Quantifying Feature Importance for Online Content Moderation

Benedetta Tessa*[a,b], Alejandro Moreo*[c], Stefano Cresci[b], Tiziano Fagni[b], Fabrizio Sebastiani[c]

[a] *Dipartimento di Informatica, Università di Pisa, Largo Bruno Pontecorvo 3, 56127, Pisa, Italy*
[b] *Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi 1, 56124, Pisa, Italy*
[c] *Istituto di Scienza e Tecnologie dell'Informazione, Consiglio Nazionale delle Ricerche, Via Giuseppe Moruzzi 1, 56124, Pisa, Italy*

Abstract

Accurately estimating how users respond to moderation interventions is paramount for developing effective and user-centred moderation strategies. However, this requires a clear understanding of which user characteristics are associated with different behavioural responses, which is the goal of this work. We investigate the informativeness of 753 socio-behavioural, linguistic, relational, and psychological features, in predicting the behavioural changes of 16.8K users affected by a major moderation intervention on Reddit. To reach this goal, we frame the problem in terms of "quantification", a task well-suited to estimating shifts in aggregate user behaviour. We then apply a greedy feature selection strategy with the double goal of *(i)* identifying the features that are most predictive of changes in user activity, toxicity, and participation diversity, and *(ii)* estimating their importance. Our results allow identifying a small set of features that are consistently informative across all tasks, and determining that many others are either task-specific or of limited utility altogether. We also find that predictive performance varies according to the task, with changes in activity and toxicity being easier to estimate than changes in diversity. Overall, our results pave the way for the development of accurate systems that predict user reactions to moderation interventions. Furthermore, our findings highlight the complexity of post-moderation user behaviour, and indicate that effective moderation should be tailored not only to user traits but also to the specific objective of the intervention.

Keywords: Content moderation, user behaviour, quantification, feature selection, feature importance

1. Introduction

The need for effective content moderation has become increasingly important for online platforms, in order to ensure the safety and well-being of online communities [28]. Without

*These authors contributed equally to this work.

moderation, problematic phenomena such as the spread of mis- and disinformation [5, 24], political manipulation [17, 63], toxicity [1, 62], hate speech [29, 67], and coordinated harmful behaviour [13, 52], risk threatening the integrity of the online discourse. In addition, platforms are legally required to implement moderation mechanisms in order to detect and mitigate certain types of harmful content [75, 81, 85]. To moderate content and behaviours, platforms adopt a wide range of interventions that depend on the context and severity of the violation. These range from limiting the visibility of a post [12] to permanently removing a user from the platform [45]. While these interventions are generally intended to restore the integrity of the online ecosystems, their effectiveness is not always guaranteed. Recent research has highlighted that some interventions may be ineffective, or even lead to unintended consequences, such as user churn, migration to less regulated platforms, or even an increase in toxicity [15, 36, 41, 83].

One of the factors that currently limits the effectiveness of content moderation is the *one-size-fits-all* approach to the deployment of moderation interventions, which do not account for individual user differences [18]. Indeed, recent studies showed that different users respond in different ways to the same intervention, which reduces the overall effectiveness of user-independent moderation [84]. A crucial step towards user-centred or personalized moderation is the identification of those user characteristics that are responsible for the different reactions to moderation interventions. Therefore, in this work we aim to shed light on the main drivers behind behavioural change that follows a moderation intervention. A deeper understanding of what drives users to react in certain ways can pave the way to the development of personalized interventions, thus enhancing the overall efficiency and effectiveness of the content moderation process [18]. To this end, we explore different behavioural and cognitive characteristics of a large set of social media users that were affected by a massive moderation intervention on Reddit. Among the features we consider are user activity, toxicity, writing style, emotional tone, and psychological traits. In order to assess which features are most relevant, we adopt an approach based on machine learning (ML). By training models and assessing the relative importance of different feature groups in predicting behavioural change after the intervention, we identify the most informative such groups.

Technically, our problem is complex due to the fact that the *IID assumption*,[1] on which essentially all supervised ML methods are based, is not verified here. This is because, in the context of online moderation, the set of data on which models are trained and the set of data to which the trained models are then applied are typically distributionally different [83, 84]. In these cases, the two distributions from which the two sets of data are sampled are said to be characterized by *dataset shift* [56, 68, 79]. The strategy we adopt thus consists of recasting our problem in terms of *quantification*, the supervised learning task of predicting the prevalence of the classes in a set of unlabelled datapoints characterized by (some kind of) dataset shift [25]. It is well known that solving this task by classifying all the unlabelled datapoints and checking how many have been assigned to each class (the "classify and count" method) is a naïve, suboptimal quantification method [20], since standard methods for training classifiers are based on the IID assumption. Therefore, we adopt a strategy

[1]The assumption that the training data and the test data are "identically and independently distributed."

based on quantification methods because our problem can indeed be formulated in terms of class prevalence prediction, and because current quantification algorithms are, as hinted above, robust *by construction* to dataset shift, or to some specific type of it.

Contributions. To assess the main drivers of behavioural change, we implement and experiment with 753 features divided into 9 groups and 69 subgroups, reflecting a broad array of socio-behavioural, linguistic, relational, and psychological aspects of online human behaviour. We leverage these features to "quantify" the changes in activity, toxicity, and participation diversity that 16,828 users exhibited following the Great Ban, a massive deplatforming intervention occurred on Reddit in 2020 [15]. Overall, our work provides the following main contributions:

- We cast the prediction of changes in post-moderation behaviour as a quantification problem. Then, we combine it with a greedy feature selection algorithm to identify the most informative user-level features.
- We carry out an extensive empirical study involving 16.8K users and 753 features for three prediction tasks: quantify behavioral changes in activity, toxicity, and participation diversity following a moderation intervention.
- For each task, we identify the most informative set of features. Our results reveal a small set of features that are consistently informative, and shows that the rest are either task-specific or of limited utility.
- We discuss practical implications for effective moderation, including the need to adjust interventions to the context of the moderation, the aim of the intervention, and the characteristics of the moderated users.

2. Related Work

2.1. Descriptive Content Moderation

Most research works on the effects of moderation interventions have adopted descriptive approaches (i.e., post-hoc observational designs) for estimating effects, focusing on how interventions influence user behaviour, community dynamics, and discourse visibility. Several works have examined community-level interventions such as bans, quarantines, and visibility restrictions. For example, analyses of "The Great Ban", a large-scale deplatforming intervention occurred on Reddit in 2020, revealed that subreddit removals produced heterogeneous effects across communities and user types, with top contributors often leaving and casual users changing the way they wrote [16, 86]. Further analyses showed that sequential interventions on `r/The_Donald` shaped user activity, toxicity, and information credibility, often inducing counter-intuitive effects such as increased polarisation [83]. Community-level interventions aimed at reducing the visibility of problematic user groups are the focus of [12, 76], which demonstrate that such interventions decrease new user influx but do not significantly

change the way existing users communicate. Finally, others analysed deplatforming interventions targeting influencers, and showed a significant drop in public attention towards them and a reduction of public presence of certain problematic ideas [40, 45].

A complementary line of work focused on user-level, rather than community- or platform-level, behavioural responses to moderation. Cima et al. [15] identified substantial inter-user heterogeneity following a deplatforming intervention, with a small but notable group of users reacting negatively to the ban by posting highly toxic comments. Similar results are described in [84], which found multiple outliers that diverged from community average estimates of intervention effects. Then, studies in [31, 93] examined temporary suspensions, showing that punishment duration and user status influence recidivism and behavioural contagion among peers. Overall, these findings emphasize the need for moderation strategies tailored to user profiles.

Lastly, some recent studies highlighted that user perceptions and bystander responses are increasingly recognized as critical in shaping moderation efficacy. Jhaver [44] found that people are more likely to support bans and warnings if they believe harmful content affects others more than themselves, while Zhao and Hobbs [95] showed that moderation decisions voted by groups of users may help in reducing toxic comments and improving community standards. Other studies showed that user perceptions of moderation effectiveness vary between moderated and bystander users [14, 38].

2.2. Predictive Content Moderation

Some works explored predictive approaches to content moderation, focusing not only on early identification of violations, but also on the consequences of moderation interventions, across different levels of analysis. One line of research focused on the preemptive detection of content violations. Kurdi et al. [50] analysed over 73,000 YouTube videos to identify features predictive of future removal, achieving promising results that support the feasibility of pre-emptive moderation. [65] proposed LAMBRETTA, a learning-to-rank system for the early identification of moderated tweets; by leveraging semantic keyword extraction from known misinformation tweets, LAMBRETTA identified up to 20 times more policy-violating content than manual methods, with minimal human input. Finally, CROSSMOD is a cross-community moderation tool for Reddit that infers macro-norm violations via transfer learning from 100 subreddits, enabling effective detection in communities lacking historical moderation data [11]. At the community level, [35] proposed a framework to identify subreddits likely to face moderation restrictions. Their models use structural and behavioural indicators to predict future bans up to nine months in advance, thus enabling scalable proactive moderation.

Beyond content, other works aimed at predicting users and community effects of moderation interventions, ahead of their application. For example, this predictive approach was applied to detect and forecast user ban evasion on Wikipedia using behavioural and psycholinguistic features [61]. Others tackled the task of forecasting user abandonment following a moderation intervention [82]. The present work is particularly related to [61] and [82], in that, similarly to the cited works, we carry out a ML task to predict the effects of a moderation intervention. However, contrary to previous work, we consider a broader range of

possible reactions to an intervention, including changes in activity, toxicity, and diversity of user participation, and we do that with the aim of identifying the most informative features for predicting such effects, rather than for optimizing prediction accuracy.

Finally, some recent studies explored the utility of large language models (LLMs) for predictive moderation. Among them, Kumar et al. [49] assessed the viability of using GPT-3.5 as a moderation agent. They applied it to 95 Reddit communities, finding that its performance varied significantly depending on community rules and content context. Others proposed the *policy-as-prompt* paradigm, enabling moderation decisions directly from policy instructions embedded in LLM prompts [64]. This approach reduces retraining needs but introduces challenges in prompt design and governance. Other works investigated small language models (SLMs) as lightweight alternatives to LLMs. The study in [94] demonstrated that fine-tuned SLMs may outperform LLMs in domain-specific moderation tasks; [89] introduced instead STAND-Guard, an SLM framework trained via cross-task learning that generalizes across novel moderation tasks.

2.3. Applications of Quantification

Quantification (a.k.a. "learning to quantify") is the supervised learning task of predicting the prevalence of the classes in a set of unlabelled datapoints characterized by dataset shift [25]. Quantification finds applications in several downstream tasks whenever these latter need to be carried out under dataset shift. Examples of such tasks are improving the accuracy of classifiers [70], measuring the fairness of classifiers [21] and rankers [43], predicting the accuracy of classifiers [87, 88], and calibrating classifiers [57].

Aside from these tasks, quantification is useful in all the disciplines that have an inherent focus on *aggregate* data (i.e., that are focused on phenomena at the "macro" level) and where the object of study are *populations* and not individuals, as in our case. In these fields, the prevalence estimates are interesting in themselves, and not just for solving a different task. Examples of such fields are epidemiology [46], the social sciences and political science [39], market research [73], and ecological modelling [34]. While no applications of quantification to online content monitoring have yet been discussed in the literature, the application we deal with in this paper squarely falls in this category.

Many quantification methods have been proposed to date ([20, 32] are two recent surveys on the subject). Most such methods have been tested under (and proved their value for addressing) prior probability shift, often considered the "paradigmatic case" of dataset shift [96]. Instead, other types of dataset shift, such as covariate shift and concept shift, have received considerably less attention in the quantification literature [33, 80].

3. Quantifying Feature Importance for Online Content Moderation

3.1. Problem Definition

Our overarching goal is to identify the set of features most relevant in determining user reactions to a moderation intervention, and to estimate their degree of relevance. To reach this goal, we develop a quantification model capable of estimating the class distribution that represents behavioural variations in an online community following an intervention.

In particular, we consider changes across multiple dimensions of user behaviour: activity, toxicity, and the diversity of user participation across subreddits. Quantifying changes in each of these dimensions constitutes a distinct task, requiring separate models trained to capture the specific dynamics associated with each behavioural aspect. In the remainder of the paper, we refer to each behavioural dimension as a distinct task. This multi-dimensional approach enables a more comprehensive assessment of feature relevance, as it allows us to identify whether certain feature groups are informative in general, or only in relation to specific types of behavioural change. The adoption of a quantification-based approach allows us to assess the informativeness of features at the population level, focusing on how well they capture aggregate shifts in behaviour, rather than at the level of individual predictions. This approach is well suited to the moderation context, where the objective is to understand broad user trends and effects, rather than to classify individual responses with high precision.

3.1.1. Class Definition

For each task (i.e., activity, toxicity, diversity), the classes correspond to discrete levels of variation in that dimension. In detail, we define the following five-point scale: HighlyDecreased, ModeratelyDecreased, NoVariation, ModeratelyIncreased, HighlyIncreased.[2] This 5-point scale generates an *ordinal* quantification context, one in which a total order is defined on our set of classes. As a consequence, mis-assigning some prevalence mass—say, from HighlyDecreased to HighlyIncreased—is a more serious mistake than mis-assigning the same prevalence mass from HighlyDecreased to ModeratelyDecreased.

The three behavioural dimensions that define the tasks are measured as follows. Given a fixed period of time, *Activity* is measured as the total number of comments posted by a user in that time period; *Toxicity* is defined as the 75th percentile of toxicity scores across the user's comments posted during that time period, and *diversity* measures the variety of communities (i.e., subreddits) in which a user has actively participated during that time period. The latter is computed via the well-known *Hill diversity index* [84]

$$^{q}H = \left(\sum_{i=1}^{R} p_i^q \right)^{\frac{1}{1-q}}$$

where p_i is the proportion of comments in subreddit i, and R is the total number of distinct subreddits the user participated in, with parameter $q = 1.5$.

3.1.2. Task Formulation

We frame the problem as a supervised learning task. Accordingly, we assume access to a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X} \subset \mathbb{R}^M$ represents a user in the social network. Here, x_i is an M-dimensional vector of covariates, while $y_i \in \mathcal{Y}$ represents the (true) label assigned to x_i, indicating one of the five variation levels for a particular behavioural aspect. For brevity, we encode these five levels as $\mathcal{Y} = \{1, 2, 3, 4, 5\}$. A vector x of covariates is a

[2]The way in which these classes are computed is thoroughly described in Section 4.2 "Ground-truth Labelling."

concatenation of different feature blocks $x = (f_1; \ldots; f_F)$, with $f_i \in \mathbb{R}^{d_i}$ a subvector of d_i real-valued components, such that $\sum_{i=1}^{F} d_i = M$. In our case we have $M = 753$ and $F = 69$.

The main objective is thus to select a subset of feature blocks that results in a more effective quantification system. Formally, if we consider $I = \{1, \ldots, F\}$ the set of indices of all available feature blocks, our goal can be stated as identifying

$$S^* = \arg\min_{S \subset I} \mathcal{L}(q(S), D) \tag{1}$$

where $\mathcal{L}$ is the loss function measuring the performance of a quantifier q applied to instances from D restricted to the selected features $x' = (f_{s_1}; \ldots; f_{s_{|S^*|}})$, with $s_i \in S^*$. Since the number of possible feature block combinations is $2^{69} \approx 6 \times 10^{20}$, the problem of examining all possible combinations of feature blocks is obviously intractable. To explore this space we thus resort to search heuristics, as detailed in Section 3.2.1.

3.1.3. Quantification Accuracy as a Proxy of Feature Importance

We treat the accuracy of our quantification model as a proxy of the quality of the set of feature blocks on which it was trained. This accuracy is measured in terms of a loss function, which must thus reflect the performance of the quantifier in a broad scenario.

To this end, we split the dataset D into a labelled set L (consisting of 8,000 randomly drawn instances), to be used to train the quantification model, and an unlabelled set U (the remaining instances), to be used for evaluation purposes. In order to reduce the experimental bias caused by one specific partition, we repeat this process 5 times.

To reflect scenarios characterised by different training set sizes, we divide the training set L into 16 batches $L_1, \ldots, L_{16}$ of 500 instances each. This setup allows us to monitor the behaviour of the quantifier in a series of experiments where we use, as the training data (here noted as T), progressively higher portions of L, from 500 ($T_1 = L_1$), 1,000 ($T_2 = L_1 \cup L_2$), ..., up to 8,000 instances ($T_{16} = L = \cup_{i=1}^{16} L_i$).

Since our ultimate goal is to obtain a quantifier that performs well in estimating variations in prevalence, we subject each experiment to a *stress test*, in which we evaluate the quantifier according to its ability to accurately predict widely varying class prevalence values. To this aim, we adopt the *Artificial Prevalence Protocol* (APP) [20, 25], in which the test set U is used to generate many test samples with different class distributions. Specifically, we generate 1,000 test samples $U_1, \ldots, U_{1000}$ as follows. We first generate 1,000 random prevalence vectors $\mathbf{p}_1, ..., \mathbf{p}_{1000}$, uniformly sampled from the probability simplex Δ^4, corresponding to the space of all possible distributions over 5 classes (which has 4 degrees of freedom) using the Kraemer sampling algorithm [78]. For each vector $\mathbf{p}_i$, we draw from U a test sample U_i of 500 instances that satisfies the specified prevalence values. The quantifier is then asked to estimate the class distribution of the 1,000 test samples, thus producing a series of estimates $\hat{\mathbf{p}}_1, ..., \hat{\mathbf{p}}_{1000}$. As our evaluation measure, we use the mean quantification error across all 1,000 test samples.

3.2. Evaluation Measures

Given that we tackle a set of ordinal quantification tasks, we measure the prediction error the models incur into when estimating the true prevalence $\mathbf{p}$ as $\hat{\mathbf{p}}$, by means of a standard

evaluation measure for ordinal quantification—namely, via the *Normalized Match Distance* (NMD), defined in [71] as

$$\mathrm{NMD}(\mathbf{p}, \hat{\mathbf{p}}) = \frac{1}{n-1}\,\mathrm{MD}(\mathbf{p}, \hat{\mathbf{p}}) \tag{2}$$

where $\frac{1}{n-1}$ is a normalization factor that allows NMD to range between 0 (best prediction) and 1 (worst prediction). Here, n is the number of available classes (in our case, $n = 5$), and MD is the well-known *Match Distance* [91], defined as

$$\mathrm{MD}(\mathbf{p}, \hat{\mathbf{p}}) = \sum_{i=1}^{n-1} d(y_i, y_{i+1}) \cdot |P(y_i) - \hat{P}(y_i)| \tag{3}$$

where $P(y_i) = \sum_{j=1}^{i} p(y_j)$ is the prevalence of y_i in the cumulative distribution of $\mathbf{p}$, $\hat{P}(y_i) = \sum_{j=1}^{i} \hat{p}(y_j)$ is an estimate of it, and $d(y_i, y_{i+1})$ is the "semantic distance" between consecutive classes y_i and y_{i+1}, i.e., the cost we incur in mistaking y_i for y_{i+1} or vice versa. Throughout this paper, we assume $d(y_i, y_{i+1}) = 1$ for all $i \in \{1, 2, 3, 4\}$. MD is *the* standard measure of evaluation in ordinal quantification [8, 10]. Since, as mentioned above, for each choice of a group S of features and a task T we run 16 experiments (one for each training set size in {500, 1000, ..., 7500, 8000}), we compute the error incurred by the system when using feature group S for task T as the mean of the NMD scores across these 16 setups, which we here denote by $\mathrm{MNMD}_S^T(\mathbf{p}, \hat{\mathbf{p}})$. In order to evaluate the overall contribution of feature group S for task T, we perform ablation experiments on task T in which we remove group S from the set of all features, and measure this contribution by the *relative increase in error* (RIE) that derives from removing feature group S from the set of all features

$$\mathrm{RIE}_{S'}^T = \frac{\mathrm{MNMD}_{S/S'}^T - \mathrm{MNMD}_S^T}{\mathrm{MNMD}_S^T}$$

3.2.1. Greedy Exploration for Feature Block Selection

The greedy exploration for selecting features consists of carrying out a series of tests in which the contribution of each feature block is evaluated. The exploration begins from an initial *candidate* configuration (i.e., an initial selection of feature groups) $C \subseteq I$. The initial configuration is selected by picking the best-performing configuration across different setups $G_1, \ldots, G_9$, each consisting of a specific superset of the 9 groups of feature groups that we will describe in Section 3.3, plus one additional configuration $G_0 = I$ that contains all indexes. Note that the groups $G_1, \ldots, G_9$ form a partition over the feature groups, that is $G_i \cap G_j = \emptyset$ if $i \neq j$ and $\bigcup_{i=1}^{9} G_i = G_0 = I$.

The exploration then goes on by evaluating different feature group in turns. Each evaluation of a feature block f_i consists of testing whether removing the index i from C (if $i \in C$) or whether adding the index i to C (if $i \notin C$) reduces the loss. In such case, the modification is retained, and otherwise is reverted. Since the number of feature groups is relatively high, and it is likely that unimportant groups are discarded early in the process, we restrict feature addition to the first three rounds of exploration. The subsequent rounds

Algorithm 1 Greedy selection of feature groups.

$C \leftarrow \arg\min_{G_i \in \{G_0,\ldots,G_9\}} \mathcal{L}(q(G_i), D)$ ▷ Initial configuration
best-loss $= \mathcal{L}(C, D)$
sorted $\leftarrow$ sort$(\{f_1, \ldots, f_{69}\}, \mathcal{L}, D)$ ▷ Sort candidate feature groups
improvement $\leftarrow$ True
round $\leftarrow 0$
while improvement **do**
  improvement $\leftarrow$ False
  for $f'_i \in$ sorted **do**
    if $f'_i \in C$ **then**
      $C' \leftarrow C - f'_i$ ▷ Remove the feature block
    else
      if round < 3 **then**
        $C' \leftarrow C \cup f'_i$ ▷ Add the feature block (only during the first 3 rounds)
      else
        continue ▷ The refinement rounds (round ≥ 3) only consider ablation tests
      end if
    end if
    new-loss $\leftarrow \mathcal{L}(C', D)$
    if new-loss $<$ best-loss **then**
      $C \leftarrow C'$ ▷ Keep the last change
      best-loss $\leftarrow$ new-loss
      improvement $\leftarrow$ True
    end if
  end for
  round $\leftarrow$ round $+ 1$
end while
return C

focus on *refining* the selected set of features (presumably reduced) and only test for feature removal. Refinement continues until no further improvement is observed.

The order in which the feature groups are presented corresponds to the increasing performance (i.e., descending loss) of a system trained using only that feature block individually. In other words, the order of f_i corresponds to the inverse rank of $\mathcal{L}(q(\{i\}), D)$ across all $i \in I$. This means that the least promising candidates, those yielding the highest loss, are processed first, while the most promising ones come last. The entire process is repeated through three rounds, and the final configuration C is returned as an estimate $\hat{S}$ of S^*. The pseudocode describing this greedy exploration is presented in Algorithm 1.

3.3. Feature Extraction

We extracted a total of 753 features, divided into the following categories: `EMBEDDINGS` (384, 51.0%), `LIWC` (112, 14.9%), `WRITING_STYLE` (81, 10.8%), `RELATIONAL` (36, 4.8%), `ACTIVITY` (34, 4.5%), `SENTIMENT` (32, 4.2%), `SOC_PSY` (31, 4.1%), `TOXICITY` (24, 3.2%), and `EMOTIONS` (19, 2.5%). Within each group, features sharing similar characteristics were further divided into subgroups. This hierarchical organisation allows for a fine-grained analysis of which specific subdimensions are most predictive. Each category of features is briefly

introduced in the following. Additionally, a detailed summary of all groups of features and their corresponding subgroups is provided in Table A.4 in the Appendix.

3.3.1. Embeddings

This group of features is the most numerous, as it contains more than 50% of the overall total. For each user, we computed a 384-dimensional embedding vector for each comment using the sentence transformer model ALL-MINILM-L6-V2[3], and we then averaged these vectors to obtain a single representation per user. Embeddings are widely used due to their ability to capture semantic and linguistic information. In particular, they have proven particularly effective in tasks related to detecting harmful content on social media, such as hate speech detection [72], deepfake identification [22], and cyberbullying [48]. Since these tasks aim to understand user behaviour on social media, embeddings could also be effective in capturing linguistic patterns that are indicative of how users react to an intervention, hence our experimentation.

3.3.2. Linguistic Inquiry and Word Count

Linguistic Inquiry and Word Count (LIWC) is a lexicon that maps textual content to psycholinguistic categories, including cognitive processes, social concerns, and affective expression [7]. LIWC has been widely used in the literature due to its ability to capture psychologically meaningful aspects of language. For instance, it has been effectively used in tasks such as hate speech detection [69], misinformation detection [26], and personal value prediction [77]. For this reason, LIWC can offer valuable insights into the psychological dimensions that may influence how users react to moderation interventions. We use the latest version of the lexicon, i.e., LIWC-22.[4]

3.3.3. Writing Style

Previous studies on the effects of content moderation suggest that users may modify their linguistic style in response to moderation interventions. For instance, users who continued to engage on the platform were found to shift from first- to third-person plural pronouns [41], while other research reported a decrease in ingroup-oriented language [86]. Inspired by these insights, this group of features includes part-of-speech distributions and readability indices, which may help detect subtle linguistic changes following moderation.

3.3.4. Activity

A fundamental aspect of user behaviour is their activity on social media, as platforms' revenues strongly depend on user engagement [9, 36]. However, certain interventions caused users to reduce their activity or even abandon the platform [15, 16, 84, 86], which justifies the adoption of this group of features. In fact, previous work already showed that activity trends, number of comments, as well as temporal indicators, are among the most predictive features of user churn following moderation interventions [82].

[3] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

[4] `https://www.liwc.app/`

3.3.5. Relational

At the core of any social media platform lies the interaction between users and communities, as these are the drivers of user activity and engagement. However, moderation interventions may change these interactions [35, 83]. Following this intuition, we include a group of relational features aimed at capturing how users interact with communities. For instance, we measure a user's influence in both banned and non-banned communities as well as initiative and adaptability indicators [53]. The rationale behind these features stems from the idea that a user's level of involvement in non-banned communities, as well as their interactions with other users affected by the ban, may influence their response to a moderation intervention.

3.3.6. Sentiment

Sentiment analysis is a widely adopted task in social media research, as it allows to assess the positive, negative, or neutral tone and sentiment of user-generated content. In particular, it is widely used in hate speech detection, as negative polarity often correlates with toxicity [3]. Here, we computed user sentiment with VADER [42], a rule-based sentiment analyser specifically developed for social media. It assigns to each comment a sentiment score that can either be negative, neutral, positive, and a compound score that summarizes overall sentiment. To obtain user-level features, we aggregated these scores across all comments posted by each user. Similarly, we computed the sentiment of the emojis used by each user, leveraging the dictionary developed in [47], which returns the same scores returned by VADER for each emoji.

3.3.7. Socio-Psychological

Predicting the effects of moderation interventions is related to predicting user behaviour, which is known to be influenced by psychological and demographic factors [19, 92]. Additionally, several studies have shown that such characteristics are associated with online behaviours, including toxic and aggressive tendencies [18, 29]. On top of that, recent studies have demonstrated that incorporating socio-psychological features leads to significant improvements in hate speech detection tasks [69]. Among the features included in this group are users' demographic attributes, such as age and gender, inferred by training a classifier on the PANDORA dataset [30]. We also computed personality traits based on the OCEAN model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), obtained by averaging the scores returned for each comment using a pre-trained classifier.[5] Finally, for each user we assessed their moral values as resulting from their comments, by using the MORAL FOUNDATIONS framework, a well-established theory that identifies five core dimensions of moral reasoning: care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and purity/degradation [51]. For each comment, we extracted for each dimension two scores, i.e., bias, which indicates the direction (e.g., toward care or harm), and intensity, expressing how strongly a dimension is conveyed. We then averaged each score for each user.

[5] `https://github.com/jkwieser/personality-prediction-from-text`

3.3.8. Toxicity

One of the most common reasons behind the ban of online communities is the spread of toxic content [35, 40, 86]. Hence, measuring user toxicity levels prior to the ban may serve as a strong predictor of their post-ban behaviour. Here, we computed toxicity scores with DETOXIFY [37], a widely adopted multilingual deep learning classifier, largely used in tasks related to the study or detection of toxicity on social media [16, 27, 82]. For each comment, DETOXIFY returns a range of scores such as toxicity, severe toxicity, obscenity, insult, identity attack, and threat that we aggregated at the user level.

3.3.9. Emotions

This group includes features aimed at capturing the emotions expressed by users in their comments, which can be indicative of their psychological state and may shape their response to moderation. Emotion analysis have been successfully used in tasks such as hate speech detection or assessing the wellbeing of online users [69]. Here, we leveraged two lexicon-based resources: the NRC EMOTION INTENSITY LEXICON (NRC-EIL) [55], which provides word-level intensity scores for anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, and the NRC VALENCE-AROUSAL-DOMINANCE (NRC-VAD) [54] lexicon, which quantifies words along continuous affective dimensions of valence (positivity), arousal (intensity), and dominance (control). To complement these lexicon-based features, we incorporated EMOATLAS [74], a computational framework that assigns a z-score to each comment for discrete emotions including joy, trust, fear, surprise, sadness, disgust, anger, and anticipation.

4. Experimental Settings

Here we describe the experiments we have carried out. The code that reproduces our experiments is available on GitHub.[6]

4.1. Dataset

For our analyses we leverage the dataset described in [15, 16]. It consists of approximately 16M Reddit comments posted by 16,828 users affected by the Great Ban on June 29, 2020. The dataset spans a 14 months period of time, including seven months before and seven months after the ban, which enables thorough studies of the ban's effects. The users in the dataset represent those who were consistently active prior to the ban in at least one of 15 large banned subreddits. Table 1 lists the 15 selected subreddits along with the number of subscribers and active users therein. The dataset does not include bots, who were filtered out via a semi-automatic procedure based on filtering rules and manual validation [16]. The 16M user comments include 2.2M comments that the 16.8K users posted in the seven months prior to the ban *within* the banned subreddits, and 13.8M comments that the same users posted in the 14 months period *outside* of the banned subreddits. Since the selected subreddits were permanently shut down with the ban, no post-ban activity exists within

[6] `https://github.com/AlexMoreo/EffectPrediction`

subreddit	subscribers	users	comments
r/ChapoTrapHouse	159,185	9,295	1,368,874
r/The_Donald	792,050	4,262	619,434
r/ConsumeProduct	64,937	1,730	60,073
r/DarkHumorAndMemes	421,506	1,632	35,561
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 1: List of banned subreddits used in this study. For each subreddit, we indicate the number of subscribers, the number of unique users, and the total number of comments.

them. Therefore, user activity outside of the banned subreddits is necessary to assess the effects of the intervention.

Based on this dataset, we computed user features using comments posted prior to the ban both inside and outside of the banned subreddits. Instead, the labelling of users (i.e., as those who increased/decreased activity/toxicity/diversity) was based solely on their comments outside the banned subreddits, for which we have both pre- and post-ban data. Additionally, to avoid possible spurious fluctuations, when computing features we excluded all comments made on the day of the ban. As a result, 753 features were computed for 16,522 users. Finally, since both toxicity and diversity are computed from user comments, their reliability depends on having a sufficient level of post-ban activity [15]. Therefore, to ensure the robustness of our measurements, we restricted the labelling of toxicity and diversity changes to users who posted at least ten comments in the post-ban period.[7]

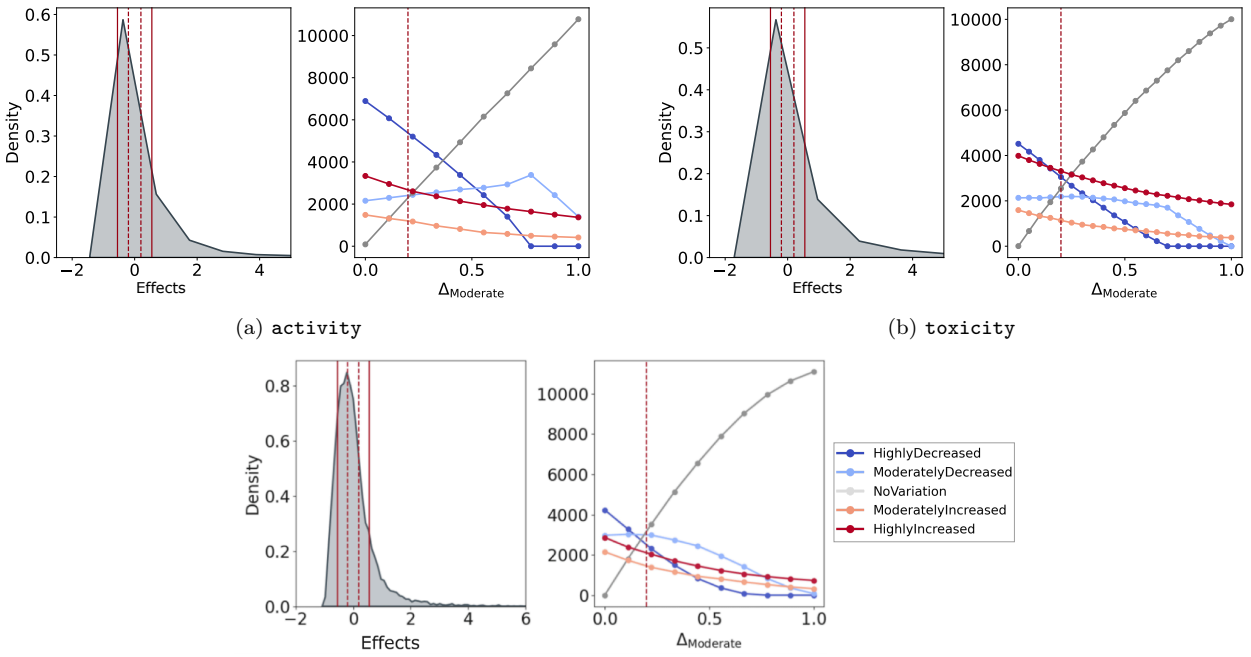
4.2. Ground Truth Labelling

To evaluate the effects of the intervention, for each user we compute the relative change they exhibit along three tasks, according to the measure

$$\text{Effect}_T = \frac{T_{\text{post}} - T_{\text{pre}}}{T_{\text{pre}}} \quad \text{with } T \in \{\text{activity}, \text{toxicity}, \text{diversity}\}$$

The resulting values are continuous numbers, which we need to discretize into five classes in order to carry out the supervised ordinal quantification task. To this end, we define two

[7]Activity in the pre-ban period is guaranteed by design, since the original dataset only contains users who were consistently active before the ban [16].



(a) `activity` (b) `toxicity`

(c) `diversity`

Figure 1: Effects distribution and class distribution as a function of the parameter Δ for each of the considered tasks: activity, diversity, and toxicity. For the effect scores, we report the kernel density estimation (KDE) of the continuous values. The red dashed lines indicate the thresholds $\pm\Delta_{\text{Moderate}}$, while the red solid lines indicate $\pm\Delta_{\text{High}}$.

thresholds Δ_{Moderate}, $\Delta_{\text{High}} > 0$, with $\Delta_{\text{Moderate}} < \Delta_{\text{High}}$. Then, for each task T, we label users as

$$\text{Label}_T = \begin{cases} \textsf{HighlyIncreased} & \text{if} \quad \Delta_{\text{High}} < \text{Effect}_T \\ \textsf{ModeratelyIncreased} & \text{if} \quad \Delta_{\text{Moderate}} < \text{Effect}_T \leq \Delta_{\text{High}} \\ \textsf{NoVariation} & \text{if} \quad -\Delta_{\text{Moderate}} \leq \text{Effect}_T \leq \Delta_{\text{Moderate}} \\ \textsf{ModeratelyDecreased} & \text{if} \quad -\Delta_{\text{High}} \leq \text{Effect}_T < -\Delta_{\text{Moderate}} \\ \textsf{HighlyDecreased} & \text{if} \quad \text{Effect}_T < -\Delta_{\text{High}} \end{cases}$$

The distributions of user-level effects Effect_T across the three tasks approximately follow Gaussian curves centred around zero, indicating that most users experienced only modest changes, while fewer exhibited pronounced increases or decreases. When defining the Δ_{Moderate} and Δ_{High} thresholds for class labels, we considered two key criteria. First, thresholds were chosen to reflect practically meaningful changes in user behaviour, in line with the real-world implications of user churn, rising toxicity, or shifting participation patterns. Second, we aimed to produce a relatively balanced class distribution, ensuring that all five classes were sufficiently represented. Figure 1 shows the class distributions, in terms of number of users per class, for each task, as a function of Δ_{Moderate}, with $\Delta_{\text{High}} = \Delta_{\text{Moderate}} + 0.35$. Based on this preliminary analysis, we set $\Delta_{\text{Moderate}} = 0.2$ and $\Delta_{\text{High}} = 0.55$ for all tasks.

5. Results

5.1. Quantifying with the Full Feature Set

We begin by evaluating the performance of several well-known quantification methods using the full set of features available. This allows us to compare how effectively each method estimates class distributions across the three tasks—prediction of changes in activity, toxicity, and participation diversity—when provided with maximum information. The best-performing method from this comparison is then adopted for all subsequent experiments on feature selection. We follow this strategy because repeating the entire feature evaluation pipeline for multiple quantifiers would come at a considerable computational cost. The results presented here thus serve both to identify an overall strong-performing quantifier and to establish a point of reference against which improvements from feature optimization can be measured.

We consider the following three well-known quantification methods:

- `CC`: Classify and Count is a weak quantifier that simply *(i)* trains a classifier on the training data, *(ii)* uses the classifier to issue label predictions for all the instances in the test sample, and *(iii)* computes the percentages of test items attributed to each class.

- `PACC`: Probabilistic Classify and Count [6] is a quantifier that corrects an initial raw count estimate by taking into account the class-conditional error rates of the classifier (estimated via cross-validation). In contrast to `CC`, the underlying classifier of `PACC` is a probabilistic one that returns estimates of posterior probabilities for each data point.

- `EMQ`: Expectation Maximisation for Quantification [70] is a mutually recursive iterative algorithm that alternates between two steps: *(i)* the *expectation* step, in which the class prevalence values are estimated as the average of the posterior probabilities, and *(ii)* the *maximisation* step, in which the posterior probabilities are updated to reflect the most recent class prior estimates. This process is repeated until convergence.

In addition to these, we also considered more recent methods such as `KDEy` [59], but ultimately excluded them due to their high computational complexity. Since our overarching goal is not to achieve the best possible predictive performance but to assess feature importance, we prioritize methods that offer a good trade-off between effectiveness and scalability.

All three above quantifiers rely on an underlying classification model to generate label predictions or probabilities, which are then used to estimate class prevalence. Following a common trend in the quantification literature, we rely on Logistic Regression (`LR`) for this purpose. We optimize the hyperparameters of `LR` independently for each quantifier and training set, via grid-search exploration, following the quantification-oriented model selection protocol proposed in [60]. We rely on `scikit-learn` for the implementation of `LR` and on `QuaPy` [58] for the implementations of `CC`, `PACC`, and `EMQ`.

In addition to the above quantifiers trained with the full feature set, we also report the results of the Maximum Likelihood Prevalence Estimator (`MLPE`), a naïve method that

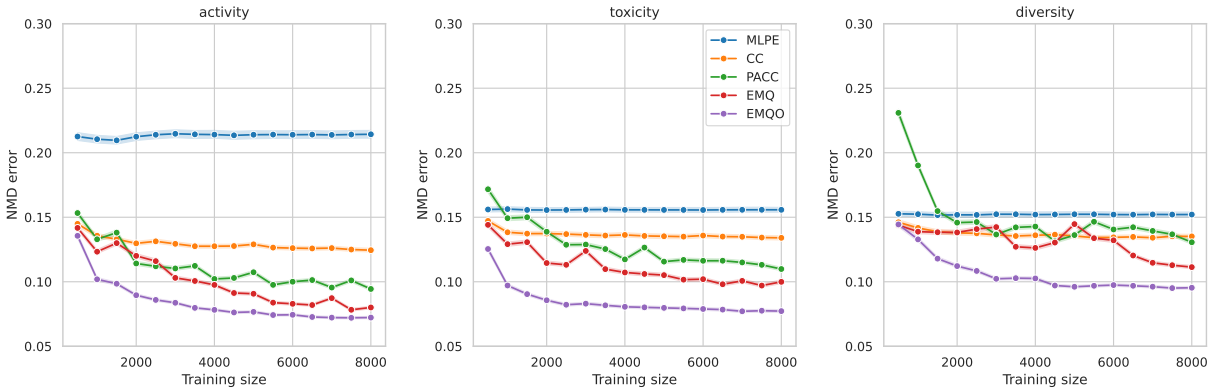


Figure 2: Quantification error of the models as a function of training set size. The curves report the error trends across incremental subsets for the `activity`, `toxicity`, and `diversity` tasks.

assumes training and test data are IID. `MLPE` simply takes the class prevalence observed in the training data as an estimate for any incoming test set, regardless of its actual distribution. Finally, we also report the results of `EMQO`, an optimized `EMQ` model equipped with the best feature groups identified for each dataset with the feature selection process. Although the feature selection process is discussed in a later section, here we include the results of `EMQO` as a point of reference, serving as a lower bound on quantification error and providing context for interpreting the performance of models that use the full feature set.

Figure 2 displays the quantification error of each model as the training set size increases, allowing us to assess how performance scales with data availability. Each curve shows the trend in error across incremental training subsets, from 500 to 8,000 instances, for all three behavioural tasks. The figure reveals that `MLPE` and `CC` yield poor estimates of class prevalence. Instead, `PACC` and, especially, `EMQ`, attain much lower errors. In addition, while the simpler models exhibit a relatively stable performance independently of the training set size, the most powerful ones consistently reduce the quantification error when provided with more training data. These results are robust to the task at hand. Based on these results, in the following sections we adopt `EMQ` as our quantifier of choice for estimating feature importance. In addition to the previous findings, Figure 2 also shows a marked and consistent improvement of `EMQO` with respect to `EMQ` in all three tasks, which demonstrates the goodness of the feature selection process and the usefulness of an optimized set of features.

5.2. Selecting the Most Informative Features

Initializing the Feature Selection Algorithm. Before running the greedy feature selection algorithm, we must choose an initial configuration of feature groups from which the search begins. This is done by evaluating the performance of each feature group individually, along with the full feature set, and selecting the configuration that yields the lowest quantification error. The initial evaluation is carried out by running the evaluation procedure described in Section 3.1.3. This provides a strong starting point for the iterative search algorithm that follows.

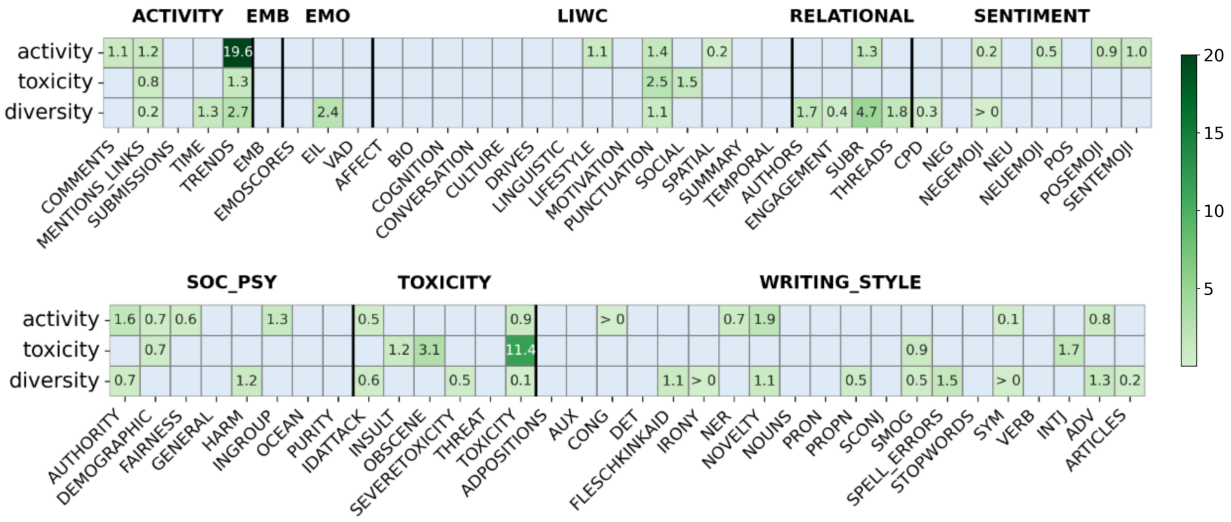


Figure 3: Percentage RIE_S^T for each subgroup across each task. Darker green-coloured cells indicate higher importance of a subgroup for a given task, whereas blue-coloured cells denote that the subgroup was not used in that task.

Table 2 reports the quantification error made by each configuration in each task. In table, the best result (i.e., lowest error) for each task and for the mean performance across all tasks is shown in bold, and the second-best is underlined. As shown, the best configuration for the activity and toxicity tasks correspond to the namesake feature groups. Instead, for the task of predicting changes in participation diversity, the best configuration is the one consisting of `ALL` feature groups. The results in Table 2 highlight some interesting patterns. As expected, the feature groups most directly related to the target variables, `ACTIVITY` and `TOXICITY`, are the best predictors for their respective tasks, confirming the relevance of those features. In contrast, no single group is clearly aligned with the diversity task, and as a result, using `ALL` feature groups together yields the best performance in that case. Interestingly, the full feature set also ranks second-best for both activity and toxicity, suggesting that while other groups contain useful information, combining everything indiscriminately is suboptimal, which underscores the importance of feature selection. Moreover, the overall results in table show that certain feature groups, like `EMBEDDINGS`, perform consistently well across tasks, while others are highly specialized. Examples of the former are the aforementioned `ACTIVITY` and `TOXICITY` features, which exhibit large variability in performance. Finally, the consistently higher quantification errors observed for the diversity task indicate that it is inherently more difficult given the available features.

Ordering the Subgroups of Features. Once the initial set of features for each task has been selected as per the results of Table 2, the 69 subgroups of features that make up our full feature set are processed sequentially in order to assess if and how much they contribute to the overall prediction, separately for each task. The order in which the subgroups of features are processed corresponds to their performance when used in isolation. Such performance and the resulting order is reported, for each task, in Table B.6 in the Appendix.

	task			overall	
feature group	*activity*	*toxicity*	*diversity*	*mean*	*stdev*
ACTIVITY	**711.4**	1198.5	1171.7	1027.2	273.8
EMBEDDINGS	1025.4	948.1	986.3	986.6	38.7
EMOTIONS	1376.9	1230.6	1211.8	1273.1	90.4
LIWC	1077.3	987.9	1058.3	1041.2	47.1
RELATIONAL	795.3	1188.0	995.3	992.9	196.4
SENTIMENT	1182.3	914.8	1292.6	1129.9	194.3
SOC_PSY	1166.2	1060.8	1174.1	1133.7	63.3
TOXICITY	947.8	**733.1**	1311.7	997.5	292.5
WRITING_STYLE	977.6	1004.9	1019.3	1000.6	21.2
ALL	748.8	830.5	**984.1**	**854.5**	119.5

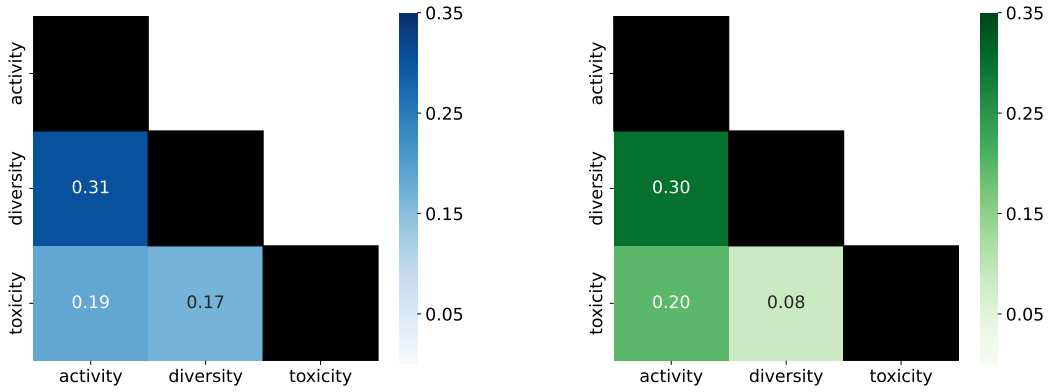
Table 2: Group-wise initial exploration. Bold indicates the best result for each dataset, in terms of MNMD, and also indicates the initial feature block used as a reference value for the heuristic feature selection procedure.

	task		
	activity	*toxicity*	*diversity*
num. selected subgroups	22 (32%)	10 (14%)	25 (36%)
MNMD EMQ (all features)	748.803	830.468	984.148
MNMD EMQO (selected features)	619.986	626.650	787.219
relative error reduction	17.20%	24.54%	20.01%

Table 3: Final selection of feature subgroups for each task. It reports the number of selected subgroups, the MNMD of the non-optimized EMQ models using all features, the MNMD of the EMQO models based on the selected subgroups, and the relative error reduction achieved by EMQO compared to EMQ.

Selecting the Subgroups of Features. The greedy feature selection strategy described in Section 3.2.1 is then applied to the ordering of Table B.6. Figure 3 shows the final results of the feature selection process where, for each task, the selected subgroups of features are represented by the green-coloured cells. In detail, the figure reports the percentage feature importance of each subgroup, where darker cells indicate higher importance, and blue cells denote that a subgroup was not chosen for the corresponding task. Additionally, Table 3 provides summarized results per task, reporting the number of selected feature subgroups, the reference MNMD value of the non-optimized EMQ models that rely on all features, and the final MNMD corresponding to the EMQO models that make use of the selected feature subgroups. Finally, the relative error reduction of the EMQO models with respect to the non-optimized EMQ ones is reported in the bottommost row.

As already noticed in Figure 2, aggregated results in Table 3 reveal the EMQO models leveraging an optimized feature set markedly reduce the quantification error, with a reduc-



(a) Jaccard index.

(b) Rank-Biased Overlap (RBO).

Figure 4: Pairwise similarity between the sets of features selected for the three behavioural prediction tasks: activity, toxicity, and diversity. Results are shown as heatmaps of the Jaccard index (left) and Rank-Biased Overlap (right). Larger values indicate greater similarity. Both measures reveal overall low similarity, highlighting that each task relies on largely distinct sets of predictive features.

tion of 17.20% in the activity task, 23.76% in the toxicity task, and 18.15% in the diversity one. Then, the number of selected feature subgroups varies considerably across tasks: only 10 (14%) for toxicity compared to 22 (32% of all available subgroups) for activity and 25 (36%) for diversity. The relatively parsimonious feature selection for toxicity suggests that a limited subset of features is sufficient to capture the behavioural changes in this dimension. Concerning Figure 3, we notice that feature importance values are generally low and often close to zero, indicating a multitude of weakly informative features. Notable exceptions are the `TRENDS` subgroup that shows a large importance for the activity task and the `TOXICITY` subgroup that achieves a similar result for the toxicity task. These results align with the definitions of the tasks and are consistent with recent findings [82]. In fact, the `TRENDS` subgroup alone achieves 19.8% feature importance in the activity task, while the sum of the remaining features only accounts for 18.21%. This dominance is reflected in a Gini coefficient $g = 0.65$, indicating that this single subgroup largely outweighs the others. A similar pattern is observed for the toxicity task, where the `TOXICITY` subgroup achieves 11.4% feature importance compared to 13.7% of the remaining features summed together, with a Gini coefficient $g = 0.53$. This shows that the `TOXICITY` subgroup is dominant in the namesake task, although less markedly so than the corresponding one in the activity task. In spite of the similar patterns however, the combined analysis of Figure 3 and Table 3 highlights some differences between the activity and toxicity tasks. In fact, despite using as much as 22 feature groups, the optimized activity model only improves on the non-optimized one by 17.20%. This result suggests that beyond activity trends, the remaining features collectively provide little information for the activity task, contrary to the toxicity one whose improvement is 23.76% with just 10 selected feature subgroups. Figure 4 reports an analysis of the degree of similarity between the sets of features selected for the three tasks. We assess similarity using two complementary indicators. The Jaccard index J captures the extent

of overlap between feature sets, providing an intuitive measure of how many features are shared across tasks. Additionally, the Rank-Biased Overlap (RBO) considers the relative ordering of features, with higher weight given to those at the top of each ranking [90]. Both indicators provide values $\in [0, 1]$, where 0 denotes no overlap and 1 perfect overlap. Results in Figure 4 show weak similarity across tasks, with mean $J = 0.22$ and mean RBO $= 0.19$. Across both indicators, activity and diversity appear as the most similar tasks, while activity and toxicity as the most dissimilar. Overall, these results highlight the importance of designing specific rather than generic features for each task, and suggest that predicting possible changes in toxicity represents a simpler task than predicting changes in activity. Besides these results, diversity remains the most challenging task. This is reflected in Table 3 by the largest set of selected features out of all tasks, and the largest quantification error for both the optimized and non-optimized models. Moreover, Figure 3 shows that no subgroup of features stands out as significantly more important than the others, resulting in the lowest Gini coefficient $g = 0.52$ out of the three tasks. These results highlight the intrinsic difficulty of the task as it is not straightforward to identify features that are strongly predictive of changes to participation diversity. Taken together, these findings suggest that predicting changes in participation diversity, and to a lower extent in activity, remains challenging with the available features, and is less amenable to performance gains through feature selection alone.

6. Discussion and Conclusions

We explored the ways in which a wide range of socio-behavioural, stylistic, and psychological features relate to user responses to a moderation intervention on Reddit. Our findings uncover several noteworthy patterns in how different types of user attributes correspond to distinct post-intervention behavioural changes.

6.1. Consistently Informative and Uninformative Features

Across all three behavioural prediction tasks—activity, toxicity, and participation diversity—a small set of staple feature groups consistently emerged as informative, i.e., `MENTIONS_LINKS`, `TRENDS`, `PUNCTUATION`, and `TOXICITY`. Their persistent selection, regardless of the outcome being modelled, highlights their broad relevance in capturing user behaviour following a moderation intervention. For example, the inclusion of `TOXICITY` features in all tasks reflects the central role that toxic expression plays in shaping user responses on Reddit, especially in the context of the Great Ban, where affected communities were removed due to aggressive and harmful conduct.[8] This suggests that toxic behaviour is not only predictive of future toxicity, but also intertwined with users' overall engagement and community participation following a moderation intervention. Similarly, the consistent use of mention-related features points to the importance of social embeddedness—how users interact with others, reference communities, or cite external sources—as a general indicator of behavioural patterns. Activity and toxicity `TRENDS` also surfaced as widely useful, aligning with prior findings that

[8] `https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/`

noted the predictive power of such temporal dynamics in anticipating user reactions to moderation [82]. Finally, the selection of PUNCTUATION features across all tasks supports recent work showing that punctuation is not merely structural, but often carries stylistic or pragmatic meaning—helping to signal tone, stance, or identity in online discourse [2]. Together, these features capture a blend of content, context, and expression that appears robustly predictive across multiple dimensions of user response to moderation.

While some feature groups proved consistently informative, a substantial portion (33% of all feature groups) were never selected for any task. This outcome partly reflects the breadth of our approach, which deliberately incorporated a wide array of features drawn from recent literature on online behaviour and moderation [3, 27, 69, 83]. Naturally, not all of these features are equally useful. Still, this result is informative in that it suggests that many features commonly used in related studies may offer limited value when it comes to predicting user reactions to moderation interventions. As an example, many of the features used herein were recently shown to boost online hate speech detection [69], a strictly related task. However, rather than dismissing these features outright, our results suggest that future work could benefit from prioritizing feature implementation by focusing on the groups and subgroups that show the most promise based on the task at hand. In any case, the overall sparsity of Figure 3, where truly informative signals are relatively rare and for any given task most available features contribute little to predictive performance, underscores the difficulty of the tasks. These findings highlight the complexity of predicting user responses to moderation [82], and underscore the need for continued scientific exploration into this practically important and socially impactful task.

6.2. Task-Specific Feature Relevance

When introducing our results in Section 5, we noticed some relevant differences in predictive performance across the three behavioural tasks, with changes in activity and toxicity being estimated more accurately than changes in participation diversity. This gap in performance suggests that the diversity task might be inherently more challenging. Part of the difficulty could stem from the lack of a dedicated set of predictive features for the diversity task. In contrast, both the activity and toxicity tasks benefited from feature groups that were explicitly aligned with their respective behavioural dimensions, such as past activity levels and expressions of toxic language. This is relevant since our feature set reflects the current state of the art in modelling online behaviour [61, 69, 77]. It thus emerges that participation diversity represents a dimension of online user behaviour that has received so far limited attention compared to other more studied phenomena, despite diversity and plurality being widely recognized as important in a growing body of literature on online moderation and algorithmic fairness [23, 66, 84]. The lower performance on the diversity task highlights a research gap, and calls for further investigation into the behavioural signals that drive shifts in user participation.

Beyond differences in task difficulty, our analysis also revealed that the sets of features selected for each task are largely distinct. This suggests that predicting changes in activity, toxicity, and participation diversity, requires attending to different underlying drivers of user behaviour. For example, RELATIONAL features—capturing interactions with other users and

communities—were almost entirely absent in the selected sets for the activity and toxicity tasks. Yet all relational subgroups were included for the diversity task. Indeed, participation diversity is inherently related to the structure and breadth of a user's engagement across subreddits. `WRITING_STYLE` features also emerged as more relevant to diversity: 60% of writing style -related subgroups were selected for this task, compared to only 25% and 20% for activity and toxicity, respectively. This observation resonates with recent findings that different communities often exhibit distinctive linguistic norms [76, 84], suggesting that variation in writing style may reflect the diversity of contexts in which a user participates. Such features, therefore, offer indirect signals about a user's exposure to and integration within varied online spaces. Overall, these findings emphasize the importance of tailoring feature design to the specific behavioural dimension under study, and point toward more nuanced, context-aware modelling approaches for capturing the multifaceted ways users respond to moderation [14, 18]. Other than this, our results also have practical implications regarding the generalizability of predictive models in this domain. Our findings suggest that distinct aspects of user behaviour, such as activity levels, toxicity, or participation diversity, are influenced by different sets of features, and therefore require dedicated predictors. This might limit the feasibility of using a single, unified model to estimate the overall impact of a moderation intervention. Instead, accurately capturing the full spectrum of possible user responses might require training and maintaining separate predictive systems for each behavioural outcome of interest, which may complicate large-scale or real-time applications.

6.3. Toward Objective-Specific Moderation Strategies

The results presented in Figure 3 highlighted the nuanced role that different user features play in predicting behavioural changes following moderation interventions. While some feature subgroups were consistently selected across tasks and others were never selected, these two extremes account for only about 40% of all considered subgroups. The majority of features fall into an intermediate space, i.e., they are predictive for some outcomes but not others. This variability suggests that user reactions to moderation are multifaceted and manifest distinct behavioural signals depending on the dimension being targeted. This finding carries implications for how moderation strategies are designed and deployed. For instance, interventions aimed at encouraging broader participation across communities—such as efforts to reduce echo chambers or increase exposure to diverse viewpoints—should be grounded in the behavioural signals most relevant to participation diversity. Conversely, strategies targeting reductions in toxic behaviour should be informed by a different set of user characteristics. In other words, our results point to the need for *objective-specific moderation*, where each intervention is tailored not only to the user it targets, but also to the specific behavioural outcome it seeks to influence.

This perspective adds a new layer to the growing call for more user-centred and personalized moderation approaches. While prior work has emphasized that moderation should consider the traits and histories of the users involved [19, 18], our findings complement and extend this view by arguing that the design of interventions should also be guided by the nature of the intended effect. A one-size-fits-all approach, such as blanket bans applied indiscriminately across cases, may be ill-suited to achieve diverse moderation goals. In contrast,

a more principled, data-driven strategy that aligns intervention types with the behavioural changes they aim to foster offers a promising path forward. In a content moderation landscape that continues to rely heavily on a narrow set of interventions [85, 75], our findings underscore the importance of expanding both the methodological and conceptual toolkits used to design them. Tailoring interventions not just to who the user is, but to what the intervention is meant to accomplish, represents a favourable step toward effective, fair, and goal-sensitive moderation.

6.4. Limitations and Future Work

Our analysis focuses on Reddit's 2020 Great Ban, a large-scale deplatforming intervention [16]. Even if such intervention affected a large number of users and communities, the reliance on a single type of moderation (i.e., a community ban) at a single point in time, nonetheless represents a limitation of our work. Similarly, our results are derived solely from Reddit data, a platform with a large user base, but with specific structural and cultural characteristics. Together, these two factors may limit the generalizability of our findings to other moderation contexts or platforms, though such constraints are standard in empirical computational research on content moderation. Our results are also based on the set of quantification algorithms and user features we experimented with. While we selected well-established methods and drew on an extensive feature set, alternative algorithms or unexplored features could yield additional insights. Moreover, our models identify correlational patterns between user attributes and post-intervention behaviour, but do not support causal claims. Determining whether specific features cause users to react in certain ways would require dedicated experimental or quasi-experimental designs that could be the goal of future research in this domain. Finally, our approach does not incorporate temporal or longitudinal modelling, which presents a further promising direction for future work.

Adding to the previous research directions, our results pave the way for future predictive systems capable of accurately modelling user reactions to moderation, whether via classification, quantification, or regression approaches. A parallel research avenue involves extending this analysis to other forms of moderation and additional platforms, though progress in this direction may be constrained by the limited accessibility of platform data [4]. Towards the medium or long term, our work also sets the stage for the development of moderation strategies that are both context-aware and tailored to specific intervention goals, marking a step toward more nuanced and effective content moderation frameworks.

Acknowledgments

AM's and FS's work has been partially supported by project "Future Artificial Intelligence Research" (FAIR – CUP B53D22000980006), project "Quantification under Dataset Shift" (QuaDaSh – CUP B53D23026250001), and project "Strengthening the Italian RI for Social Mining and Big Data Analytics" (SoBigData.it – CUP B53C22001760006), all funded by the European Union under the NextGenerationEU funding scheme. SC's, BT's, and TF's work has been partly supported by the European Union – Next Generation EU, Mission 4 Component 1, for project PIANO (CUP B53D23013290006) and for the ERC project DEDUCE under grant #101113826.

References

[1] Lorenzo Alvisi, Serena Tardelli, and Maurizio Tesconi. Mapping the Italian Telegram ecosystem: Communities, toxicity, and hate speech. *arXiv preprint arXiv:2504.19594*, 2025.

[2] Jannis Androutsopoulos. Punctuating the other: Graphic cues, voice, and positioning in digital discourse. *Language & Communication*, 88:141–152, 2023.

[3] Anjum and Rahul Katarya. Hate speech, toxicity detection in online social media: A recent survey of state of the art and opportunities. *International Journal of Information Security*, 23(1):577–608, 2024.

[4] Dennis Assenmacher, Derek Weber, Mike Preuss, André Calero Valdez, Alison Bradshaw, Björn Ross, Stefano Cresci, Heike Trautmann, Frank Neumann, and Christian Grimme. Benchmarking crisis in social media analytics: A solution for the data sharing problem. *Social Science Computer Review*, pages 1–27, 2021.

[5] Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244, 2023.

[6] Antonio Bella, Cèsar Ferri, José Hernández-Orallo, and María José Ramírez-Quintana. Quantification via probability estimators. In *Proceedings of the 11th IEEE International Conference on Data Mining (ICDM 2010)*, pages 737–742, Sydney, AU, 2010.

[7] Ryan L. Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin*, 10:1–47, 2022.

[8] Mirko Bunse, Alejandro Moreo, Fabrizio Sebastiani, and Martin Senz. Regularization-based methods for ordinal quantification. *Data Mining and Knowledge Discovery*, 38(6):4076–4121, 2024.

[9] Caitlin Ring Carlson and Luc S Cousineau. Are you sure you want to view this community? Exploring the ethics of Reddit's quarantine practice. *Journal of Media Ethics*, 35(4):202–213, 2020.

[10] Alberto Castaño, Pablo González, Jaime A. González, and Juan J. del Coz. Matching distributions algorithms based on the Earth mover's distance for ordinal quantification. *IEEE Transactions on Neural Networks and Learning Systems*, 35(1):1050–1061, 2024.

[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. Crossmod: A cross-community learning-based system to assist Reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30, 2019.

[12] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert. Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 29(4):1–26, 2022.

[13] Lorenzo Cima, Lorenzo Mannocci, Marco Avvenuti, Maurizio Tesconi, and Stefano Cresci. Coordinated behavior in information operations on Twitter. *IEEE Access*, 2024.

[14] Lorenzo Cima, Alessio Miaschi, Amaury Trujillo, Marco Avvenuti, Felice Dell'Orletta, and Stefano Cresci. Contextualized counterspeech: Strategies for adaptation, personalization, and evaluation. In *The 34th ACM Web Conference (The Web Conf.'25)*, pages 5022–5033. ACM, 2025.

[15] Lorenzo Cima, Benedetta Tessa, Amaury Trujillo, Stefano Cresci, and Marco Avvenuti. Investigating the heterogeneous effects of a massive content moderation intervention via difference-in-differences. *Online Social Networks and Media*, 48:100320, 2025.

[16] Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. The Great Ban: Efficacy and unintended consequences of a massive deplatforming operation on Reddit. In *Companion Publication of the 16th ACM Web Science Conference*, pages 85–93, 2024.

[17] Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. A criticism to society (as seen by Twitter analytics). In *The 34th IEEE International Conference on Distributed Computing Systems Workshops (ICDCS'14 Workshops)*, pages 194–200. IEEE, 2014.

[18] Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. Personalized interventions for online moderation. In *ACM HT*, pages 248–251, 2022.

[19] Elżbieta Drażkiewicz. Study conspiracy theories with compassion. *Nature*, 603(7903):765–765, 2022.

[20] Andrea Esuli, Alessandro Fabris, Alejandro Moreo, and Fabrizio Sebastiani. *Learning to quantify*. Springer Nature, Cham, CH, 2023.

[21] Alessandro Fabris, Andrea Esuli, Alejandro Moreo, and Fabrizio Sebastiani. Measuring fairness under unawareness of sensitive attributes: A quantification-based approach. *Journal of Artificial Intelligence Research*, 76:1117–1180, 2023.

[22] Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. Tweepfake: About detecting deepfake tweets. *PLoS One*, 16(5), 2021.

[23] Sina Fazelpour and Maria De-Arteaga. Diversity in sociotechnical machine learning systems. *Big Data & Society*, 9(1), 2022.

[24] Emilio Ferrara, Stefano Cresci, and Luca Luceri. Misinformation, manipulation, and abuse on social media in the era of COVID-19. *Journal of Computational Social Science*, 3:271–277, 2020.

[25] George Forman. Quantifying counts and costs via classification. *Data Mining and Knowledge Discovery*, 17(2):164–206, 2008.

[26] Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. The anatomy of conspiracy theorists: Unveiling traits using a comprehensive Twitter dataset. *Computer Communications*, 217:25–40, 2024.

[27] Shlok Gilda, Luiz Giovanini, Mirela Silva, and Daniela Oliveira. Predicting different types of subtle toxicity in unhealthy online conversations. *Procedia Computer Science*, 198:360–366, 2022.

[28] Tarleton Gillespie. *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[29] Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. Human and LLM biases in hate speech annotations: A socio-demographic analysis of annotators and targets. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 653–670, 2025.

[30] Matej Gjurkovic, Mladen Karan, Iva Vukojevic, Mihaela Bosnjak, and Jan Snajder. Pandora talks: Personality and demographics on Reddit. In *Proceedings of the 9th International Workshop on Natural Language Processing for Social Media (SocialNLP@NAACL 2021)*, pages 138–152, Online event, 2021.

[31] Jeffrey Gleason, Alex Leavitt, and Bridget Daly. In suspense about suspensions? The relative effectiveness of suspension durations on a popular social platform. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2025.

[32] Pablo González, Alberto Castaño, Nitesh V. Chawla, and Juan José del Coz. A review on quantification learning. *ACM Computing Surveys*, 50(5):74:1–74:40, 2017.

[33] Pablo González, Alejandro Moreo, and Fabrizio Sebastiani. Binary quantification and dataset shift: An experimental investigation. *Data Mining and Knowledge Discovery*, 38(4):1670–1712, 2024.

[34] Pablo González, Alberto Castaño, Emily E. Peacock, Jorge Díez, Juan José Del Coz, and Heidi M. Sosik. Automatic plankton quantification using deep features. *Journal of Plankton Research*, 41(4):449–463, 2019.

[35] Hussam Habib, Maaz Bin Musa, Muhammad Fareed Zaffar, and Rishab Nithyanand. Are proactive interventions for Reddit communities feasible? In *Proceedings of the*

International AAAI Conference on Web and Social Media, volume 16, pages 264–274, 2022.

[36] Hussam Habib and Rishab Nithyanand. Exploring the magnitude and effects of media influence on Reddit moderation. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2022)*, volume 16, pages 275–286, 2022.

[37] Laura Hanu and Unitary team. Detoxify, 2020.

[38] Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. Outcome-constrained large language models for countering hate speech. In *The 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP'24)*, pages 4523–4536, 2024.

[39] Daniel J. Hopkins and Gary King. A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247, 2010.

[40] Manoel Horta Ribeiro, Shagun Jhaver, Jordi Cluet-i Martinell, Marie Reignier-Tayar, and Robert West. Deplatforming norm-violating influencers on social media reduces overall online attention toward them. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–25, 2025.

[41] Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. Do platform migrations compromise content moderation? Evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

[42] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225, 2014.

[43] Thomas Jaenich, Alejandro Moreo, Alessandro Fabris, Graham McDonald, Andrea Esuli, Iadh Ounis, and Fabrizio Sebastiani. Quantifying query fairness under unawareness. arXiv:2506.04140 [cs.IR], 2025.

[44] Shagun Jhaver. Bans vs. warning labels: Examining bystanders' support for community-wide moderation interventions. *ACM Transactions on Computer-Human Interaction*, 32(2):1–30, 2025.

[45] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. Evaluating the effectiveness of deplatforming as a moderation strategy on Twitter. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–30, 2021.

[46] Gary King and Ying Lu. Verbal autopsy methods with multiple causes of death. *Statistical Science*, 23(1):78–91, 2008.

[47] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PloS One*, 10(12):e0144296, 2015.

[48] Agashini V. Kumar, Deepa Gupta, and Manju Venugopalan. Cyberbullying text classification for social media data using embedding and deep learning approaches. In *Proceedings of the 14th International Conference on Computing Communication and Networking Technologies (ICCCNT 2023)*, pages 1–6, 2023.

[49] Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. Watch your language: Investigating content moderation with large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):865–878, May 2024.

[50] Maram Kurdi, Nuha Albadi, and Shivakant Mishra. "Video unavailable": Analysis and prediction of deleted and moderated YouTube videos. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 166–173. IEEE, 2020.

[51] Haewoon Kwak, Jisun An, Elise Jing, and Yong-Yeol Ahn. Frameaxis: Characterizing microframe bias and intensity with word embedding. *PeerJ Computer Science*, 7:e644, 2021.

[52] Lorenzo Mannocci, Michele Mazza, Anna Monreale, Maurizio Tesconi, and Stefano Cresci. Detection and characterization of coordinated online behavior: A survey. *arXiv preprint arXiv:2408.01257*, 2024.

[53] Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Investigating the difference between trolls, social bots, and humans on Twitter. *Computer Communications*, 196:23–36, 2022.

[54] Saif M. Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of The Annual Conference of the Association for Computational Linguistics (ACL 2018)*, Melbourne, AU, 2018.

[55] Saif M. Mohammad. Word affect intensities. In *Proceedings of the 11th Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, JP, 2018.

[56] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaíz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.

[57] Alejandro Moreo. On the interconnections of calibration, quantification, and classifier accuracy prediction under dataset shift. arXiv:2505.11380 [cs.LG], 2025.

[58] Alejandro Moreo, Andrea Esuli, and Fabrizio Sebastiani. QuaPy: A Python-based framework for quantification. In *Proceedings of the 30th ACM International Conference on Knowledge Management (CIKM 2021)*, pages 4534–4543, Gold Coast, AU, 2021.

[59] Alejandro Moreo, Pablo González, and Juan José del Coz. Kernel density estimation for multiclass quantification. *Machine Learning*, 114(4), 2025.

[60] Alejandro Moreo and Fabrizio Sebastiani. Re-assessing the "classify and count" quantification method. In *Proceedings of the 43rd European Conference on Information Retrieval (ECIR 2021)*, volume II, pages 75–91, Lucca, IT, 2021.

[61] Manoj Niverthi, Gaurav Verma, and Srijan Kumar. Characterizing, detecting, and predicting online ban evasion. In *Proceedings of the ACM Web Conference 2022*, pages 2614–2623, 2022.

[62] Gianluca Nogara, Erfan Samieyan Sahneh, Matthew R DeVerna, Nick Liu, Luca Luceri, Filippo Menczer, Francesco Pierri, and Silvia Giordano. A longitudinal analysis of misinformation, polarization and toxicity on BlueSky after its public launch. *arXiv preprint arXiv:2505.02317*, 2025.

[63] Vishnuprasad Padinjaredath Suresh, Gianluca Nogara, Felipe Cardoso, Stefano Cresci, Silvia Giordano, and Luca Luceri. Tracking fringe and coordinated activity on Twitter leading up to the US Capitol attack. In *The 18th International AAAI Conference on Web and Social Media (ICWSM'24)*, pages 1557–1570. AAAI, 2024.

[64] Konstantina Palla, Jos'e Luis Redondo Garc'ıa, Claudia Hauff, Francesco Fabbri, Henrik Lindström, Daniel R. Taber, Andreas Damianou, and Mounia Lalmas. Policy-as-prompt: Rethinking content moderation in the age of large language models. *arXiv preprint arXiv:2502.18695*, 2025.

[65] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Lambretta: Learning to rank for Twitter soft moderation. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 311–326. IEEE, 2023.

[66] Dino Pedreschi, Luca Pappalardo, Emanuele Ferragina, Ricardo Baeza-Yates, Albert-László Barabási, Frank Dignum, Virginia Dignum, Tina Eliassi-Rad, Fosca Giannotti, János Kertész, et al. Human-AI coevolution. *Artificial Intelligence*, 339, 2025.

[67] Francesco Pierri. Drivers of hate speech in political conversations on Twitter: The case of the 2022 Italian general election. *EPJ Data Science*, 13(1):63, 2024.

[68] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset shift in machine learning*. The MIT Press, Cambridge, US, 2009.

[69] Rohan Raut and Francesca Spezzano. Enhancing hate speech detection with user characteristics. *International Journal of Data Science and Analytics*, 18(4):445–455, 2024.

[70] Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002.

[71] Tetsuya Sakai. Comparing two binned probability distributions for information access evaluation. In *Proceedings of the 41st International ACM Conference on Research and Development in Information Retrieval (SIGIR 2018)*, pages 1073–1076, Ann Arbor, US, 2018.

[72] S. Santhiya, S. Uma, N. Abinaya, P. Jayadharshini, S. Priyanka, and M.N. Dharshini. A comparative exploration in text classification for hate speech and offensive language detection using BERT-based and GLOVE embeddings. In *Proceedings of the 2nd International Conference on Disruptive Technologies (ICDT 2024)*, pages 1506–1509, 2024.

[73] Fabrizio Sebastiani. Market research, deep learning, and quantification. In *Presented at the ASC Conference on the Application of Artificial Intelligence and Machine Learning to Surveys*, London, UK, 2018. `http://goo.gl/JvWU7A`.

[74] Alfonso Semeraro, Salvatore Vilella, Riccardo Improta, Edoardo S. De Duro, Saif M. Mohammad, Giancarlo Ruffo, and Massimo Stella. EmoAtlas: An emotional network analyzer of texts that merges psychological lexicons, artificial intelligence, and network science. *Behavior Research Methods*, 57(2):77, 2025.

[75] Gautam Kishore Shahi, Benedetta Tessa, Amaury Trujillo, and Stefano Cresci. A year of the DSA transparency database: What it (does not) reveal about platform moderation during the 2024 European Parliament election. In *ICWSM Workshops*, 2025.

[76] Qinlan Shen and Carolyn P Ros'e. A tale of two subreddits: Measuring the impacts of quarantines on political engagement on Reddit. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 932–943, 2022.

[77] Amila Silva, Pei-Chi Lo, and Ee Peng Lim. On predicting personal values of social media users using community-specific language features and personal value correlation. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2021)*, volume 15, pages 680–690, 2021.

[78] Noah A. Smith and Roy W. Tromble. Sampling uniformly from the unit simplex. Technical report, Johns Hopkins University, 2004. `https://www.cs.cmu.edu/~nasmith/papers/smith+tromble.tr04.pdf`.

[79] Amos Storkey. When training and test sets are different: Characterizing learning transfer. In Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors, *Dataset shift in machine learning*, pages 3–28. The MIT Press, Cambridge, US, 2009.

[80] Dirk Tasche. Class prior estimation under covariate shift: No problem? In *Proceedings of the 2nd International Workshop on Learning to Quantify (LQ 2022)*, pages 11–26, Grenoble, IT, 2022.

[81] Benedetta Tessa, Denise Amram, Anna Monreale, and Stefano Cresci. Improving regulatory oversight in online content moderation. In *ECMLPKDD Workshops*, 2025.

[82] Benedetta Tessa, Lorenzo Cima, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci. Beyond trial-and-error: Predicting user abandonment after a moderation intervention. *Engineering Applications of Artificial Intelligence*, 162:112375, 2025.

[83] Amaury Trujillo and Stefano Cresci. Make Reddit Great Again: Assessing community effects of moderation interventions on r/The_Donald. *Proceedings of the ACM on Human-computer Interaction*, 6(CSCW2):1–28, 2022.

[84] Amaury Trujillo and Stefano Cresci. One of many: Assessing user-level effects of moderation interventions on r/The_Donald. In *Proceedings of the 15th ACM Web Science Conference 2023*, pages 55–64, 2023.

[85] Amaury Trujillo, Tiziano Fagni, and Stefano Cresci. The DSA transparency database: Auditing self-reported moderation actions by social media. In *Proceedings of the ACM on Human-Computer Interaction*, volume 9, pages 1–28, 2025.

[86] Milo Z. Trujillo, Samuel F. Rosenblatt, Guillermo De Anda Jauregui, Emily Moog, Briane Paul V. Samson, Laurent Hébert-Dufresne, and Allison M. Roth. When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban. *arXiv preprint arXiv:2106.16207*, 2021.

[87] Lorenzo Volpi, Alejandro Moreo, and Fabrizio Sebastiani. LEAP: Linear equations for classifier accuracy prediction under prior probability shift. *Machine Learning*, 2025. Forthcoming.

[88] Lorenzo Volpi, Alejandro Moreo, and Fabrizio Sebastiani. QuAcc: Using quantification to predict classifier accuracy under prior probability shift. *Intelligenza Artificiale*, 19(2):141–157, 2025.

[89] Minjia Wang, Pingping Lin, Siqi Cai, Shengnan An, Shengjie Ma, Zeqi Lin, Congrui Huang, and Bixiong Xu. STAND-Guard: A small task-adaptive content moderation model. In *The 31st International Conference on Computational Linguistics (COLING'25)*, pages 1–20, 2025.

[90] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.

[91] Michael Werman, Shmuel Peleg, and Azriel Rosenfeld. A distance metric for multidimensional histograms. *Computer Vision, Graphics, and Image Processing*, 32:328–336, 1985.

[92] Emma J. Williams, Amy Beardmore, and Adam N. Joinson. Individual differences in susceptibility to online influence: A theoretical review. *Computers in Human Behavior*, 72:412–421, 2017.

[93] Kenji Yokotani and Masanori Takano. Effects of suspensions on offences and damage of suspended offenders and their peers on an online chat platform. *Telematics and Informatics*, 68:101776, 2022.

[94] Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. Slm-mod: Small language models surpass LLMs at content moderation. *arXiv preprint arXiv:2410.13155*, 2024.

[95] Andy Zhao and Will Hobbs. The effects and non-effects of social sanctions from user jury-based content moderation decisions on Weibo. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2025.

[96] Albert Ziegler and Paweł Czyż. Bayesian quantification with black-box estimators. *Transactions on Machine Learning Research*, 2024, 2024.

Appendix A. Features

Table A.4 reports the cardinality of each subgroup of features, while Table A.5 provides their description.

group	subgroup	num
EMBEDDINGS	EMBEDDINGS	384
LIWC	LINGUISTIC	21
	SOCIAL	12
	COGNITION	12
	AFFECT	11
	MOTIVATION	9
	SUMMARY	8
	LIFESTYLE	6
	BIO	6
	DRIVES	4
	CULTURE	4
	SPATIAL	4
	TEMPORAL	4
	CONVERSATION	4
WRITING_STYLE	NER	5
	FLESCH-KINKAID	4
	SMOG	4
	SPELL_ERRORS	4
	NOUNS	4
	ARTICLES	4
	ADPOSITIONS	4
	IRONY	4
	NOVELTY	4
	STOPWORDS	4
WRITING_STYLE	SYM	4
	CONG	4
	SCONJ	4
	PROPN	4
	PRON	4
	INTJ	4
	DET	4
	ADV	4
	VERB	4
	AUX	4
ACTIVITY	SUBMISSIONS	14
	COMMENTS	9
	TIME	6
	MENTIONS_LINKS	3
	TRENDS	2
RELATIONAL	ENGAGEMENT	10
	THREADS	10
	AUTHORS	10
	SUBR	6
SENTIMENT	NEG	4
	NEU	4
	POS	4
	CPD	4
	NEG-EMOJI	4
	NEU-EMOJI	4
	POS-EMOJI	4
	SENT-EMOJI	4
SOC_PSY	OCEAN	5
	AUTHORITY	4
	FAIRNESS	4
	GENERAL	4
	HARM	4
	INGROUP	4
	PURITY	4
	DEMOGRAPHIC	2
TOXICITY	TOXICITY	4
	SEVERE-TOXICITY	4
	OBSCENE	4
	THREAT	4
	INSULT	4
	ID-ATTACK	4
EMOTIONS	EIL	8
	EMOSCORES	8
	VAD	3

Table A.4: Feature groups, subgroups, and their cardinality.

Group	Subgroup	Description
EMB	EMBEDDINGS	Numerical vector representations of text generated using a sentence-transformer
LIWC	LINGUISTIC	Linguistic features such as pronouns, articles, function words, and verb tenses
	SOCIAL	Words that reflect social processes and relationships, including terms related to family, friends, and people
	COGNITION	Vocabulary linked to cognitive functions such as insight, causation, certainty, and tentativeness
	AFFECT	Terms expressing affective states like positive and negative emotions, anxiety, anger, and sadness
	MOTIVATION	Words related to drives and needs, such as achievement, power, reward, and affiliation
	SUMMARY	Aggregate measures computed by LIWC, such as analytical thinking, clout, authenticity, and emotional tone
	LIFESTYLE	Terms associated with everyday life domains, such as work, home, money, and leisure activities
	BIO	Words referring to biological processes, including health, body, and sexual content
	DRIVES	Lexical indicators of psychological drives such as affiliation, achievement, and risk
	CULTURE	References to cultural terms including religion, morality, and traditions
	SPATIAL	Words indicating spatial relations and references to locations or movements
	TEMPORAL	Expressions related to time, including past, present, and future references
	CONVERSATION	Lexical items typically used in conversations, such as assent, fillers, and nonfluencies
SENTIMENT	NEGATIVE	Negative sentiment expressed in text, capturing feelings such as sadness, anger, or dissatisfaction
	NEUTRAL	Neutral or objective tone without strong emotional charge
	POSITIVE	Positive sentiment reflected in language conveying happiness, approval, or optimism
	COMPOUND	Overall sentiment score combining positive and negative expressions into a single value
	NEGATIVE EMOJI	Use of emojis conveying negative emotions such as sadness, anger, or frustration
	NEUTRAL EMOJI	Use of emojis expressing neutral or ambiguous emotions
	POSITIVE EMOJI	Use of emojis conveying positive emotions like happiness, love, or encouragement
	SENTIMENT EMOJI	Overall sentiment expressed through emojis combining positive, neutral, and negative cues
TOXICITY	TOXICITY	Presence of toxic language, including insults, offensive words, or harmful content
	SEVERE TOXICITY	Strong forms of toxic content characterized by extreme insults or threats

Table A.5 continues on next page

Table A.5 (continued from previous page)

Group	Subgroup	Description
TOXICITY	OBSCENE	Use of obscene or vulgar language and expressions
	THREAT	Statements containing threats of harm or violence
	INSULT	Language intended to offend or demean another individual or group
	IDENTITY ATTACK	Offensive comments targeting a person's identity attributes such as race, gender, or religion
SOC_PSY	OCEAN	Personality traits based on the Big Five model: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
	AUTHORITY	Words related to respect for authority, hierarchy, and social order
	FAIRNESS	Expressions concerning justice, equity, and fairness in social contexts
	GENERAL	General social psychological terms that do not fit into more specific categories
	HARM	Terms related to harm, suffering, or protection from harm in social situations
	INGROUP	Words denoting belonging to or identification with social groups
	PURITY	Language reflecting concepts of purity, sanctity, and moral cleanliness
	DEMOGRAPHIC	Terms related to demographic characteristics such as age, gender, or ethnicity
ACTIVITY	SUBMISSIONS	Information about users' submissions made on Reddit
	COMMENTS	Information about users' comments made on Reddit
	TIME	Temporal aspects of user activity, including frequency and timing of posts
	MENTIONS_LINKS	Number of comments where users mention other users, communities or include hyperlinks in their content
	TRENDS	Slope representing the increases or decreases in activity and toxicity
RELATIONAL	ENGAGEMENT	Measures of user interaction computed as the difference between upvotes and downvotes to a comment
	THREADS	Conversation chains or discussion threads where users participate
	AUTHORS	Information about how much users interact with other users
	SUBREDDITS	Data related to the specific subreddit communities where users post or comment
EMOTIONS	EIL	Categories of basic emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, trust) identified using the EIL lexicon
	EMOSCORES	Numerical scores quantifying the intensity of emotions (anger, trust, surprise, disgust, joy, sadness, fear, anticipation)
	VAD	Dimensions of affective meaning: Valence (positive–negative), Arousal (calm–excited), Dominance (submissive–dominant)
WR_STYLE	FLESCH-KINKAID	Readability index based on the flesch-kinkaid score
	SMOG	Readability index estimating years of education required to understand text
	SPELL_ERRORS	Frequency of spelling mistakes in text
	VERBS	Distribution of verbs used in comments

Table A.5 continues on next page

Table A.5 (continued from previous page)

Group	Subgroup	Description
WRITING_STYLE	NOUNS	Distribution of nouns used in comments
	ARTICLES	Distribution of articles (e.g., "the", "a") used in comments
	ADVERBS	Distribution of adverbs used in comments
	ADPOSITIONS	Distribution of adpositions (e.g., prepositions, postpositions) used in comments
	PRONOUNS	Distribution of pronouns used in comments
	IRONY	Estimated presence of ironic expressions in text
	NER	Named entity recognition categories (e.g., persons, organizations, geopolitical entities, laws, groups)
	NOVELTY	Measure of textual originality compared to reference corpus
	STOPWORDS	Distribution of stopwords (common non-content words such as "and", "the", "is")
	SYM	Distribution of symbols present in text (e.g., %, $,)
	CONG	Distribution of coordinating conjunctions (e.g., "and", "but")
	SCONJ	Distribution of subordinating conjunctions (e.g., "because", "although")
	PROPN	Distribution of proper nouns in text
	PRON	Distribution of pronouns in text
	INTJ	Distribution of interjections (e.g., "wow", "oh")
	DET	Distribution of determiners (e.g., "this", "those")
	ADV	Distribution of adverbs (fine-grained category from POS tagging)
	VERB	Distribution of verbs (fine-grained category from POS tagging)
	AUX	Distribution of auxiliary verbs (e.g., "is", "have", "will")

Table A.5: Brief description of each subgroup of features.

Appendix B. Quantification Performance of the Subgroups of Features

Table B.6 reports the order of inspection of the feature subgroups for each task. This order is used during the greedy feature selection process described in Algorithm 1.

	activity		toxicity		diversity	
rank	feature name	MNMD	feature name	MNMD	feature name	MNMD
1	SOC:AUTHORITY	1657.735	REL:ENGAGEMENT	1368.431	SOC:AUTHORITY	1311.063
2	LIW:CULTURE	1644.752	REL:AUTHORS	1366.133	REL:AUTHORS	1271.090
3	WRI:CONG	1599.881	LIW:COGNITION	1328.143	SOC:FAIRNESS	1265.083
4	WRI:SYM	1585.881	REL:THREADS	1316.311	LIW:SUMMARY	1262.627
5	LIW:DRIVES	1583.815	LIW:BIO	1312.633	LIW:COGNITION	1234.172
6	SOC:HARM	1564.193	LIW:PUNCTUATION	1311.661	LIW:LINGUISTIC	1231.092
7	ACT:MENTIONS_LINKS	1557.762	EMO:EIL	1279.451	WRI:SMOG	1227.331
8	LIW:MOTIVATION	1540.192	SOC:DEMOGRAPHIC	1270.266	WRI:PRON	1224.617
9	EMO:VAD	1520.968	EMO:EMOSCORES	1260.259	REL:THREADS	1215.066
10	LIW:PUNCTUATION	1515.679	ACT:COMMENTS	1254.685	EMO:EIL	1205.383
11	LIW:TEMPORAL	1514.538	LIW:LINGUISTIC	1241.606	ACT:SUBMISSIONS	1205.228

Table B.6 continues on next page

Table B.6 (continued from previous page)

rank	activity		toxicity		diversity	
	feature name	MNMD	feature name	MNMD	feature name	MNMD
12	EMO:EIL	1508.496	SEN:POS-EMOJI	1220.949	WRI:FLESCH-KINKAID	1198.309
13	SOC:OCEAN	1498.040	LIW:SOCIAL	1218.462	WRI:NER	1195.571
14	SEN:NEU-EMOJI	1475.566	LIW:CONVERSATION	1212.394	SOC:PURITY	1190.555
15	SOC:INGROUP	1469.145	ACT:SUBMISSIONS	1212.151	LIW:DRIVES	1185.555
16	SEN:SENT-EMOJI	1469.101	WRI:NOVELTY	1211.304	ACT:TRENDS	1181.687
17	SEN:POS-EMOJI	1464.278	WRI:NER	1209.736	REL:ENGAGEMENT	1179.924
18	WRI:AUX	1445.009	ACT:TIME	1208.817	SEN:NEG	1174.586
19	SEN:NEG-EMOJI	1443.660	SOC:GENERAL	1205.825	WRI:SCONJ	1168.685
20	WRI:SCONJ	1436.706	SOC:OCEAN	1204.764	ACT:COMMENTS	1168.595
21	LIW:SPATIAL	1430.364	WRI:DET	1201.019	SOC:GENERAL	1167.661
22	ACT:TIME	1425.299	LIW:MOTIVATION	1199.623	WRI:SYM	1162.521
23	TOX:SEVERE-TOXICITY	1420.325	LIW:LIFESTYLE	1195.578	SOC:OCEAN	1158.915
24	LIW:LIFESTYLE	1413.914	LIW:AFFECT	1194.761	TOX:ID-ATTACK	1158.599
25	SOC:DEMOGRAPHIC	1408.487	LIW:SPATIAL	1190.505	EMO:VAD	1158.457
26	WRI:FLESCH-KINKAID	1381.244	EMO:VAD	1190.196	EMO:EMOSCORES	1158.257
27	WRI:NER	1377.075	WRI:STOPWORDS	1187.248	LIW:SOCIAL	1156.770
28	EMO:EMOSCORES	1375.618	LIW:CULTURE	1187.248	SOC:INGROUP	1156.378
29	TOX:OBSCENE	1375.335	SEN:NEU-EMOJI	1185.784	LIW:BIO	1155.709
30	SEN:POS	1368.405	REL:SUBR	1184.372	TOX:THREAT	1155.021
31	SEN:CPD	1365.942	WRI:ARTICLES	1182.534	WRI:AUX	1152.659
32	LIW:COGNITION	1365.580	WRI:SYM	1180.736	ACT:TIME	1151.716
33	WRI:SMOG	1361.447	SEN:NEG-EMOJI	1178.618	WRI:ARTICLES	1150.436
34	SOC:FAIRNESS	1360.234	WRI:NOUNS	1177.560	SOC:HARM	1149.303
35	LIW:LINGUISTIC	1357.386	WRI:ADV	1175.680	LIW:PUNCTUATION	1145.945
36	LIW:CONVERSATION	1350.528	ACT:TRENDS	1172.241	SEN:NEG-EMOJI	1145.679
37	WRI:PRON	1348.809	WRI:FLESCH-KINKAID	1168.360	SEN:CPD	1144.436
38	WRI:SPELL_ERRORS	1348.611	WRI:CONG	1168.104	SEN:NEU-EMOJI	1144.349
39	LIW:AFFECT	1348.413	SOC:PURITY	1163.298	SEN:SENT-EMOJI	1143.934
40	WRI:INTJ	1346.151	SEN:SENT-EMOJI	1157.149	WRI:ADV	1142.218
41	WRI:IRONY	1345.676	SOC:AUTHORITY	1157.075	WRI:CONG	1140.672
42	TOX:THREAT	1339.511	WRI:PROPN	1156.584	ACT:MENTIONS_LINKS	1139.670
43	WRI:ADV	1338.324	WRI:SPELL_ERRORS	1154.185	LIW:MOTIVATION	1139.582
44	SEN:NEG	1337.666	WRI:IRONY	1151.980	WRI:DET	1138.912
45	SOC:GENERAL	1331.506	ACT:MENTIONS_LINKS	1150.571	TOX:INSULT	1137.612
46	WRI:VERB	1318.154	WRI:SMOG	1146.108	SOC:DEMOGRAPHIC	1137.423
47	WRI:PROPN	1313.759	LIW:DRIVES	1145.194	LIW:AFFECT	1136.799
48	ACT:COMMENTS	1310.718	SOC:HARM	1143.599	TOX:SEVERE-TOXICITY	1136.324
49	SEN:NEU	1305.519	WRI:PRON	1143.352	SEN:POS-EMOJI	1134.120
50	WRI:NOUNS	1304.647	LIW:TEMPORAL	1141.857	LIW:CULTURE	1133.908
51	WRI:STOPWORDS	1300.541	WRI:AUX	1141.625	TOX:OBSCENE	1133.743
52	ACT:SUBMISSIONS	1292.198	SOC:INGROUP	1136.889	TOX:TOXICITY	1133.316
53	SOC:PURITY	1274.784	SEN:POS	1136.698	LIW:LIFESTYLE	1130.407
54	ACT:TRENDS	1269.451	SOC:FAIRNESS	1134.151	SEN:NEU	1128.792
55	WRI:NOVELTY	1269.434	WRI:INTJ	1132.721	LIW:TEMPORAL	1128.338
56	WRI:ADPOSITIONS	1264.317	WRI:VERB	1121.263	WRI:IRONY	1126.011
57	WRI:DET	1253.132	WRI:SCONJ	1115.567	WRI:NOVELTY	1124.884

Table B.6 continues on next page

Table B.6 (continued from previous page)

	activity		toxicity		diversity	
rank	feature name	MNMD	feature name	MNMD	feature name	MNMD
58	LIW:SUMMARY	1250.382	LIW:SUMMARY	1113.113	WRI:INTJ	1124.002
59	WRI:ARTICLES	1238.723	SEN:NEU	1094.396	WRI:SPELL_ERRORS	1120.098
60	TOX:INSULT	1232.596	WRI:ADPOSITIONS	1085.154	LIW:SPATIAL	1118.784
61	LIW:SOCIAL	1219.891	TOX:THREAT	1060.968	WRI:STOPWORDS	1117.319
62	TOX:TOXICITY	1214.411	SEN:CPD	1049.432	WRI:VERB	1111.365
63	LIW:BIO	1212.964	EMB:HIDDEN	948.059	WRI:PROPN	1110.323
64	REL:SUBR	1155.066	TOX:ID-ATTACK	922.653	WRI:ADPOSITIONS	1100.310
65	REL:AUTHORS	1142.419	SEN:NEG	908.108	LIW:CONVERSATION	1099.490
66	TOX:ID-ATTACK	1129.271	TOX:SEVERE-TOXICITY	889.121	SEN:POS	1095.805
67	REL:THREADS	1110.959	TOX:OBSCENE	825.793	WRI:NOUNS	1093.251
68	EMB:HIDDEN	1025.402	TOX:TOXICITY	792.252	REL:SUBR	1082.163
69	REL:ENGAGEMENT	959.841	TOX:INSULT	787.978	EMB:HIDDEN	986.320

Table B.6: Feature subgroup order of inspection for each task. Feature subgroups are color-coded according to their parent group (identified by the prefix before the colon).