only a few subreddits. By contrast, we use RECRO to build a broader model predicting radicalization risk for all Reddit users. Second, we incorporate temporal information to track how risk and its contributing factors evolve. Third, while we use the same RECRO stages, we rely on different features, apart from some similar linguistic features during reflection. Finally, our experiments address distinct research questions: "Do later RECRO stages influence radicalization risk more?" and "Is risk prevalent outside controversial communities?".

3 Data

We download historical Reddit data from the PushShift repository [2], comprised of comments and posts from 1 January 2016 to 31 December 2020 taken from the most popular 20k subreddits. Reddit is organized into subreddits: forums for specific topics [27]. Users can join subreddits to share and receive information, while whether a user subscribes to a subreddit is not public. We can only collect data for users who have posted or commented on a subreddit. For our risk-prediction model, we categorize Reddit users as "high-risk" and "low-risk" cohorts based on the subreddits they actively participated in.

Over the years, numerous controversial subreddits have emerged to spread radical ideologies and hateful content [16,35]. Reddit introduced quarantining for subreddits violating their policies, eventually banning persistent violators [37]. In our study, we use these banned subreddits as proxies for communities with high radicalization risk. To identify banned subreddits, we scrape the homepages of all subreddits in the PushShift dataset [2] and send requests to each URL. If a subreddit is banned, the request returns a ban warning message. We retain only those banned for radicalization-related reasons (e.g., "promoting hate" or "inciting violence") and exclude subreddits banned for other reasons such as "graphic content" or "copyright violations"[5]. Ultimately, we retain 115 subreddits banned between Jan 1, 2016, and Dec 31, 2020. Following [27], users with at least five posts or 25 comments in a banned subreddit are labeled "high-risk."

For the "low-risk'" cohort, we select users from neutral communities (`/r/neutralnews` and `/r/NeutralPolitics`), focused on polite, empirical discussions of news and politics. Despite their intent, some users in these communities

[5] https://redditinc.com/policies/content-policy

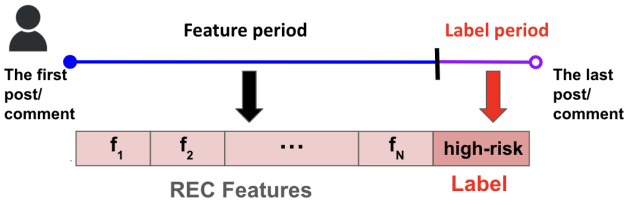

Fig. 1: Feature and label periods on user journeys.

are also active in banned ones. To ensure dataset integrity, we exclude such users and include the remaining neutral community users in the "low-risk'" cohort, applying the same cutoff as for "high-risk'" users.

We consider users' entire Reddit history including all the subreddits they participated in. Each user's historical data is split into an observation (feature) period and a subsequent label period (Figure 1) [21] . Users posted or commented in banned subreddits during the label period are labeled "high-risk," others as "low-risk." This separation lets us predict future participation in banned subreddits, identifying users particularly at risk. Notably, even previously "high-risk" users are labeled "low-risk" if they never participated in a banned community during the label period. Consequently, risk scores are higher for users likely to engage in banned communities soon, and lower for those who never do or cease participating.

4 Measuring Radicalization Features

4.1 Reflection

The "reflection" phase is described as "the triggers, needs, and vulnerabilities that an individual may have, which increase one's susceptibility towards alternative belief systems" [29]. Personal traits, issues, and social life can shape one's openness to exploring new online viewpoints, including those supporting violent extremism. In this study, we use several language markers in Reddit posts and comments as proxies for psychological states and personality traits to analyze the reflection phase.

Linguistic features. Psycho-linguistic features in users' online behavior, such as comments expressing anger, anxiety, and heightened emotions are markers of radicalization [33,38,16].

To extract linguistic features, we use Empath [8], an open-source tool that analyzes text through lexical categories. We include Empath scores for four prebuilt categories: *anger, hate, nervousness, and positive emotion.* We calculate the average empath score for user u and a given category α as

$$R^{\alpha}_{\text{emp}}(u) = \frac{\sum_{i=1}^{m} e(p_i) + \sum_{j=1}^{n} e(c_j)}{m+n}, \tag{1}$$

where $e(p_i)$ and $e(c_j)$ are the empath scores of the i^{th} post and j^{th} comment, and m and n are the number of posts and comments.

We expect all factors except "positive emotion" to positively correlate with radicalization risk scores. Conversely, "positive emotion" is anticipated to negatively correlate, indicating that higher-risk users show less positivity in their Reddit posts and comments.

Moral foundations. Moral judgments have shown to be important features for understanding political polarization [44], group formation [13], and radicalization [1]. We analyze the moral content in the text using the extended Moral Foundations Dictionary (eMFD) [19], which includes five core aspects of human morality according to Moral Foundations Theory [17,12]: *care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation.*

For a given text, the eMFD tool provides a moral foundation scoring in each category which suggests the probability of the text including annotations in that category [19]. To characterize the reflection phase of the use u, we compute the moral foundation scores in each category β as

$$R^{\beta}_{\text{moral}}(u) = \frac{\sum_{i=1}^{m} f(p_i) + \sum_{j=1}^{n} f(c_j)}{m+n}, \tag{2}$$

where $f(p_i)$ and $f(c_j)$ are the moral foundation scores of the i^{th} post and j^{th} comment, and m and n are the number of posts and comments.

4.2 Exploration

In RECRO, the "exploration" phase is when individuals search online for alternative belief systems, exposing them to radical narratives and new worldviews [29]. We posit that the spread of radical ideas often relies on disinformation, thereby increasing individuals' vulnerability to exposure to misinformation. Studies also link conspiracy theories to extremism, suggesting conspiracy communities can serve as gateways to radicalization [35,33].

To incorporate misinformation and conspiracy features, we use Media Bias Fact Check (MBFC),[6] an independent organization rating media sources on political bias, factual reporting, and conspiracy levels. MBFC, the largest online news credibility evaluator, covers about 5300 web pages [4,43]. We crawl MBFC to collect factuality and conspiracy scores for news sources.

Exposure to misinformation. We quantify exposure to misinformation by estimating the misinformation content in subreddits that the user engages with. If a user posts or comments in a subreddit, we assume they are likely exposed to its content, either directly through subscription or indirectly through algorithmic recommendations.

We adopt an approach similar to the one presented by [43]. We gather data from MBFC's Web pages, which are organized into various categories such as "left bias," "right bias," "questionable sources," and "conspiracy-pseudoscience." We collect the names, URLs, and factuality ratings of each news source. MBFC rates factual reporting on a 6-point Likert scale, from "very low" to "very high."

[6] https://mediabiasfactcheck.com

We convert these categorical ratings to numerical values in the range $[-2, 3]$, where "very low" corresponds to -2 and "very high" to 3. We calculate the factuality score for a subreddit as the average score of all links shared there, excluding unrated links. Finally, we quantify the exposure to misinformation for a user, denoted as u, as follows:

$$E_{\text{mis}}(u) = -\frac{\sum_{i=1}^{k} f(s_i)}{k}, \tag{3}$$

where $f(s_i)$ is the factuality score of the i^{th} subreddit and k is the number of subreddits that user u has posted or commented.

Seeking conspiracy. We measure exposure to conspiracy-related content similarly to misinformation. We extract and annotate the news sources categorized under "Conspiracy-Pseudoscience" on the MBFC website. Additionally, we review sources in the "Questionable Source" category, annotating those with "conspiracy" as well. A subreddit's conspiracy score is the ratio of conspiracy-labeled links to all MBFC-categorized links posted in that subreddit. Finally, we quantify the exposure to misinformation for user u as

$$E_{\text{consp}}(u) = \frac{\sum_{i=1}^{k} c(s_i)}{k}, \tag{4}$$

where $c(s_i)$ is the conspiracy score of the i^{th} subreddit and k is the total number of subreddits where user u has written.

4.3 Connection

During the "connection" phase, individuals deepen their radical perspective by interacting with like-minded communities [29]. We assume users who frequently post or comment in Reddit communities with hateful, offensive, or discriminatory discussions are more likely to bond with and adopt the views of those groups.

To identify problematic subreddits, one approach might involve using banned communities or those with similar discussions, often moderated for reasons like "promoting hate" or "inciting violence." However, since our training data for high-risk users comes exclusively from banned communities, and low-risk users never participate in them, using similar communities to characterize the connection phase risks data leakage.

Instead, we measure the language toxicity in the community to analyze whether its users frequently include abusive and harmful content in their posts and discussions. For this, we use Google's Perspective API,[7] a widely adopted tool for toxicity assessment [23,36]. The API provides scores (0 to 1) for attributes like "toxicity," "insult," "threat," and "sexually explicit." We focus on the "toxicity," attribute, defining a subreddit's toxicity level, $t(s_i)$, as the average toxicity score of all its posts and comments.

[7] https://perspectiveapi.com

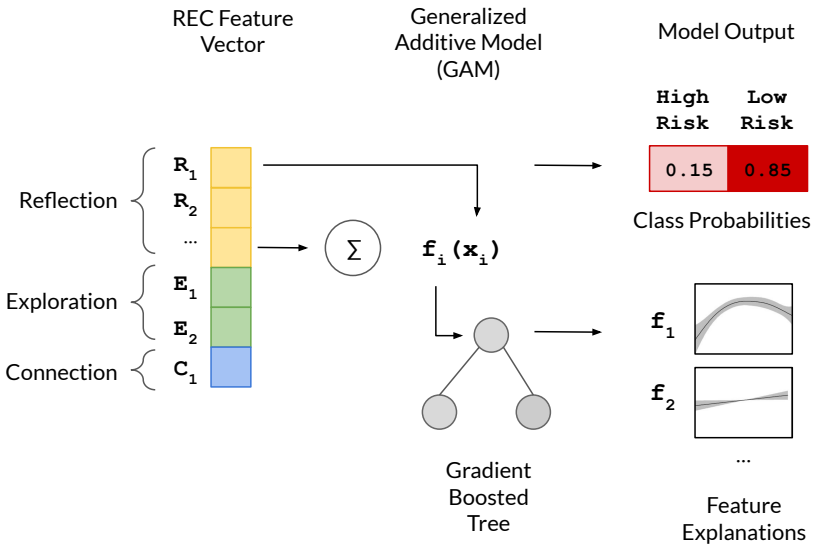

Fig. 2: Illustration of the risk-prediction model. It takes REC features as input, producing risk probabilities and feature attributions (adapted from [45]).

$$t(s_i) = \frac{\sum_{i=1}^{m} t(p_i) + \sum_{j=1}^{n} t(c_j)}{m+n}, \tag{5}$$

where $t(p_i)$ and $t(c_i)$ are the toxicity scores of the i^{th} post and j^{th} comment and m and n are the total number of posts and comments published in subreddit s_i.

Finally, the engagement score of a given user u, is computed as the average number of posts and comments across the list of subreddits in which u participated, with each subreddit being weighted with its toxicity score

$$C_{\text{eng}}(u) = \frac{\sum_{i=1}^{k} t(s_i)(m_i + n_i)}{k}, \tag{6}$$

where $t(s_i)$ is the toxicity score of the i^{th} subreddit, m_i and n_i are the number of posts and comments made by u in s_i respectively, and k is the total number of subreddits in which u has been active.

While Russo et al. [36] use user-level "language toxicity" to operationalize the reflection phase, we argue that toxic language does not inherently indicate vulnerability to radicalization. Instead, individuals adopt more toxic behavior after joining radical communities. Therefore, we focus on community-level toxicity to spot problematic subreddits likely to be radicalizing, thereby capturing the connection phase.

5 Risk-Prediction Model

Generalized additive models (GAMs) [18] are a category of models that offer greater flexibility and efficacy compared to linear models while still being more

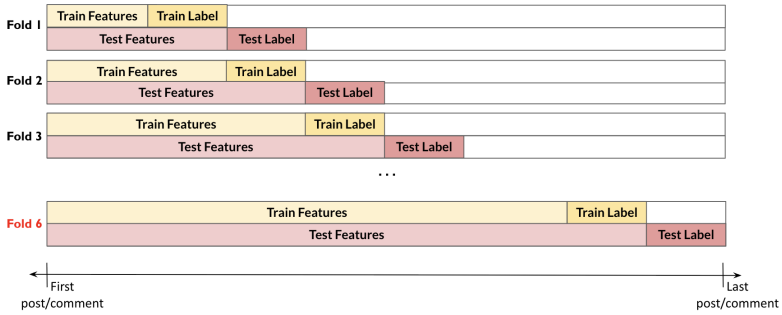

Fig. 3: Temporal cross-validation procedure.

interpretable than complex models such as neural networks [46]. A GAM models the target variable as the sum of non-linear functions of single input features [46]. The impact of each term on the final prediction can be visualized by plotting the target variable as the output of each function and the given term [6].

In particular, we employ Explainable Boosting Machines (EBMs) [30], which are based on the GA^2M algorithm [25,5], and belong to the family of GAMs of the following form

$$g(E[y]) = \beta_0 + \sum f_j(x_j) + \sum f_{ij}(x_i, x_j), \tag{7}$$

where $g(\cdot)$ is the link function depending on the ML task, x_j is the j^{th} feature, f_j is the shape function for feature x_j, and f_{ij} is a function for pairwise interactions between x_i and x_j. Each input feature x_j is a REC feature (e.g., R_{tox} or E_{mis}) presented in Section 4. The shape function f_j is learned by using gradient-boosted ensembles of bagged trees. Since we formulate our problem as a binary classification model that predicts users' risk of radicalization as "high risk" or "low risk," the link function $g(\cdot)$ is a logit (inverse of the logistic function).

Figure 2 shows a depiction of the model. EBM improves on traditional GAM by capturing feature interactions and employing tree-based ensemble learning. It achieves performance on par with other tree-based boosting methods [30] while remaining highly interpretable by decomposing its prediction into each feature's contribution. We use the EBM implementation provided by InterpretML [30].

6 Results

This section presents results from the collected data and risk-prediction model. We first show performance comparisons between the EBM and other ML models. Next, we analyze how radicalization features influence user risk and their evolving importance over time. Finally, we compare risk distributions across communities with differing viewpoints.

6.1 Predictive Performance

We frame our problem as a temporal binary classification task to predict users' radicalization risk. To accurately capture temporal relations, we use a temporal

cross-validation procedure [22,34,40]. This involves dividing the data into six temporal folds (see Figure 3). Each fold contains a training and testing set, with the training window always preceding the testing window, ensuring that future information doesn't influence the prediction of past. Each training and testing split also has consecutive feature and label periods (explained in Section 3). REC features are generated from user posts and comments during the feature period, while users are labeled "high-risk" or "low-risk" based on their posting or commenting activity in banned subreddits during the label period. This setup allows for early prediction of radicalization risk.

As stated by Neo [29], the stages of RECRO are not distinctly separated and can occur simultaneously. Thus, all 12 features characterizing the REC phases are included in each fold. Reddit users in our data do not have uniform timelines regarding the first and last dates of their activity, total duration, and posting frequency. Therefore, we split the temporal folds based on the percentage of total time spent on Reddit, rather than specific dates. This approach allows us to track how risk changes for each user over their timeline, instead of focusing on specific calendar times.

For final predictions (Table 1) and feature analysis (Figure 8), the first five folds are used to select the best parameters, while the last fold is held out to report final results.

Figure 4 shows the temporal-cross-validation results from EBM and three other classification models (Random Forest, XGBoost, and Logistic Regression) using the same set of REC features. Each fold shows the average ROC-AUC scores obtained from each hyperparameter tuned for that model. Additionally, Table 1 shows the final prediction performance on four metrics: ROC-AUC, accuracy, precision, and recall. In terms of prediction performance, EBM exhibits performance on par with state-of-the-art ML methods such as Random Forest and XGBoost. This result aligns with observations in previous studies [30,15]. All methods achieve a high classification performance, thus, showing that it is possible to automatically distinguish Reddit users at higher radicalization risk from the ones at lower risk over time by using the REC features.

Figure 5 shows how risks progress over time among different types of users. This figure presents the average risk scores for users from high- and low-risk cohorts, as described in Section 3. The risk scores of users in the high-risk cohort steadily increase on average, while they remain relatively stable among low-risk users. Note that users within the same cohort do not exhibit identical activity

Table 1: Risk prediction performance of different ML models.

Model	ROC-AUC	Accuracy	Precision	Recall
EBM	0.8179	0.8186	0.6223	0.8165
XGBoost	0.7978	0.7892	0.5741	0.8165
Random Forest	0.7649	0.7794	0.5673	0.7339
Logistic Regression	0.7846	0.8039	0.6090	0.7431

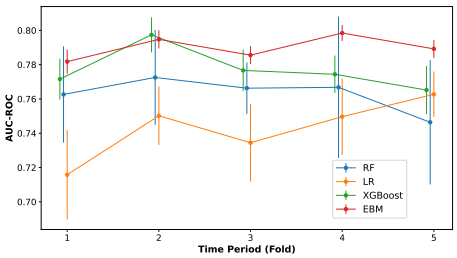

Fig. 4: Temporal cross-validation results for each method

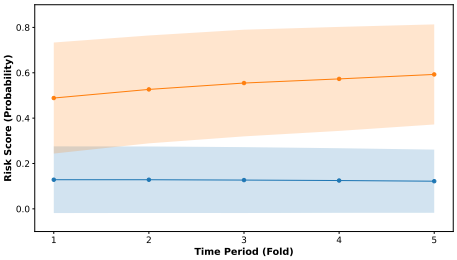

Fig. 5: Average risk over time for high and low-risk cohorts.

on Reddit, and we cannot assume they follow the stages of RECRO in the same manner. Some users may enter problematic communities at much earlier stages and then cease engaging with these communities, while others may experience a slower progression but become more deeply connected with these communities [26]. Nevertheless, the overall trend indicates that the risk of radicalization progressively advances on average during users' online journeys, aligning with the prevailing social theories about online radicalization [42,29].

6.2 Feature Analysis

The additive property of the EBM model enables measuring each input feature's contribution to the prediction. Each term $f(x_i)$ in the model returns a score (i.e., a log odds ratio), which is subsequently added to the users' predicted risk [5]. More specifically, $f(x_i) > 0$ implies that x_i contributes to a positive label (high risk), while $f(x_i) < 0$ implies that x_i contributes to a negative label (low risk).

Figure 6 displays the importance of REC features in models trained with temporal cross-validation. The y-axis shows the mean absolute value of feature importance scores across the training set [5,30], computed by scoring each data point using one feature at a time and averaging the absolute values. The figure illustrates how importance scores for different REC phases evolve over time.

The most influential feature is typically a connection phase feature measuring engagement with toxic communities, closely followed by an exploration feature quantifying exposure to misinformation. Since users experience REC phases simultaneously, some engage with banned communities early and are labeled high-risk, which may explain why the connection feature remains influential even in the early stages. In contrast, conspiracy seeking, the second exploration feature, behaves differently: it ranks third in importance during the first three periods but drops significantly to fifth by the end.

Initially, feature importance scores are closer to each other across all phases, but over time, the gap between the connection and reflection phase features widens. Combined with Figure 5 showing an increasing trend among risk scores over time, this finding provides empirical support for RECRO, postulating that Internet-mediated radicalization is a process evolving over successive phases [29].

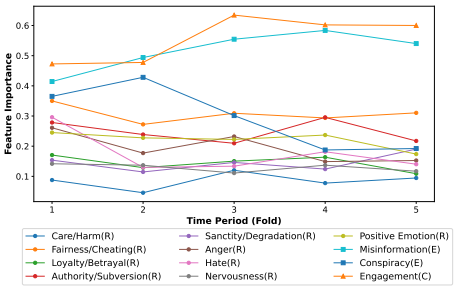

Fig. 6: Evolution of feature importance scores for each REC feature.

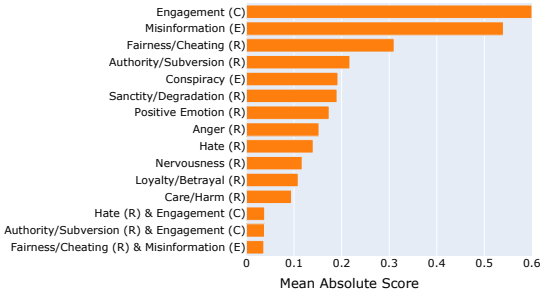

Fig. 7: Feature importance table of the final EBM model.

Figure 7 presents the ranking of REC features and their pairwise interactions in the final prediction model based on their importance. Engagement and misinformation remain the most influential features, followed by two morality features (fairness/cheating and authority/subversion), characterizing the reflection phase. Linguistic features, such as anger, hate, and nervousness, contribute significantly less to radicalization risk in comparison. This is notable as many previous studies have emphasized language signals as strong indicators of problematic behavior and have widely used them to analyze radicalization [31,38].

We further analyze the effect of REC features on risk by plotting their shape functions in Figure 8. The y-axis delineates the scores (log odds ratios) predicted by the model, while the x-axis shows the values of the input REC feature [5]. The error bars represent the standard deviation of the risk score, determined via 100 rounds of bagging [5,30].

The first graph shows the risk scores concerning the connection feature (i.e., engagement), which stands out as the most influential feature in our model. It exhibits a near-logistic growth, with the growth rate peaking between the feature values of 0.11 and 0.17. That is, engagement scores lower and higher than these values push predictions toward low and high-risk classes, respectively.

Misinformation, as an exploration feature and the second most important factor, exhibits a mostly positive trend. This confirms our assumption that exposure to misinformation is significantly associated with radicalization. Such a finding should urge online platforms to enhance their tools and content moderation strategies to limit exposure to misinformation and fake news, thereby mitigating the progression of radicalization. Similarly, the conspiracy shape function, overall suggests an elevated risk for users who explore content from conspiracy-related subreddits and has a positive impact on the prediction target.

Most contributing reflection terms—fairness/cheating, authority/subversion, and positive emotions—exhibit inverse correlation with risk scores, showcasing high-risk users use less fairness and authority-related vocabularies and express less positive emotion in their posts and comments. Remarkably, other reflection terms such as loyalty (i.e., "us versus them" thinking) [19], care/harm, anger, hate, and nervousness, known to commonly permeate radical discourse in offline

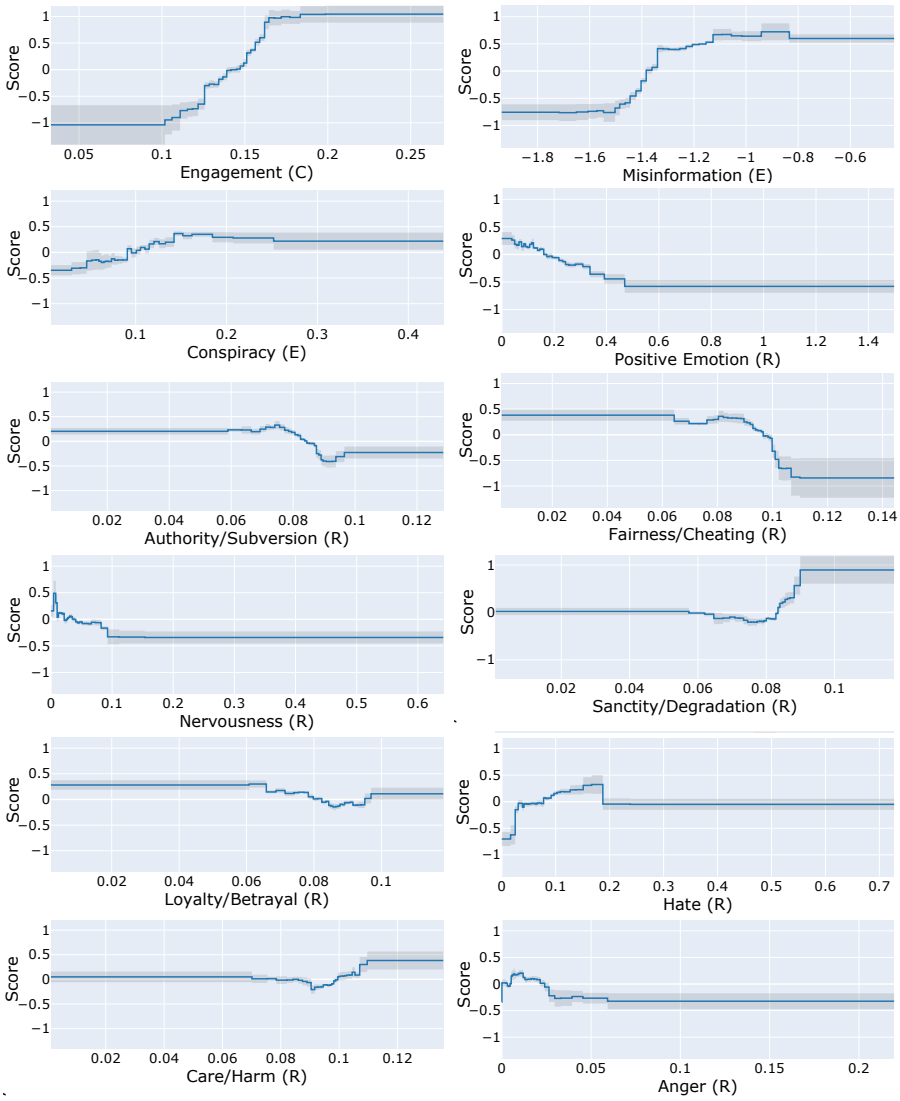

Fig. 8: Shape functions learned by the EBM model that each corresponding a REC feature.

contexts [38,16], have lower impact on prediction results. Their respective shape plots reveal markedly low and frequently inconsistent effects in the logit space.

Note that all these linguistic features exhibit highly skewed distributions, with mean values below 0.1. Both the Empath and eMFD models use word frequencies for specific categories, normalized against a large volume of posts and comments, leading to small values. More advanced language models might yield different results.

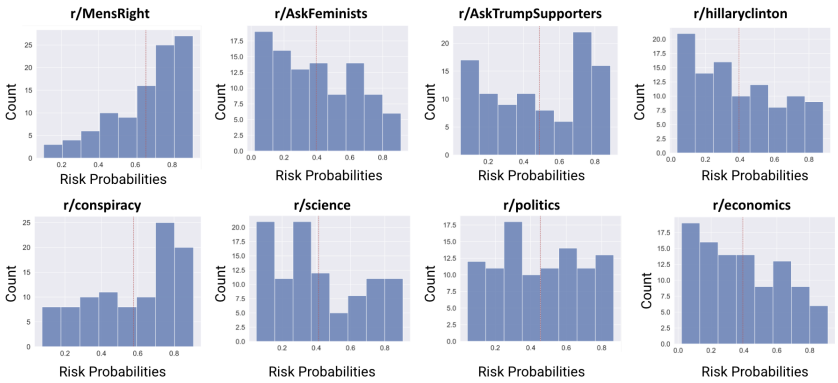

Fig. 9: Distributions of estimated risk probabilities across subreddits. Even on mainstream subreddits, there is a significant portion of users at high risk.

6.3 Subreddit Analysis

As an application of our model, we compute the risk predictions for users engaged in several communities. We select three active subreddits (`/r/MensRights`, `/r/AskTrumpSupporters`, and `/r/conspiracy`) known for controversial opinions, resembling their banned counterparts (`/r/incels`, `/r/The_Donald`, and `/r/thegreatawakening`). Additionally, we choose `/r/AskFeminists`, `/r/hillaryclinton`, and `/r/science` as contrasting communities, informed by previous literature [41]. We also include `/r/politics` and `/r/economics` for more general discussions on similar topics with a diverse audience.

We randomly select 100 users from each subreddit among the 1000 most active users and predict their risk scores with our model. Figure 9 shows the distribution of their risk probabilities. The average risk is considerably higher among users in controversial communities compared to their antipodes or more general communities. However, a substantial number of users in communities such as `/r/AskFeminists`, `/r/hillaryclinton`, and `/r/politics`, are also at risk. This indicates that extreme views are not exclusive to banned or overtly controversial communities, but can be prevalent even in "neutral" or "generic" ones such as `/r/science` and `/r/economics`. These findings emphasize the importance of critical thinking when consuming online content and engaging with users.

7 Conclusion

In this paper, we studied the radicalization risk of Reddit users by combining a data-driven approach with a social theory framework, RECRO. We developed a risk-prediction model using a GAM-based machine-learning method, enabling detailed analysis of features contributing to risk. The model was trained on historical data, with users from banned and neutral communities selected and annotated using a temporal validation design.

Our study analyzed how features within the RECRO framework impact radicalization risk over time. Features from later RECRO phases have stronger

predictive power, particularly during later periods of users' online journeys. We also examined risk in communities similar to banned ones, their opposite groups, and general topic subreddits. Results showed many users in controversial communities are at high risk, and even generic communities have notable numbers of at-risk users, highlighting the importance of critical thinking online.

Limitations. Radicalization risk is subjective and hard to quantify. We use banned communities as proxies for heightened risk, based on frequent violations of Reddit's rules against inciting violence or hate. However, "low risk" does not rule out radical beliefs, and "high risk" does not confirm actual radicalization.

We carefully prevent data leakage, ensuring none of the REC features rely on the target variable or future information. Still, the model relies on curated data from clearly distinct communities, and real-world risk distributions may not be so bimodal. Furthermore, radicalization can be a multifactorial process [3]. We frame it as a binary classification to leverage ML models for risk prediction. Future research should refine this approach.

Ethical considerations. This research addresses the sensitive subjects of radicalization and online extremism, requiring careful use of ML models and interpretation of results. Our goal is to identify key factors that increase susceptibility to radicalization risk on online platforms and draw attention to safety and awareness. We strive to ensure our research outcomes are accurate, transparent, and framed to avoid perpetuating harmful narratives or stereotypes. Our dataset comes from the PushShift archive [2], a widely used public resource in over a hundred peer-reviewed studies. No personally identifiable information is used, and all analyses are aggregated to ensure user anonymity. This study involved no human interventions and required no IRB review.

Acknowledgments. This research was funded by the ERC Advanced Grant REBOUND (834862), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

1. Alizadeh, M., Weber, I., Cioffi-Revilla, C., Fortunato, S., Macy, M.: Psychology and morality of political extremists: evidence from Twitter language analysis of alt-right and antifa. EPJ Data Science **8**(1), 1–35 (2019)
2. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., Blackburn, J.: The pushshift reddit dataset. In: ICWSM. vol. 14, pp. 830–839 (2020)
3. Berger, J.M.: Extremism. MIT Press (2018)
4. Bozarth, L., Saraf, A., Budak, C.: Higher ground? how groundtruth labeling impacts our understanding of fake news about the 2016 us presidential nominees. In: ICWSM. vol. 14, pp. 48–59 (2020)
5. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: KDD. pp. 1721–1730 (2015)
6. Chang, C.H., Tan, S., Lengerich, B., Goldenberg, A., Caruana, R.: How interpretable and trustworthy are gams? In: KDD. pp. 95–105 (2021)

7. Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The Echo Chamber Effect on Social Media. PNAS **118**(9), e2023301118 (2021)
8. Fast, E., Chen, B., Bernstein, M.: Empath: Understanding Topic Signals in Large-Scale Text. In: CHI. pp. 4647–4657 (May 2016)
9. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Quantifying Controversy in Social Media. In: WSDM (2016)
10. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Reducing Controversy by Connecting Opposing Views. In: WSDM (2017)
11. Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: WWW (2018)
12. Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S.P., Ditto, P.H.: Moral foundations theory: The pragmatic validity of moral pluralism. In: Adv. Exp. Soc. Psychol., vol. 47, pp. 55–130. Elsevier (2013)
13. Graham, J., Haidt, J., Nosek, B.A.: Liberals and conservatives rely on different sets of moral foundations. J. Pers. Soc. Psychol. **96**(5), 1029 (2009)
14. Guess, A.M., Nyhan, B., Reifler, J.: Exposure to untrustworthy websites in the 2016 us election. Nature human behaviour **4**(5), 472–480 (2020)
15. Guldogan, E., Yagin, F.H., Pinar, A., Colak, C., Kadry, S., Kim, J.: A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris. Sci. Rep. **13**(1), 22189 (2023)
16. Habib, H., Srinivasan, P., Nithyanand, R.: Making a radical misogynist: How online social engagement with the manosphere influences traits of radicalization. CSCW **6**(the CSCW2), 1–28 (2022)
17. Haidt, J.: The new synthesis in moral psychology. science **316**(5827), 998–1002 (2007)
18. Hastie, T., Tibshirani, R.: Generalized Additive Models: some applications. J. Amer. Statist. Assoc. **82**(398) (1987)
19. Hopp, F.R., Fisher, J.T., Cornell, D., Huskey, R., Weber, R.: The extended moral foundations dictionary (emfd): Development and applications of a crowd-sourced approach to extracting moral intuitions from text. Behav. Res. Methods. **53**, 232–246 (2021)
20. Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D., Watts, D.: Examining the consumption of radical content on YouTube. PNAS **118**(32) (2021)
21. Hyndman, R.J., Athanasopoulos, G.: Forecasting: principles and practice. OTexts (2018)
22. Kumar, A., Rizvi, S.A.A., Brooks, B., Vanderveld, R.A., Wilson, K.H., Kenney, C., Edelstein, S., Finch, A., Maxwell, A., Zuckerbraun, J., et al.: Using machine learning to assess the risk of and prevent water main breaks. In: KDD. pp. 472–480 (2018)
23. Kumar, D., Hancock, J., Thomas, K., Durumeric, Z.: Understanding the behaviors of toxic accounts on Reddit. In: WebConf. pp. 2797–2807 (2023)
24. Lerman, K., Karnati, A., Zhou, S., Chen, S., Kumar, S., He, Z., Yau, J., Horn, A.: Radicalized by thinness: Using a model of radicalization to understand pro-anorexia communities on Twitter. arXiv:2305.11316 (2023)
25. Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: KDD. pp. 623–631 (2013)
26. Monti, C., Aiello, L. M., De Francisci Morales, G, Bonchi, F.: The language of opinion change on social media under the lens of communicative action. Sci. Rep. **12**, 17920 (2022)

27. Monti, C., D'Ignazi, J., Starnini, M., De Francisci Morales, G.: Evidence of demographic rather than ideological segregation in news discussion on Reddit. In: WebConf. pp. 2777–2786 (2023)
28. Necaise, A., Williams, A., Vrzakova, H., Amon, M.J.: Regularity versus novelty of users' multimodal comment patterns and dynamics as markers of social media radicalization. In: Hypertext. pp. 237–243 (2021)
29. Neo, L.S.: An internet-mediated pathway for online radicalisation: RECRO. In: Violent Extremism: Breakthroughs in Research and Practice (2019)
30. Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
31. Nouh, M., Nurse, J.R., Goldsmith, M.: Understanding the radical mind: Identifying signals to detect extremist content on Twitter. In: IEEE ISI. pp. 98–103. IEEE (2019)
32. Papadamou, K., et al.: "How over is it?" Understanding the incel community on YouTube. CSCW **5** (2021)
33. Phadke, S., Samory, M., Mitra, T.: Pathways through conspiracy: the evolution of conspiracy radicalization through engagement in online conspiracy discussions. In: ICWSM. vol. 16, pp. 770–781 (2022)
34. Ramachandran, A., Kumar, A., Koenig, H., De Unanue, A., Sung, C., Walsh, J., Schneider, J., Ghani, R., Ridgway, J.P.: Predictive analytics for retention in care in an urban hiv clinic. Sci. Rep. **10**(1), 6421 (2020)
35. Rollo, C., De Francisci Morales, G., Monti, C., Panisson, A.: Communities, gateways, and bridges: Measuring attention flow in the reddit political sphere. In: SocInfo (2022)
36. Russo, G., Horta Ribeiro, M., Casiraghi, G., Verginer, L.: Understanding online migration decisions following the banning of radical communities. In: WebSci (2023)
37. Russo, G., Verginer, L., Ribeiro, M.H., Casiraghi, G.: Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. In: ICWSM. vol. 17, pp. 742–753 (2023)
38. Shrestha, A., Kaati, L., Cohen, K.: Extreme adopters in digital communities. J. Threat Assess. Manag **7**(1-2), 72 (2020)
39. Thompson, R.: Radicalization and the use of social media. J. Strateg. Secur. **4**(4), 167–190 (2011)
40. Vajiac, C., Frey, A., Baumann, J., Smith, A., Amarasinghe, K., Lai, A., Rodolfa, K.T., Ghani, R.: Preventing eviction-caused homelessness through ml-informed distribution of rental assistance. In: AAAI. vol. 38, pp. 22393–22400 (2024)
41. Waller, I., Anderson, A.: Quantifying social organization and political polarization in online platforms. Nature **600**(7888), 264–268 (2021)
42. Webber, D., Kruglanski, A.W.: Psychological factors in radicalization: A "3n" approach. The handbook of the criminology of terrorism pp. 33–46 (2016)
43. Weld, G., Glenski, M., Althoff, T.: Political bias and factualness in news sharing across more than 100,000 online communities. In: ICWSM. vol. 15, pp. 796–807 (2021)
44. Wolsko, C., Ariceaga, H., Seiden, J.: Red, white, and blue enough to be green: Effects of moral framing on climate change attitudes and conservation behaviors. J. Exp. Soc. Psychol **65**, 7–19 (2016)
45. Xenopoulos, P., Freeman, W.R., Silva, C.: Analyzing the differences between professional and amateur esports through win probability. In: WebConf. pp. 3418–3427 (2022)

46. Zhuang, H., Wang, X., Bendersky, M., Grushetsky, A., Wu, Y., Mitrichev, P., Sterling, E., Bell, N., Ravina, W., Qian, H.: Interpretable ranking with generalized additive models. In: WSDM. pp. 499–507 (2021)