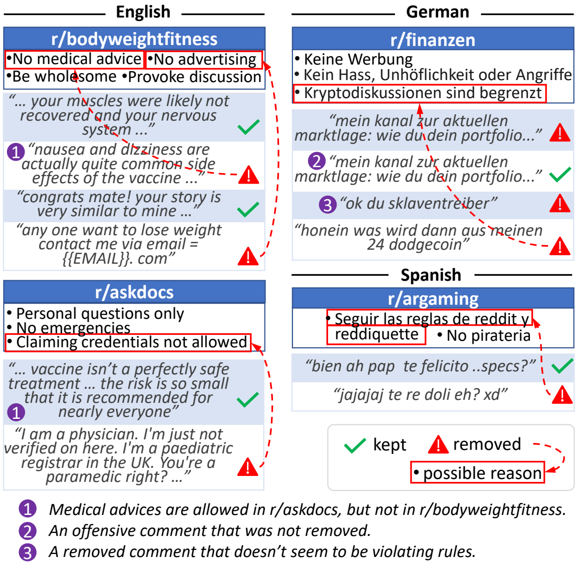
Multilingual Content Moderation: A Case Study on Reddit

Meng Ye[1], Karan Sikka[1], Katherine Atwell[2], Sabit Hassan[2]
Ajay Divakaran[1], Malihe Alikhani[2]
[1]Center for Vision Technologies, SRI International
[2]Department of Computer Science, University of Pittsburgh
{first.last}@sri.com, {kaa139, sah259, malihe}@pitt.edu

Abstract

Content moderation is the process of flagging content based on pre-defined platform rules. There has been a growing need for AI moderators to safeguard users as well as protect the mental health of human moderators from traumatic content. While prior works have focused on identifying hateful/offensive language, they are not adequate for meeting the challenges of content moderation since 1) moderation decisions are based on violation of rules, which subsumes detection of offensive speech, and 2) such rules often differ across communities which entails an adaptive solution. We propose to study the challenges of content moderation by introducing a multilingual dataset of **1.8 Million** Reddit comments spanning **56** subreddits in English, German, Spanish and French[1]. We perform extensive experimental analysis to highlight the underlying challenges and suggest related research problems such as cross-lingual transfer, learning under label noise (human biases), transfer of moderation models, and predicting the violated rule. Our dataset and analysis can help better prepare for the challenges and opportunities of auto moderation.



r/bodyweightfitness
- No medical advice
- No advertising
- Be wholesome
- Provoke discussion

"... your muscles were likely not recovered and your nervous system ..." ✓

1 "nausea and dizziness are actually quite common side effects of the vaccine ..." ⚠

"congrats mate! your story is very similar to mine ..." ✓

"any one want to lose weight contact me via email = {{EMAIL}}. com" ⚠

r/askdocs
- Personal questions only
- No emergencies
- Claiming credentials not allowed

"... vaccine isn't a perfectly safe treatment ... the risk is so small that it is recommended for nearly everyone" 1 ✓

"I am a physician. I'm just not verified on here. I'm a paediatric registrar in the UK. You're a paramedic right? ..." ⚠

r/finanzen
- Keine Werbung
- Kein Hass, Unhöflichkeit oder Angriffe
- Kryptodiskussionen sind begrenzt

"mein kanal zur aktuellen marktlage: wie du dein portfolio..." ⚠

2 "mein kanal zur aktuellen marktlage: wie du dein portfolio..." ✓

3 "ok du sklaventreiber" ⚠

"honein was wird dann aus meinen 24 dodgecoin" ⚠

r/argaming
- Seguir las reglas de reddit y reddiquette
- No pirateria

"bien ah pap te felicito ..specs?" ✓

"jajajaj te re doli eh? xd" ⚠

1 *Medical advices are allowed in r/askdocs, but not in r/bodyweightfitness.*
2 *An offensive comment that was not removed.*
3 *A removed comment that doesn't seem to be violating rules.*

Figure 1: Content moderation on Reddit is challenging as it 1) depends on rules of each community, 2) requires contextualized understanding of the comments, and 3) is affected by moderator biases.

1 Introduction

Being able to moderate user-generated content is critical for online social media platforms. Several platforms employ human moderators to monitor user content to prevent the spread of misinformation, adverse effects of hateful speech, fraud, etc (Geiger and Ribes, 2010; Dosono and Semaan, 2019; Wang et al., 2022; Jhaver et al., 2017). The moderators' task is to remove improper content and/or suspend users posting such content. However, reviewing and moderating each user comment is practically infeasible due to limited resources, especially during time-critical and large-scale events. More importantly, such moderation work can cause damage to moderators' mental health due to burnout from volunteer work and exposure to harmful content (Gillespie, 2020; Roberts, 2014; Dosono and Semaan, 2019; Wohn, 2019).

In recent years, researchers have put effort into collecting hate speech or offensive language datasets from social platforms such as Twitter (Zampieri et al., 2019; Gautam et al., 2020; Golbeck et al., 2017), YouTube (Dinakar et al., 2011), or a mixture of those platforms (Kennedy et al., 2017; Bhattacharya et al., 2020). These datasets contain annotations of hateful speech and are used to train NLP systems to remove harmful comments (Pamungkas et al., 2020; Chung et al., 2019; Ranasinghe and Zampieri, 2021). We refer to this problem as Offensive Language Identification (OLI). Instead of simply removing such content or suspending users, some works propose countering hate speech with discourse-aware style

[1] https://github.com/mye1225/multilingual_content_mod.git

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3828–3844
May 2-6, 2023 ©2023 Association for Computational Linguistics

transfer (Atwell et al., 2022), expert-based counter-narratives (Chung et al., 2019), by learning to intervene with conversational context (Qian et al., 2019), or by proposing a holistic conceptual framework (Chaudhary et al., 2021).

However, despite remarkable progress in OLI, it is not sufficient to tackle content moderation for two key reasons: (1) OLI is a subset of content moderation, as the latter involves flagging content that is not only hateful but also violates the rules of the platform. For example, Reddit has site-wise rules that not only requires users to be civil, but also prohibits actions such as posting illegal content, spamming, and revealing personal information[2]. (2) Prohibiting offensive/hateful speech is universal to all communities (Chandrasekharan et al., 2018), while content moderation requires a system to be adaptive to rules that change dynamically across different communities. It is often the case that content that is allowed in one subreddit might be disallowed in another. For example, it is allowed to post medical advice on `r/askdocs` while it is prohibited on `r/bodyweightfitness`. We thus believe it is useful to study Reddit comments from a moderation perspective as we can collect data from multiple communities and measure the effectiveness of such systems in adapting to community guidelines.

In this work, we aim to bridge this gap by introducing a dataset of Reddit comments collected from 56 communities for studying content moderation. We propose a challenge for content moderation and demonstrate the limitations of the state-of-the-art models. Our dataset will be available to researchers who agree to the terms and conditions of our data-sharing policy approved by the ethical committee of our institution. Our contributions are:

- We propose a multilingual dataset for content moderation. The dataset consists of **1.8 million** comments from **56** subreddits with moderation labels, that specify whether a comment was kept or removed by the moderator. We also provide the meta-data for each subreddit that includes its name, description, and rules.

- We show that existing offensive speech datasets are not suitable for the content moderation task because only a small portion of the removed comments are offensive. As such, the models trained on OLI datasets fail to identify removed comments that are non-offensive.

- We study the performance of models trained on moderation data under different cross-lingual settings including multilingual-language model, translate-train, and translate-test.

- We provide insights on what makes content moderation a challenging task and discuss potential research problems that can be explored with our proposed dataset.

Dataset	Platform	Type	Size	Language
OLID	T	O&H	14K	en
SWAD	T	SW	1.5K	en
OffensEval	T,R,F	O&H	9M	en, da, tr, ar, el
HatEval	T	O&H	19.6K	en, es
CDUC	Y	CYB	4.6K	en
CONAN	T	O&H	15K	en, fr, it
Norms	R	MOD	4M	en
Ours	R	MOD	1.8M	en, de, es, fr

Table 1: Comparison with published datasets. OLID (Zampieri et al., 2019). SWAD (Pamungkas et al., 2020). OffensEval-2020 (Zampieri et al., 2020). HatEval (Basile et al., 2019). CONAN (Chung et al., 2019). CDUC (Dadvar et al., 2013). Norms (Chandrasekharan et al., 2018). Platforms: (T)witter, (F)acebook, (R)eddit, (Y)outube. Types: (O)ffensive and (H)ate speech, (SW)earing, (CYB)er Bullying, (MOD)eration.

2 Related Work

2.1 Offensive Language and Hate Speech Datasets

Many existing works that have collected user comments from online social platforms. Zampieri et al. (2019) proposed Offensive Language Identification Dataset (OLID), which models not only different types of offensive language, but also the target of offensive messages in a hierarchical structure. The Swear Words Abusiveness Dataset (SWAD) developed by Pamungkas et al. (2020) is a collection of tweets selected from OLID that focuses on predicting abusiveness of a swear word in a tweet context. Chung et al. (2019) created a multilingual dataset with hate speech/counter-narrative pairs provided by experts. The idea is to fight online hate speech content with informed textual responses instead of the standard method of removing content or suspending users. OffensEval (Zampieri et al., 2020) is an offensive language identification challenge that has attracted multiple research teams. It is

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

based on the same hierarchical three-level annotations from the aforementioned OLID dataset in multiple languages. CAD (Vidgen et al., 2021) is a recent recently proposed dataset of Reddit posts and comments with manually annotated two-level abusive categories. A brief comparison between these datasets and ours is shown in Table 1.

All the mentioned works focus on detecting offensive speech. However, online content moderation involves detection of comments that violate community rules in addition to those that are offensive. For example, moderators will often remove comments that are self-promoting, spamming, or off-topic because they do not provide useful information and are harmful for the communication environment (e.g., *"I can help you, see my youtube channel"*). In both of these cases, a model trained to detect hate speech will likely fail.

2.2 Reddit Rules and Content Moderation

A pre-existing related work by Chandrasekharan et al. (2018) trained 100 subreddit classifiers on removed/un-removed comments, and used clustering analysis to discover three types of *implicit norms* from the removed comments: macro norms that are universal to all subreddits (e.g., hate speech, personal attacks), meso norms that applies to subgroups (e.g., meme responses), and micro norms that are specific to individual subreddits (e.g., offering commerce tips). Our study is different in that we focus on moderation task that depends on *explicit community rules* written by moderators. Another work (Samory, 2021) studied the problem of identifying comments *approved* by moderators. They found that approved comments and removed ones actually share many traits such as toxicity and insults, and that it is hard to distinguish them. Fiesler et al. (2018) studied rules from $1,000$ subreddits and found that rules are highly dependent on the context of each individual subreddit while sharing common characteristics across the platform.

One key limitation of these studies is that they only focused on subreddits in the English language and discarded all non-English subreddits. This could be because most data on Reddit is in English. In our work we also select subreddits in non-English languages, i.e. German, French, and Spanish. We then study the possibility of transferring a moderation model trained on subreddits in English (which generally has more content). We argue that a good AI moderator should not only focus on data in English language, but also leverage that knowledge to improve performance on other low-resource languages. While Hassan et al. (2022) study *human* moderation bias across different languages and cultures, our work focuses on the problem of *automated* moderation.

3 Dataset

3.1 Subreddit selection

We collected data from 56 subreddits[3] based on the following criteria:

Wide range of topics: Generally the popularity (and thus data points) of a subreddit depends on its topic. Being able to cover many topics will enable us to better estimate the generalizability of machine learning models on this task. The topics include news, politics, finance, sports, electronics, etc.

Subreddits on similar topics: We chose subreddits with similar topics to enable inquiry into questions relating to transferability of moderation models. For example, "Do models trained on one subreddit transfer to other subreddits?", "will models transfer better between subreddits on similar topics?", and "which subreddit should I train the model on so that it could also be better adapted to X?". Some example of similar subreddits are `r/news` and `r/worldnews`, `r/finance` and `r/personalfinance`, `r/anime` and `r/naruto`, `r/games` and `r/xboxseriesx`.

Multilingual data: Most groups on Reddit are in English, but there are also some in other languages such as Spanish, German, French, Arabic. To extend moderation models to multilingual settings we also selected a number of non-English subreddits. For example French (`r/quebec`, `r/france`, `r/moi_dlvv`), German (`r/de`, `r/finanzen`, `r/ich_iel`), and Spanish (`r/argentina`, `r/gaming`).

3.2 Data Collection Pipeline

We built our data collection pipeline based on the Python Reddit API Wrapper (PRAW) [4], which streams comments and submissions in real time from multiple subreddits. For each data record we store important fields such as the ID, author, posting time, and comment body. We follow a two step approach where we first scrape data for a week, and then check if the comment/submission has been removed or retained by the moderator or if it has

[3] A full list of all 56 subreddits can be found in Table 8.

[4] https://github.com/praw-dev/praw

split	language	#sub	#comments	removal
		Training data		
en-train	English	48	1,347,611	1.87%
en-val	English	48	74,885	1.83%
		Test data		
de	German	3	177,046	0.86%
es	Spanish	2	95,586	0.66%
en	English	48	74,893	1.85%
fr	French	3	49,780	0.23%

Table 2: Number of subreddits, comments, and percentage of removed comments in each data split.

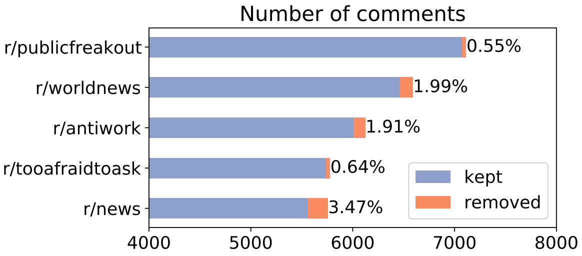


Figure 2: Number of kept and removed comments in the top 5 subreddits (% removed comments in also shown).

been deleted by the author. We use this two-step approach since once the comment is removed/deleted it does not retain its original content. Figure 5 in the appendix shows the overall procedure of our data collection pipeline.

3.3 Dataset Overview

We use the data collection pipeline to collect ~ 1.8 Million comments in a time span of three weeks.

For benchmarking, we split all of the English data randomly into 90%/5%/5% as train, validation, and test subsets. We chose all the non-English data to be part of the test set to study the cross-lingual transferability of models trained on data in English. Table 2 lists the number of subreddits, total number of comments, and the percentage of removed comments for all the data splits.

English data makes up a large proportion of the entire dataset, and is collected from 48 subreddits on different topics. For non-English languages, there are significantly fewer subreddits and we choose the most popular ones for each language. This dataset is highly imbalanced in that only a small proportion of the comments are removed by the moderators: English data has a removal rate of around 1.8% and for German, Spanish, and French it is 0.86%, 0.66%, and 0.23%, respectively. This makes the task of identifying moderated comments very challenging. For comparison, the proportion of offensive examples is around 33% in OLID, 12.5% $\sim$ 28.4% across different languages in OffensEval-2020 (Zampieri et al., 2020), and 28.4% $\sim$ 50.0% in HASOC-2020 (Mandl et al., 2021). We argue that the high percentage of offensive examples in those datasets makes them less suitable for real world applications because these offensive posts make up only a small share of moderated content.

We also observe that different subreddits have different rates of removal, from the lowest of 0.0% to the highest of 21.74%. These differences are likely explained by the discrepancies in topic and subreddit rules, as well as the level to which the moderators enforce these rules. However, the overall rate of removal is quite low (1.84%) compared to existing datasets mentioned above. We show the statistics of the top 5 active subreddits in Figure 2 (refer to Figure 6 in the appendix for more details).

Each subreddit is associated with a set of rules that are enforced by the moderators, which are normally displayed in the *About* section. Typically, each rule has a short title and a longer description to explain in more detail the types of content that are prohibited. For each subreddit, we have collected all of its rules including titles and descriptions (see Figure 7 in the Appendix for the distribution of the number of rules in our dataset).

3.4 Manual Analysis of Offensiveness

To better understand the level of offensiveness in the content moderation task, we manually annotated 1,238 comments (around 200 removed and 100 unremoved examples for each of the four languages) using the fine-grained taxonomy of offensiveness presented in Mubarak et al. (2022), with the addition of the categories of sexuality and age. The distribution of comments for different categories is shown in Figure 3. We observed that most of the comments (71.86% of removed and 80.115% of non-removed) are not offensive. Many of these non-offensive comments criticized the corresponding subreddit or disagreed with the views or goals of the subreddit. For instance, the comment *"i was disappointed when i got the halo console..."* was removed from the `r/halo` subreddit. The comment is not offensive but could have been removed due to its criticism of a product supported by the `r/halo` subreddit. Since these rules, views and goals can vary significantly across subreddits, it explains why content moderation is hard for machine learning classifiers and our moderation classifiers achieve lower scores compared to offensiveness

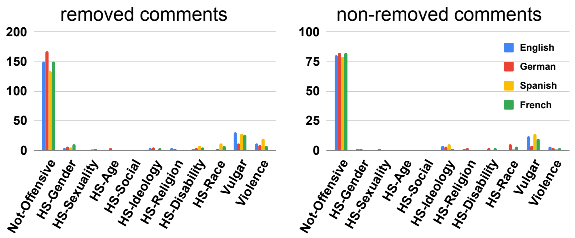


Figure 3: Distribution of different types of offensive speech for Reddit comments.

classifiers. We also see that the distributions of different types of offensive speech are similar across both removed and non-removed comments as well as the four different languages. To conclude, moderating online content is a very different task compared to identifying offensive language and hate speech and a new dataset is needed to train models and study automated content moderation.

4 Experiments

We study the task of content moderation on Reddit by formulating it as a binary classification task: given a user comment x predict whether it should be removed by moderators using $y = f(x)$, where f is the classifier, and y is the removal probability. We set the binary labels based on whether the comment was removed or retained by the moderator (Section 3.2). We report both AUC (area under the ROC curve) and F1 scores.

4.1 Evaluation with Existing Offensive Language Identification Models

A naive solution to moderation is to cast this problem as offensive language detection and directly use classifiers trained on OLI datasets. We thus evaluate models trained on OLI datasets for moderation. This experiment will help in answering two questions: 1) can these models recognize offensive comments in the context of Reddit content moderation?, and 2) do these models miss comments that should be removed but are not-offensive? Being able to answer these question will help us understand whether content moderation can be solved by using OLI or it is a necessity to collect data tailored for moderation. We investigate the above question using three models: RoBERTa (base) (Liu et al., 2019) finetuned on the HatEval (Basile et al., 2019) and OLID dataset (Zampieri et al., 2020), and an XLM-RoBERTa model fine-tuned on OLID. We compute the moderation score for each comment by averaging the output probabilities from these models.

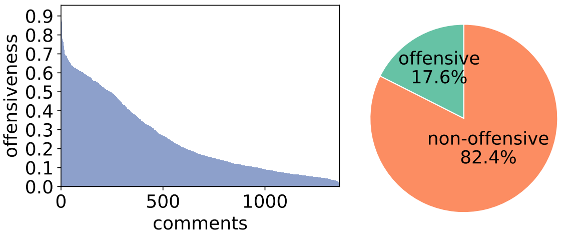


Figure 4: Distribution of offensiveness scores for removed comments (left plot shows sorted scores). Most of the removed comments are not very offensive. Only 18% of them get offensive scores higher than 0.5.

Model	Tr data	T-ori	R-off	R-mod
RoBERTa	HatEval	48.72	44.12	50.59
RoBERTa	OLID	81.25	79.20	47.76
XLM-RoBERTa	OLID	75.81	75.82	48.45

Table 3: Offensive language/hate speech classifiers with their original training (Tr) Twitter dataset and macro-F1 scores on the test split of the same dataset (T-ori), a subset of our Reddit data with offensive labels (R-off), and our Reddit data with moderation labels (R-mod).

Figure 4 shows the distribution of offensiveness scores of the removed comments. The left and the right plot show the sorted scores and the percentage of offensive and non-offensive comments using a threshold of 0.5, respectively. We observe that most of the removed comments (82.4%) are not offensive, and also note that this number matches with the manual analysis in Section 3.4.

By qualitatively studying these examples we observe that 1) OLI models can detect offensiveness in the comments, and 2) not all removed comments are offensive, and OLI models tend to miss these. For instance, a comment stating *"try contacting drewcybersupport on instagram about ..."* from `r/minecraft` was likely removed due to self-promotion (the average offensiveness of this comment is 0.01). More examples can be found in Table 9. We also evaluate the performance of these models on three datasets: the original offensive datasets that they were fine-tuned on (T-ori), a subset of our collected Reddit data with manual offensiveness annotations (R-off), and our proposed moderation dataset (R-mod). From the results in Table 3 we can observe that: 1) these models achieve good performance[5] on both Twitter and Reddit data when the task is to identify offensive language, and

[5] The model fine-tuned on the HatEval dataset has a low F1 score possibly because it focused on hate speech targeted at immigrants and women instead of generic OLI.

Setting	Language	Encoder	AUC				F1-macro			
			en	de	fr	es	en	de	fr	es
MLLM	multilingual	XLM-RoBERTa	66.83	69.42	64.25	63.90	49.60	49.78	49.94	49.83
Translate-train	German	bert-base-german-uncased	60.33	63.34	-	-	49.53	49.78	-	-
	French	flaubert-base-uncased	53.47	-	49.96	-	49.53	-	49.94	-
	Spanish	bert-base-spanish-uncased	64.16	-	-	63.29	49.58	-	-	49.99
Translate-test	English	roberta-base	67.38	71.23	71.61	64.33	49.66	50.01	50.36	50.22

Table 4: Experimental results on moderation classifier under three settings: 1) MLLM: multilingual language model embeddings, 2) translate-train, and 3) translate-test.

Setting	Language	Encoder	AUC				F1-macro			
			en	de	fr	es	en	de	fr	es
MLLM	multilingual	XLM-RoBERTa	66.14	70.10	68.32	59.81	59.97	64.46	61.15	56.78
Translate-train	German	bert-base-german-uncased	61.10	65.90	-	-	57.24	61.15	-	-
	French	flaubert-base-uncased	53.64	-	53.71	-	51.99	-	55.10	-
	Spanish	bert-base-spanish-uncased	63.08	-	-	60.52	59.33	-	-	56.87
Translate-test	English	roberta-base	66.94	70.66	71.57	61.96	61.96	64.16	64.87	58.11

Table 5: Experimental results with balanced data splits from our moderation dataset.

2) their performance drops significantly when evaluated on the moderation task.

Based on the results from both our qualitative and quantitative analyses, we conclude that there is a significant mismatch between the goals of content moderation and offensive language identification. We thus believe that it is necessary to have a moderation-specific dataset to train classifiers for content moderation.

4.2 Evaluation with Models Trained on Proposed Dataset

We use our collected data with moderation labels to train the classifier f. We use pre-trained transformer based language models as text encoders and add a shallow classifier on top: $y = f_m(f_{enc}(x))$, where f_{enc} is the encoder and f_m is the moderation classifier. To deal with multilingual data, we need to either use a multilingual encoder, that can work with inputs in any language (MLLM), or use a monolingual encoder with machine translation models[6]. We only transfer models trained on English data to non-English data and not the other way since the number of data points in English are much larger and it could be beneficial to leverage it for other languages with less data available.

We summarize our experimental results[7] in Table 11 and observe that the RoBERTa (67.38) and XLM-RoBERTa (66.83) models outperform offensive language detectors (second model in Table 3 achieved 61.11 on AUC) on the English data. This is expected because the moderation training data contains removed comments that are non-offensive, and thus the trained classifiers are able to capture these and make better predictions. We also observe that translate-train performs lower than translate-test in most cases, possibly due to the fact that noise is often introduced during translation and the negative impact is more severe in the training stage than in the test stage. We note that although the best performance on most of the test splits is achieved by translate-test, MLLM performance is close to the best in most cases. Considering the challenges of maintaining an extra translation model for every language in real-world applications, MLLM might be a better solution towards building multilingual auto moderators.

We would like to note that the (macro) F1 scores of these models are close to 50 since the classifier/predictions are overwhelmed by data imbalance (less than 2% of comments are labeled to be removed). We thus additionally report AUC scores since, being a ranking based metric, it is not affected by data imbalance. For comparison, we also propose additional data splits[8] with balanced classes and repeat the same experiments. As shown in Table 12, the performance trends for the balanced set are similar to the original splits, and the F1 scores are clearly above 50, ranging from 55.10 to 64.46 with one exception of 51.99, which is

[6] Following Artetxe et al. (2020), we use two settings with machine translation: 1) *Translate-test*: a classifier is trained on English data, and each non-English language needs to be translated into English during test, and 2) *Translate-train*: translate the English data into each target non-English language, and then do both training and test in that target language. For translation we used models from Tiedemann and Thottingal (2020)

[7] Averaged over 3 runs with different random seeds.

[8] Statistics of the balanced splits can be found in Table 10.

Group	Examples	GT/f_o/f_m	freq.
1)	*"every f****** one of those dogs ought to be put down.", "my god this sub is a perpetual b***h fest. clearly time to unsub"*	Y/Y/Y	7.35%
2)	*"prove it by sending me $5 000.00. that would be nothing for you... so go ahead and do it ...", "how do you explain exploding infections and new variants coming out of south africa where 90% of the population is vaccinated?"*	Y/N/Y	16.95%
3)	*"f**k {{name}} and his shorts.", "she has such a punchable face"*	Y/Y/N	1.85%
4)	*"was....was this written by a cat", "depends on the market. {{organization}} sold more units than switch last month in the uk."*	Y/N/N	23.85%
5)	*"oh my god you're pathetic lmao", "well you are a moron so i guess it could be worse"*	N/Y/Y	2.15%

Table 6: Different groups of comments from the English test data. 'GT', f_o, and f_m are ground truth label, offensiveness score and moderation score, respectively. "Y" stands for offensive/removed and "N" stands for non-offensive/non-removed. "freq." denote the proportion of each group in the whole subset (2000 samples).

likely caused by low translation quality in French.

4.3 Analysis on Error Types

We performed additional analysis to understand the errors made by the learned classifier by randomly sampling $1,000$ examples from both removed and non-removed comments and computing their offensiveness[9] (f_o) and moderation scores (f_m). We then organize all comments into groups based on their ground truth label and classifier predictions. Table 6 shows five of those groups and their frequencies: 1)There are 7.35% examples which are offensive and both f_o and f_m were able to identify them. 2) There are 16.95% examples on which f_o failed due to their low offensiveness while f_m was able to identify them successfully. 3) f_m missed a small portion of examples (1.85%) that are offensive and should be removed (*model-error*). 4) There is a significant portion of data points (23.85%) that were removed by moderators while both classifiers predicted them to be fine (*model-error*). This group represents the majority of removed comments that have violated rules other than the use offensive language. 5) Both classifiers do not agree with the ground truth label on 2.15% of the comments, most of them are highly offensive and should be removed but were actually approved by moderators. These instances indicate noisy labels in our dataset that are possibly caused by moderators' biases (*human-error*). Based on these observations we conclude that offensiveness classifiers and moderation classifiers have different types of errors and we believe that there is scope for improvement by using models that: 1) can incorporate knowledge from both OLI and moderation tasks and 2) are robust to label noise.

Subreddit Group	Model	AUC	F1
r/naruto	M-BERT	51.59	47.43
	XLM-Ro	64.48	37.84
r/naruto + r/anime	M-BERT	55.78	54.13
	XLM-Ro	31.08	46.39
r/judaism	M-BERT	62.08	60.65
	XLM-Ro	50.23	33.77
r/judaism + r/islam	M-BERT	68.27	63.89
	XLM-Ro	62.79	33.91
r/judaism + r/islam + r/christianity	M-BERT	67.89	62.33
	XLM-Ro	49.27	34.50
r/feminism	M-BERT	71.82	64.05
	XLM-Ro	51.27	38.22
r/feminism + r/lgbt	M-BERT	73.25	67.26
	XLM-Ro	49.67	34.10
r/feminism + r/lgbt + r/racism	M-BERT	72.95	66.16
	XLM-Ro	60.78	33.80

Table 7: Performance of classifiers within groups of subreddits. M-BERT refers to bert-base-multilingual-cased and XLM-Ro refers to XLM-RoBERTa-base.

4.4 Manual Analysis of Rule Violations

We also manually labeled 145 removed comments from the r/stock subreddit with the help of 3 annotators. The annotators were instructed[10] to select one of the five rules a removed comment violated. We decided to drop class 4 due to a small number of samples. We also dropped class 5 as this class corresponded to the "not sure" class. We selected examples where annotators had the most agreement and this resulted in 111 comments across 3 classes with 36, 54, and 21 comments for the non-civil, missing context, and spam/self-promotion categories (respectively). We performed five-fold cross-validation strategy for our experiments by training a logistic regression classifier on textual features extracted using sentence-transformers (all-mpnet-base-v2) (Reimers and Gurevych, 2019) We obtained an AUC of 91.8 ± 1.8, and the F1 scores for individual classes are 80.56 (Non-civil),

[9] Average offensiveness score from classifiers in Section 4.1

[10] The guide provided to the annotators is shown in Table 14

77.47 (Missing context), and 61.53 (Spam/self-promotion). The class-wise numbers match our intuition that examples from the "non-civil" class performs the best as it mostly includes offensive comments. This experiment shows that it is possible to learn to predict which rule was violated and thus provide some explanation to the user.

4.5 Experiments with Partitioning Data

As observed earlier, the AUC and F1-macro scores for moderation can be relatively low due to the task's complexity. We hypothesize that grouping similar subreddits together may improve the results within the groups. To verify this, we manually selected three groups of subreddits with similar topics and sample both training and test sets based on the grouping. From Table 7, we observe that grouping subreddits together may result in improved performance. For example, grouping `r/lgbt` and `r/feminism` together achieves 73.25 AUC, an improvement of 1.43 over `r/feminism` alone, and an improvement of $6 \sim 7$ compared to Table 12. However, we also observe that grouping more than two subreddits can lead to a drop in performance. For instance, adding `r/racism` to the group of `r/lgbt` and `r/feminism` decreases the AUC slightly to 72.95. This suggest that splitting the moderation task across multiple classifiers could lead to better performance with proper partitioning.

4.6 Insights on Moderation Dataset and Task

Based on the qualitative and quantitative evaluations, we summarize our insights:

Moderation is a highly imbalanced task. Unlike existing offensive language datasets, which may have $\geq 30\%$ positive examples (Mandl et al., 2021; Zampieri et al., 2019), the proportion of removed comments in our moderation dataset is $\leq 2\%$ This is challenging for most classifiers as their prediction will be biased by the majority class (see Table 11). We believe that this challenge makes our dataset distinct from others because it exposes models to real world scenarios.

Moderation labels are noisy. The moderation labels in our dataset are noisy as we found that some highly offensive comments were not removed and some benign comments were removed by the moderators. There could be multiple reasons for this such as human errors, individual biases. We believe this label noise is a characteristic of the task and researchers will be required to design algorithms with this in mind.

Moderation requires more than offensiveness detection. Our experiments reveal that pre-existing datasets with hate speech or offensive language labels are not sufficient for moderation task. This task involves referencing a set of guidelines which include not only being civil, but also other community-specific rules such as no off-topic discussions, no self-promotions, and no low-effort comments. Models trained only on offensive language datasets are able to identify non-civil content but fail to detect other cases of rule violation.

Moderation systems need to be adaptive. As Reddit communities are self-organized and moderators can make their own rules, many subreddits have very specific rules that may not be common with other subreddits. For example, "keep it halo" (`r/halo`), no "Asking for handouts or transactions" (`r/personalfinance`), "Research must be less than 6 months old" (`r/science`), and "No medical advice" (`r/bodyweightfitness`). A classifier trained on a large dataset like ours could possibly capture general rules and macro-norms, but is unlikely to transfer to novel communities with very specific rules. We thus argue that an adaptive and dynamic system that could condition its decisions on given guidelines is needed for successful auto moderation on platforms such as Reddit.

5 Conclusions

In this study we discussed the challenges in content moderation, which include but are not limited to: differences in community rules, subtlety of offensiveness in different contexts or languages, and moderator biases. Due to these challenges, prior works on offensiveness or hate speech identification are not adequate for solving moderation. Thus, we propose a multilingual dataset consisting of Reddit comments as well as subreddit meta-data including descriptions and rules. We experimented with baseline transformer based models to verify our assumptions and show that, although moderation is a challenging task, there are also many opportunities for further studies: linking removed comments to violated rules, moderating comments in context, understanding the differences between languages and culture in moderation process, etc. We believe that our work will foster more research in the area of automatic content moderation.

Limitations

Moderation within context. Our experiments using standard supervised learning methods take each comment separately and make predictions independently, while in practice context should also be taken into consideration. When collecting each comment we have recorded its parent post id, which could be used to recover the tree structure of a discussion thread, providing context for a comment.
Incorporating community rules. Our baseline model is a universal moderation classifier that did not incorporate community rules. A more adaptive model would be able to make predictions conditioned on rules that are dynamically changing among different communities. We provide metadata of each subreddit with our dataset that includes a description of the subreddit and subreddit rules so that future research can incorporate this information to build better approaches.

Ethics Statement

We have discussed in detail the data selection process while working with different subreddits. We had approval from our Institutional Review Board (IRB) for collecting social media data. We only collected publicly available data (comments, submissions, and subreddit metadata) while adhering to Reddit's policy and did not collect any user-related Personal Identifiable Information (PII) (e.g., Date of Birth) intentionally. To address the possible appearance of PII in public comments, we went through a data cleaning process using scrubadub[11] to remove word tokens or phrases that could be a person's name, email, SSN, driver's license number, etc., to further reduce the risk of PII leakage. We have kept the data and codebase on secure servers that are accessible only by persons involved in this study. Our dataset will be available to researchers who agree to the terms and conditions of our data-sharing policy approved by the ethical committee of our institution.

Acknowledgment

This project was supported by DARPA grant prime OTA No. HR00112290024 (subcontract No. AWD00005100 and SRA00002145). We also acknowledge the Center for Research Computing at the University of Pittsburgh for providing computational resources.

[11] https://scrubadub.readthedocs.io/

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. *ArXiv*, abs/2209.08207.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. *arXiv preprint arXiv:2003.07428*.

Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–25.

Mudit Chaudhary, Chandni Saxena, and Helen Meng. 2021. Countering online hate speech: An nlp perspective. *arXiv preprint arXiv:2109.02941*.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, pages 693–696. Springer.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5, pages 11–17.

Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Casey Fiesler, Joshua McCann, Kyle Frye, Jed R Brubaker, et al. 2018. Reddit rules! characterizing an ecosystem of governance. In *Twelfth International AAAI Conference on Web and Social Media*.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

R Stuart Geiger and David Ribes. 2010. The work of sustaining order in wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 117–126.

Tarleton Gillespie. 2020. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234.

Jennifer Golbeck, Zahra Ashktorab, Rashad O Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A Geller, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, et al. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on web science conference*, pages 229–233.

Sabit Hassan, Katherine J Atwell, and Malihe Alikhani. 2022. Studying the effect of moderator biases on the diversity of online discussions: A computational cross-linguistic study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Shagun Jhaver, Pranil Vora, and Amy Bruckman. 2017. Designing for civil conversations: Lessons learned from changemyview. Technical report, Georgia Institute of Technology.

George Kennedy, Andrew McCollough, Edward Dixon, Alexei Bastidas, John Ryan, Chris Loo, and Saurav Sahay. 2017. Technology solutions to combat online harassment. In *Proceedings of the first workshop on abusive language online*, pages 73–77.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2021. Overview of the HASOC track at FIRE 2020: Hate speech and offensive content identification in indo-european languages. *CoRR*, abs/2108.05927.

Hamdy Mubarak, Sabit Hassan, and Shammur A. Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *ArXiv*, abs/2201.06723.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *The 12th Language Resources and Evaluation Conference*, pages 6237–6246. European Language Resources Association.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Multilingual offensive language identification for low-resource languages. *Transactions on Asian and Low-Resource Language Information Processing*, 21(1):1–13.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Sarah T Roberts. 2014. *Behind the screen: The hidden digital labor of commercial content moderation*. University of Illinois at Urbana-Champaign.

Mattia Samory. 2021. On positive moderation decisions. In *ICWSM*, pages 585–596.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Hannah M Wang, Beril Bulat, Stephen Fujimoto, and Seth Frey. 2022. Governing for free: Rule process effects on reddit moderator motivations. In *International Conference on Human-Computer Interaction*, pages 97–105. Springer.

Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–13.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

A Appendix

A.1 Human Annotators

For better understanding of the moderation data we have performed two studies that involved human annotators: 1) Manual analysis of offensiveness in Section 3.4, and 2) manual analysis of rule violations in Section 4.4. In these two studies have used volunteered annotators including employees within our institution and university graduate students, who are all based in North America. We did not use external annotation platforms and services since these are small-scale studies ($\leq$ 200 data points) designed for better understanding of the moderation data and the annotations will not be released with our large main dataset.

A.2 Temporal Changes in Community Rules

We scraped the community rules for each subreddit at temporal intervals of 1-2 week(s) and changes in rules have been captured in our data. We checked our entire collection of scraped rules across 7.5 months and found that rule edits happened at a rate less than 0.93% among all rules from all subreddits per week. Many of those edits are simple changes (e.g., adding/removing periods, capitalizing words) that does not alter their semantic meaning and thus we consider those as not affecting moderation decisions. There are a few cases when the moderators added new rules which would affect moderation decisions of certain types of content. For example, on `r/antiwork`, the rule of "No politicians, no CEOs" was changed to "No politicians, no employers, no landlords, no cops." in between June 27 and July 04. For those cases, developing content moderation models with changing rules would be a good extension of this work.

A.3 Completeness of the checking later procedure in data collection

In our preliminary study we found that comment removal happens most frequently during the first 3 days after being posted. Thus, we set the interval to 1 week to achieve a reasonable trade-off between completeness and efficiency. We did an additional study by re-scraping the status of an extra set of $452,000$ comments collected during September. Comparing their old labels (checked 7 days after posting) with new moderation labels (scraped this week, which is more than 2 months after posting), 92.89% removed comments has already been covered. Based on our observation of typical subreddits, we also chose the time window of 7 days to limit domain shifts in the data due to change in topics, moderators or focus of the subreddit.

Language	Subreddits
English	r/anime, r/antivaxxers, r/antiwork, r/birdswitharms, r/bitcoin, r/boardgames, r/bodyweightfitness, r/breadstapledtotrees, r/christianity, r/collapse, r/coronavirus, r/covid19, r/feminism, r/fifthworldproblems, r/funny, r/futurology, r/gadgets, r/games, r/grandpajoehate, r/halo, r/havewemet, r/immigration, r/islam, r/judaism, r/lego, r/lgbt, r/lifehacks, r/mildyinfuriating, r/minecraft, r/naruto, r/news, r/nfl, r/onetruegod, r/personalfinance, r/publicfreakout, r/racism, r/science, r/scifi, r/showerorange, r/showerthoughts, r/space, r/stocks, r/streetwear, r/talesfromcavesupport, r/theoryofreddit, r/tooafraidtoask, r/worldnews, r/xboxseriesx
Spanish	r/argentina, r/argaming
German	r/de, r/finanzen, r/ich_iel
French	r/rance, r/quebec, r/moi_dlvv

Table 8: The list of all subreddits in our dataset.

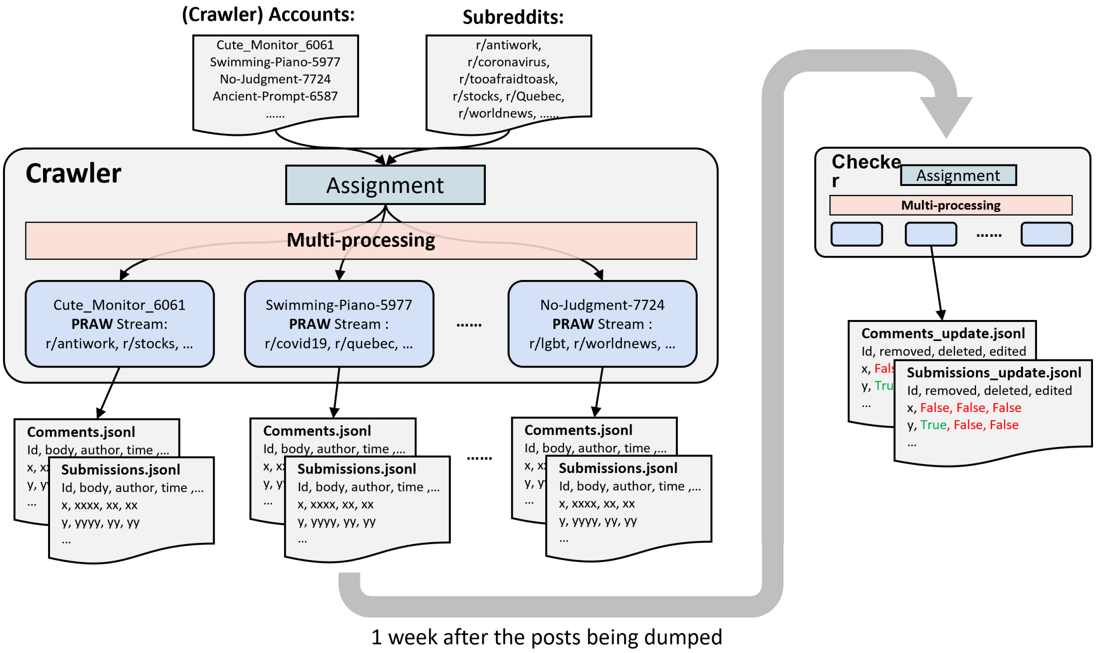


Figure 5: Data collection pipeline: 1) the crawler will assign target subreddits to a number of accounts and streaming comments and submissions concurrently and 2) after a certain period of time (around a week), the checker will go through each record and update if the comment still exists or has been removed/deleted.

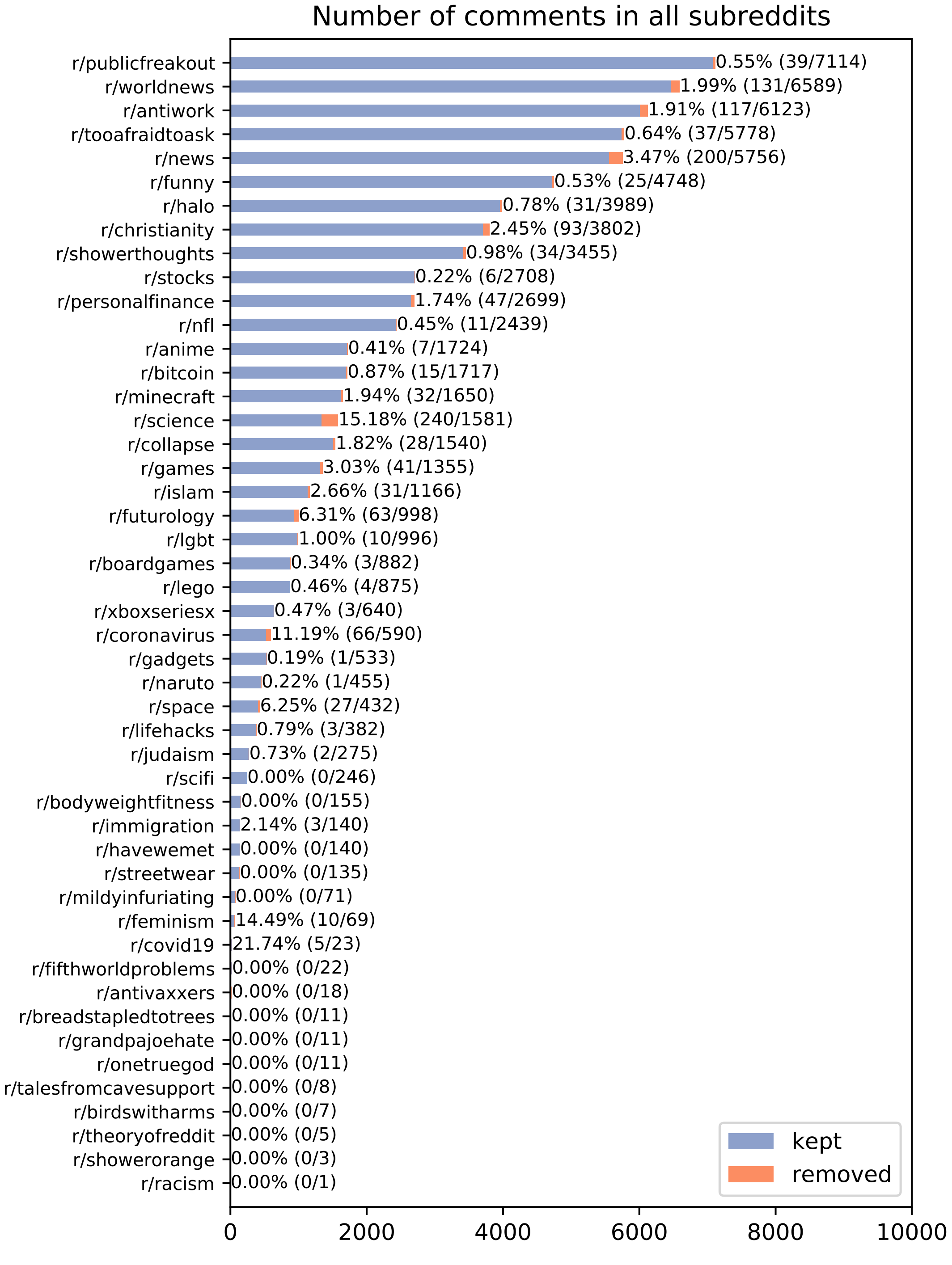


Figure 6: Number of kept and removed comments in the English validation (en-val) data.

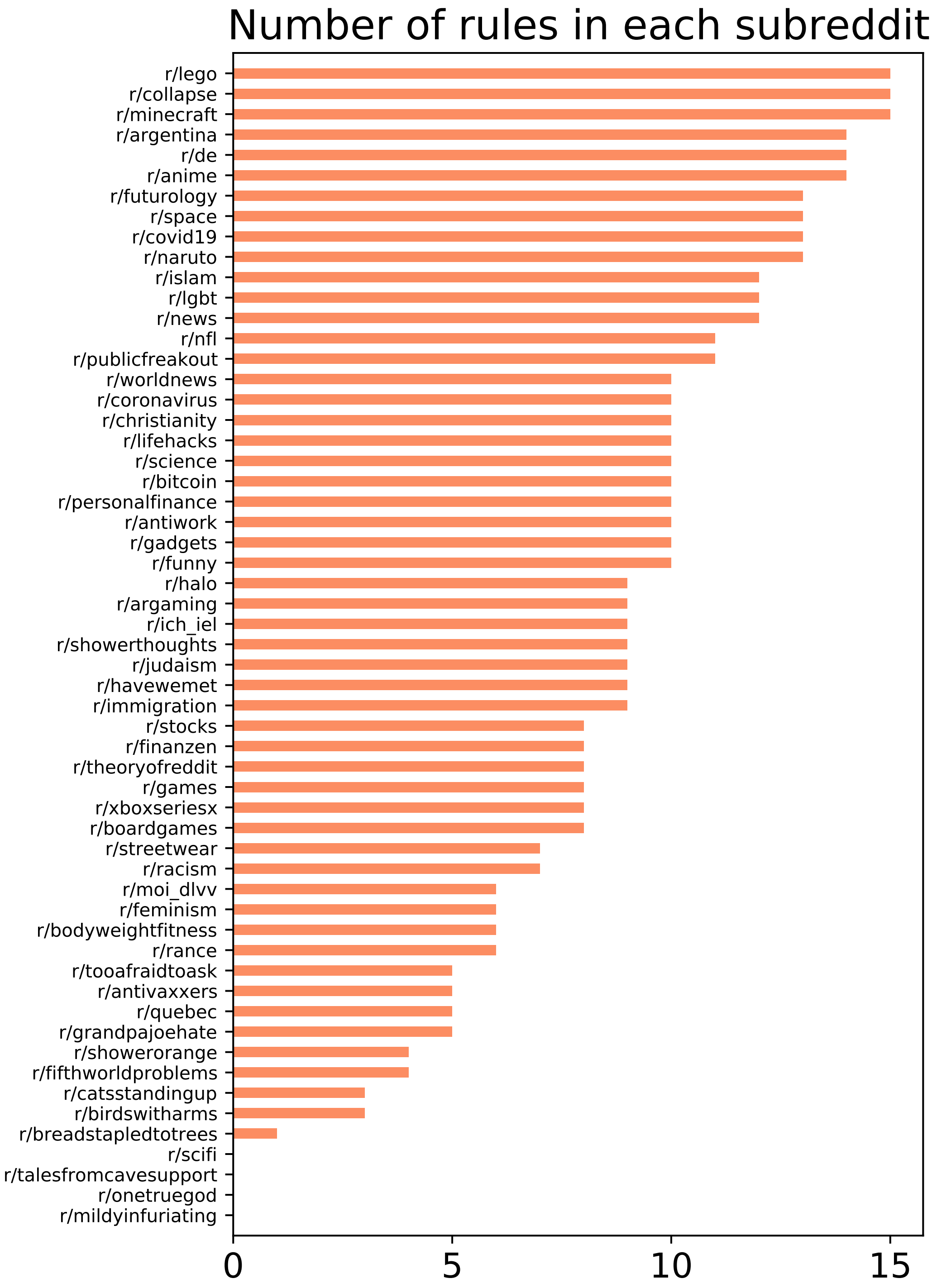


Figure 7: Number of rules in all subreddits.

	Comment	**Subreddit**	f_o
a	try contacting drewcybersupport on instagram about this he will definitely help you out	r/minecraft	0.01
b	i think there may still be hope. my company helps clients like yourself get rid of pmi. pm me and we can help you.	r/personalfinance	0.02
c	i wanna hear ur view on {{organization}} products and $sprk tokens. are ya joining anyone?	r/bitcoin	0.02
d	thanks for the good constructive conversation i really did enjoy it.	r/feminism	0.02
e	maybe it would be easier if you set up a meeting irl.	r/science	0.03
f	if anyone needs images they are all over /r/transformers right now	r/lego	0.03
g	so you have two masters degrees and youre looking to {{organization}} for help. those degrees are really paying off.	r/personalfinance	0.03
h	no not me it would be another little tiny b***h like you	r/publicfreakout	0.91
i	... your piece of s**t car so slow you'll be left behind ... d**b silly beta.	r/tooafraidtoask	0.86
j	we've come so god d**n far in this country ... f**k this s**t!! f**k this country!	r/antiwork	0.82

Table 9: Sampled removed comments and their average offensiveness score from 3 classifiers trained on offensive language identification and hate speech detection datasets.

split	**language**	**#sub**	**#comments**	**removal**
		Training data		
en-train	English	48	$1,016,386$	47.96%
en-val	English	48	$56,460$	47.96%
		Test data		
de	German	3	$57,952$	50.00%
es	Spanish	2	$18,298$	50.00%
en	English	48	$56,466$	48.09%
fr	French	3	$5,122$	44.57%

Table 10: Number of subreddits, comments, and percentage of removed comments in the additional balanced data splits.

Setting	Lang.	AUC				F1-macro			
		en	de	fr	es	en	de	fr	es
MLLM	multi	66.83$_{\pm 0.10}$	69.42$_{\pm 0.14}$	64.25$_{\pm 0.14}$	63.90$_{\pm 0.03}$	49.60$_{\pm 0.10}$	49.78$_{\pm 0.00}$	49.94$_{\pm 0.00}$	49.83$_{\pm 0.00}$
Trans-tr	de	60.33$_{\pm 0.44}$	63.34$_{\pm 0.68}$	-	-	49.53$_{\pm 0.00}$	49.78$_{\pm 0.00}$	-	-
	fr	53.47$_{\pm 0.38}$	-	49.96$_{\pm 0.61}$	-	49.53$_{\pm 0.00}$	-	49.94$_{\pm 0.00}$	-
	es	64.16$_{\pm 0.53}$	-	-	63.29$_{\pm 0.43}$	49.58$_{\pm 0.03}$	-	-	49.99$_{\pm 0.13}$
Trans-te	en	67.38$_{\pm 0.18}$	71.23$_{\pm 0.02}$	71.61$_{\pm 0.34}$	64.33$_{\pm 0.09}$	49.66$_{\pm 0.05}$	50.01$_{\pm 0.17}$	50.36$_{\pm 0.30}$	50.22$_{\pm 0.30}$

Table 11: Experimental results with standard deviation on original splits.

Setting	Lang.	AUC				F1-macro			
		en	de	fr	es	en	de	fr	es
MLLM	multi	66.14$_{\pm 0.02}$	70.10$_{\pm 0.07}$	68.32$_{\pm 0.01}$	59.81$_{\pm 0.06}$	59.97$_{\pm 0.51}$	64.46$_{\pm 0.11}$	61.15$_{\pm 0.93}$	56.78$_{\pm 0.06}$
Trans-tr	de	61.10$_{\pm 0.67}$	65.90$_{\pm 1.22}$	-	-	57.24$_{\pm 1.15}$	61.15$_{\pm 1.17}$	-	-
	fr	53.64$_{\pm 0.37}$	-	53.71$_{\pm 1.40}$	-	51.99$_{\pm 0.98}$	-	55.10$_{\pm 0.56}$	-
	es	63.08$_{\pm 0.14}$	-	-	60.52$_{\pm 0.51}$	59.33$_{\pm 0.26}$	-	-	56.87$_{\pm 0.47}$
Trans-te	en	66.94$_{\pm 0.05}$	70.66$_{\pm 0.11}$	71.57$_{\pm 0.1}$	61.96$_{\pm 0.03}$	61.96$_{\pm 0.1}$	64.16$_{\pm 0.40}$	64.87$_{\pm 1.30}$	58.11$_{\pm 0.27}$

Table 12: Experimental results with standard deviation on balanced splits.

Category	**number of examples**
Non-civil	36
Missing context	54
Spam/self-promotion	21

Table 13: Manually annotated rule violations.

Warning: Due to the nature of this study, this data may contain disturbing content, such as offensive language and hate speech.

Label	**Rule**	**Examples**
1	not being civil (hate, offensive, insulting, harassment, ...).	(1) you're complete moron. (2) you fuckin loser. (3) have fun staying poor buddy.
2	missing context or effort (Context and effort must be provided; empty posts or empty posts with links will be automatically removed).	(1) https://www.youtube.com/watch?v=... (2) let's GOOOOO. (3) GME.
3	spam or self-promotion (Spam, ads, solicitations (including referral links), and self-promotion posts or comments will be removed).	(1) Checkout my youtube channel. (2) https://www.youbube.com/xxxxxxx. (3) use this link to get $10 bonus when you register an account at xxxxx.
4	Cryptocoin discussions unrelated to stocks ("I bought bitcoins at coinbase" doesn't count, but "Coinbase sells X amount of bitcoins which is X amount of profit for the company" is fine), penny stocks (including OTC, microcaps, pump & dumps, low vol pumps and SPACs).	(1) SHIB coin is on fire. (2) $ATER! get on the train now we are going to the moon! (3) BRQS is the play now, rocketing to $2 by the end of this week.
5	Not sure why this comment was removed (a comment doesn't seem to have violated any of the rules above, it's hard to say with out context why it was removed)	None

Table 14: The guide provided to annotators.