Sensitive Content Classification in Social Media: A Holistic Resource and Evaluation

Dimosthenis Antypas[1*] Indira Sen[2*†] Carla Perez-Almendros[1]
Jose Camacho Collados[1] Francesco Barbieri[†]
[1]Cardiff NLP, School of Computer Science and Informatics, Cardiff University, UK
[2]University of Mannheim, Germany
[1]{AntypasD,PerezAlmendrosC,CamachoColladosJ}@cardiff.ac.uk
[2]indira.sen@uni-mannheim.de

Abstract

The detection of sensitive content in large datasets is crucial for ensuring that shared and analysed data is free from harmful material. However, current moderation tools, such as external APIs, suffer from limitations in customisation, accuracy across diverse sensitive categories, and privacy concerns. Additionally, existing datasets and open-source models focus predominantly on toxic language, leaving gaps in detecting other sensitive categories such as substance abuse or self-harm. In this paper, we put forward a unified dataset tailored for social media content moderation across six sensitive categories: conflictual language, profanity, sexually explicit material, drug-related content, self-harm, and spam. By collecting and annotating data with consistent retrieval strategies and guidelines, we address the shortcomings of previous focalised research. Our analysis demonstrates that fine-tuning large language models (LLMs) on this novel dataset yields significant improvements in detection performance compared to open off-the-shelf models such as LLaMA, and even proprietary OpenAI models, which underperform by 10-15% overall. This limitation is even more pronounced on popular moderation APIs, which cannot be easily tailored to specific sensitive content categories, among others.

Disclaimer

Due to the nature of the subject studied in this work, **this paper contains sensitive and potentially offensive language. Reader discretion is advised.**

1 Introduction

Consider the case of a researcher or a data analyst who needs to filter sensitive content from a large dataset. Such task is crucial to ensure that data shared or analysed does not include harmful or inappropriate material. One might initially consider using external tools like Perspective[1] or OpenAI moderation APIs[2] to assess and filter sensitive content. However, this approach often falls short, presenting important limitations for an effective identification of inappropriate content online (Udupa et al., 2023). For instance, they usually offer limited customisation capabilities (e.g., how can the model be improved if it fails on specific domains or keywords?), and limited sensitive categories coverage (lacking in detecting self-harm (Uban and Rosso, 2020), for example). Finally, these tools rely on external servers, which raises concerns about data privacy and security (Oseni et al., 2021; Gupta et al., 2023).

Alternatively, one might consider using existing datasets and open-source models for sensitive content detection. This could be a viable option if the primary focus was on detecting toxic language, given the abundance of resources available in this area. However, if the goal extends to identifying additional sensitive categories such as sexually explicit content, drugs, self-harm or spam, the situation becomes more challenging. Data on these less-explored categories is limited and sometimes outdated. For instance, those categories could be covered on datasets that are biased (Wiegand et al., 2019), old or inaccessible even in an anonymized manner (Tadesse et al., 2019; Sawhney et al., 2018), too small-scaled or rely on an handful of keywords to extract the data (Ding et al., 2016). This limited approach can result in incomplete or less accurate detection of sensitive content.

Existing solutions either require sending data to external servers or fail to address the full spectrum of sensitive content categories. In response to these challenges, this paper proposes a new holistic ap-

*Equal contribution.
† Work done while at Snap Inc.

[1]https://perspectiveapi.com/
[2]https://platform.openai.com/docs/guides/moderation

Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH), pages 17–31
August 1, 2025 ©2025 Association for Computational Linguistics

proach: **a unified dataset for detecting sensitive text across a broad range of categories**, including (1) conflictual language, (2) profanity, (3) sexually explicit material, (4) drug-related content, (5) self-harm, and (6) spam[3]. This dataset can then be used for both evaluation and fine-tuned models to all these categories under a single framework.

Our approach involves collecting and re-annotating data to ensure consistent quality across sensitive classes. The alternative of putting together a collection of existing datasets to create one single dataset would include several limitations, as (1) there would be bias towards the data distribution (different retrieval strategies, topics, source platforms), (2) annotation guidelines and quality would differ, and (3) each text would include only one sensitive dimension (even if the text includes multiple sensitive categories).

In short, we propose a holistic approach when it comes to sensitive content moderation in social media, overcoming common shortcomings of previous works and providing the following contributions:

- New dataset: We introduce the X-Sensitive dataset, manually annotated and tailored for social media content, featuring multiple categories and designed to be resilient against keyword and domain shifts.
- Sensitive category analysis: We study the interplay between sensitive categories and how these categories vary across different annotator demographics.
- Model evaluation: The best results are achieved from large language models (LLMs) with 8 billion parameters, fine-tuned on our dataset. However, smaller language models[4] (355 million parameters) show only about 2% less accuracy.
- Comparison with off-the-shelf LLMs: We find that readily available LLMs, such as gpt-4o, under-perform by 10-15% compared to fine-tuned models, highlighting the value of bespoke training on specialised datasets.

The X-Sensitive dataset, as well as the best performing models built upon it, are made openly available. X-Sensitive is available at https://huggingface.co/datasets/cardiffnlp/x_sensitive. Best multi-label and binary models are available at https://huggingface.co/cardiffnlp/twitter-roberta-large-sensitive-multilabel and https://huggingface.co/cardiffnlp/twitter-roberta-large-sensitive-binary, respectively.

[3]We use social media platform user guidelines to be sure to have reliable guidelines.

[4]Pre-trained on social media language.

2 Related Work

Our current work aims to bridge the gap between current academic research in content moderation and the needs of content moderators in realistic scenarios. While hate speech and toxic language are widely studied in NLP, there is little research on other types of sensitive content that platforms seek to detect and moderate, such as sexually explicit content or content about illicit substances (Arora et al., 2023). To that end, our work is situated at the intersection of NLP research on harmful language detection and research on platform governance and content moderation.

2.1 Automatic Detection of Harmful Language

Hate speech Detection. Automatic detection of hate speech, and related social constructs like offensive and toxic language, is an active area of research in NLP (Fortuna and Nunes, 2018; Poletto et al., 2021). However, there are several challenges, not least the lack of high quality datasets for studying such phenomena (Vidgen and Derczynski, 2020).

Self-harm and Suicidal Content Detection. Chancellor et al. (2016b) identify communities with self-harm related content, while Tejaswini et al. (2024) also look into related behaviors such as depression. Previous research has also looked into suicidal content detection (Coppersmith et al., 2018) and general self-harm (Un Nisa and Muhammad, 2021). There are generally several ethical challenges associated with studying mental health conditions, including self-harm and suicidal ideation (Chancellor et al., 2019).

Illicit Substance Abuse. Past research has looked into automated approaches for discussions of illegal or banned substances, including drugs (Buntain and Golbeck, 2015; Lavanya and Sasikala, 2022; Simpson et al., 2018).

Sexually Explicit Content. Research has also focused on developing automated systems to detect sexually explicit content (Barrientos et al., 2020), address sexual harassment (Chowdhury et al., 2019), and identify sexualised cyberbullying (Basu et al., 2021).

Spam Detection. Automatic Spam detection is widely studied in NLP as well as computer security communities. Typical automation techniques rely on expert-annotated training data used to train machine learning models (Hussain et al., 2019). However, like the other categories spam detection has rarely been studied in the context of other types of problematic content, with Founta et al. (2018) being an exception.

2.2 Content Moderation and Platform Governance

Platforms on the internet, such as web and social media sites, often employ mechanisms to curate their content and reduce problematic or harmful content through content moderation (CM). CM can take many forms, from commercial content moderation outsources to underpaid moderators in the Global South (Roberts, 2019) to artisanal solutions, some of which are led by volunteers (Caplan, 2018). Yet as content grows, platforms turn towards automated methods, often Artificial Intelligence (AI) based techniques either solve or ameliorate their moderation problem (Gorwa et al., 2020).

However, the question remains on how much of this detection is automatable? (Gillespie, 2020). There are not only several technological limitations (e.g. the dearth of AI methods for non-English content (Vidgen and Derczynski, 2020)) but also political challenges (e.g., who gets to decide what is harmful? (Fleisig et al., 2024)) and challenges at the nexus of technology and politics (e.g., how do we aggregate the potentially divergent judgements of whether something is harmful? (Fan and Zhang, 2020; Gordon et al., 2022)). On the other hand, platform studies researchers have studied which types of technological solutions, including AI-based tools, would facilitate the work of content moderators while also establishing some of the tensions of the whole practice of content moderation. However, it is unclear if those proposing technological solutions for CM are basing their solutions on the requirements of content moderators.

Categories of Sensitive Content. Several researchers have attempted to categorise what counts as 'sensitive' content on web and social media platforms, i.e., content that requires moderation (Jiang et al., 2020; Scheuerman et al., 2021).

We address one of the many challenges of automatic content moderation — lack of benchmark datasets for measuring understudied categories of problematic content like discussion related to self-harm and illicit substances, particularly drugs. We also provide a holistic benchmark of both these aforementioned understudied categories as well as widely studied categories like profanity, allowing researchers to model the associations between different types of sensitive content.

3 X-Sensitive Dataset

In order to study sensitive content in X, we construct a new dataset, X-Sensitive. As a first step, we conceptualise a topic taxonomy based on community guidelines from several social media platforms.

3.1 Taxonomy

We use the community guidelines of various social media platforms to ground our taxonomy (Scheuerman et al., 2021). Using iterative coding, we refine, merge, and fix 5 broad categories and 7 specific subcategories of sensitive content which are mapped to rules in community guidelines. Our final categories and their definitions are:

Drugs. Content that encourages, promotes or glorifies the use of regulated drugs. Also applicable to content that mentions sales, purchases, or the act of obtaining or trying to obtain regulated drugs.

Sexually Explicit Content (Sex). Pornographic or other types of sexual content. We collect and download 50+ textual abusive language datasets from hatespeechdata.com. We then use the Perspective API to label these datasets with the 'sexually explicit' endpoint and then use the labelled data for fine-tuning the XLM-T sexually explicit content classifier.

Hate speech. Attacks against protected attributes like race, colour, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight or pregnancy status.

Other conflictual language. Attacks based on other categories or without any mention of the categories mentioned.

Profanity. Language containing slurs and profanity even if they are not directed towards a specific entity.

Self-harm. Posts depicting, promoting or glorifying violence or harm against oneself, such as eating disorders or suicide.

Spam. Irrelevant content that is unsolicited; or content that aims to drive traffic or attention from a conversation on the platform to entities outside the platform.

3.2 Message Collection

Typically previous work on sensitive content detection, particularly hate speech detection, uses a small set of keywords to collect data, which may lead to limited coverage of the resultant datasets (Ousidhoum et al., 2021). To tackle this problem, we utilise a keyword expansion technique combining word embeddings (Mikolov et al., 2013), trained on tweets (Pennington et al., 2014), for keyword expansion and clustering for controlling the expanded sets. The specific algorithm is described as follows:

Algorithm 1 Keyword List Expansion Technique using Word Embeddings

```
Input: Seed list {w_1, w_2, ...}
Output: Expanded keyword list
Start with a seed list {w_1, w_2, ...}
Cluster keywords' vectors into k clusters
Check and select the relevant clusters
Compute the dot product v_i · μ_c for each word
  vector v_i and each cluster mean vector μ_c
Find n_1 words that are closest to the mean of
  each cluster
for each new word do
    Find the closest n_2 words
end for
```

Due to the variety of the categories in our sensitive content category, we look at multiple sources for seed lists (Appendix A, Table 3). The conceptual similarities between profanity, sexually explicit content, and conflictual content as well as the existence of lists that collect keywords related to these three types of discourse, make us opt for a unified seed word list covering these three categories.

For self-harm and spam, uni-grams are not as informative as they are for other categories. Therefore for the former, we look at past research on eating disorders and suicidal ideation (Chancellor et al., 2016b, 2021) and obtain phrases (n-grams with a high TF-IDF score) from the Reddit data used to train the self-harm phase 1 classifier. For spam, we use the dataset from Founta et al. (2018), particularly the tweets that were labelled containing spam and obtain n-grams from there using a similar method. We manually assess each of the keywords for all categories and remove low precision words like 'snow' for drugs. While snow may refer to cocaine in some contexts, most tweets containing it do not use it in that sense. After this manual inspection, we apply our cluster-based keyword expansion technique. We again manually assess the keywords and include only those that are relevant to the category. The final statistics of our keywords are listed in Appendix A, Table 4.

3.3 Annotation

Each entry of the dataset was annotated by at least three coders, where each coder had to answer with *yes*, *no*, or *not sure* if the tweet contained any of the sensitive classes. Specifically, for the case of conflictual language the annotators were asked to select whether the tweet contained hate speech or any other form of conflictual language. This approach aimed for a more fine grained classification of conflictual language. However, due to low agreement between coders we opted to merge the categories "Hate Speech" and "Other Conflictual Language" into a single class *Conflictual Language*.

A label was assigned to a tweet if at least one annotator answered *yes* and there was no direct opposition from the rest of the coders (i.e. the rest of the coders answered *yes* or *not sure*). We refrained from utilising a majority rule in order to create a more realistic and challenging dataset while also weighting the recall of potentially sensitive content higher.

The coders who worked on this task were selected and filtered through the Prolific.co platform based on their fluency in English. The annotation was performed through an interface created with qualtricsXM[5]. The coders were also provided with 15 examples of already annotated tweets to help them better understand the task. Finally, we utilised several filters to ensure a high quality of annotations. First, we included a set of test questions randomly inserted in the task which were used to filter out low quality coders. Additionally, coders that finished the task too quickly or provided low

[5]The annotation guidelines can be found in Figure 3, Appendix B.

quality answers (for example always selecting the same answer) were excluded.

Overall 523 coders from various demographic backgrounds took part in the annotation process. We assessed the quality of the annotation by utilising Krippendorff's Alpha (*Alpha*) (Krippendorff, 2011). The annotators achieve 0.49 *Alpha* when considering all available classes and 0.56 *Alpha* when considering only the presence of sensitive content or not. The scores are in line or better with previous similar studies on toxic and sensitive content (Muralikumar et al., 2023; Lima et al., 2024).

It is interesting to note that when looking at subgroups of annotators based on their demographics we observe higher agreement between specific groups, mainly younger (0.51 *Alpha* in multi-label setting for people 39 old and younger) and non-binary people (0.82 *Alpha* in the binary setting). More detailed results can be found in Appendix B, Tables 6 and 5).

Looking in more detail on how different demographics annotate examples a trend is noticed where younger coders and non-binary annotators tend to be more sensitive to the content and are more likely to flag a tweet as sensitive (Appendix B, Tables 7 and 8).

The discrepancies in agreement between groups indicate the inherent difficulty of the task while also providing evidence of a greater coverage of sensitive content within X-Sensitive.

3.4 Statistics

X-Sensitive contains a total of 8,000 tweets all related to sensitive content with 49% of them labelled as one or more of the six sensitive classes available making it a challenging dataset. On average tweets flagged as sensitive are assigned 1.4 labels with maximum assigned labels to a single tweet being 4.

Our dataset displays a skewed distribution of classes as seen if Table 1 with *profanity* being the most populated class present (30.4%). This uneven distribution represents a realistic representation of sensitive content in social media as seen in previous similar studies (Beknazar-Yuzbashev et al., 2022) where it estimated a 5% - 7% of content displayed is inappropriate, making X-Sensitive ideal for usage in real world applications.

At the same time *profanity* being the most frequent class is also expected. Due to the multi-label nature of the dataset, we expect high overlap between *profanity* and other classes as seen in Figure

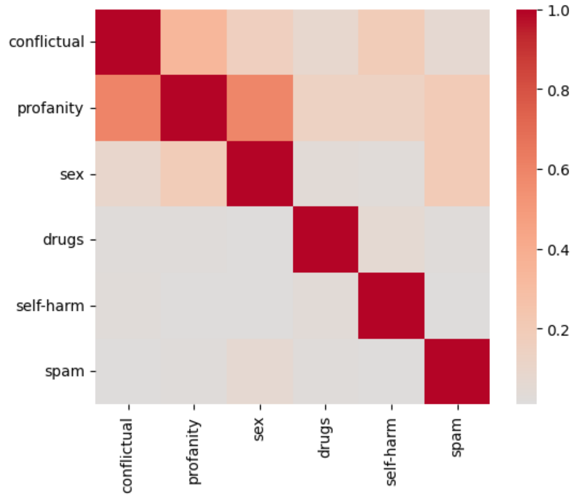

Figure 1: Overlap of classes.

1. Particularly there is a high overlap between *profanity* tweets and those labelled as *sexual explicit content*, and *conflictual*.

In general differences between the classes are revealed even when looking at basic statistics such as the average length of tweets and the presence of emojis in them. As seen in Table 1 tweets labelled as *spam* tend to be longer on average and include a higher number of emojis, characteristics frequently found on spam messages (Robinson and Mago, 2022). Similarly, a higher usage of emojis is observed in tweets flagged as *sexual explicit content*, as specific emojis are often used as representation of sexual acts (Thomson et al., 2018). Furthermore, when examining the top terms of each class based on lexical specificity scores (Camacho-Collados et al., 2016), a clear distinction between the classes is observed, which also serves as a sanity check for the quality of our dataset.

4 Experimental Setting

In this section, we set out the common experimental framework which serve as the basis of the evaluation.

4.1 Data and Settings

To evaluate X-Sensitive and establish baselines of its difficulty we establish two distinct settings: binary and multi-label classification. In the binary setting, tweets will be classified into one of two categories, distinguishing between sensitive and not sensitive content. This approach simplifies the classification process, focusing on the presence or absence of sensitive characteristics in general. In

Category	L	Emo	%	Top Terms
Conflictual	188.67	0.26	17.3	fucking racist nigga white shut
Profanity	173.05	0.43	30.4	fucking shit bitch fuck as
Sex	160.79	0.66	9.7	cock pussy dick horny cum
Drugs	155.62	0.21	3.9	drug weed cbd thc mushroom
Self-harm	166.35	0.44	3.0	suicide suicidal attempt commit ideation
Spam	200.30	1.12	3.4	dm project airdrop solana solanaairdrop
Not Sensitive	176.39	0.32	51.2	physically depressing triggering mental depression
Overall	174.77	0.37		

Table 1: General lexical statistics for each class. The averages of the length of tweet, emojis count are reported. The distribution of each class along with the top five terms based on their lexical specificity are also displayed.

the multi-label setting, tweets can belong to multiple sensitive categories simultaneously, allowing for a more fine-grained analysis that captures the complexity of the content. This dual approach enables a comprehensive evaluation of our dataset's versatility and the classifier's robustness in handling varying degrees of complexity in sensitive content detection.

For both settings we use a split the dataset in train/validation/test sets of 6,000/1,000/2,000 tweets while ensuring that the distribution of classes is similar in each split. To investigate the generalisability capabilities of the models, an additional constraint check is enforced where we ensure that approximately half of the test set, 1,016 tweets, do not share any of the keywords used for collection with tweets from the train set.

4.2 Comparison Systems

For the evaluation, we are interested in comparing three types of approaches: fine-tuning on the same dataset (Section 4.2.1), LLMs with in-context learning either zero- or few-shot (Section 4.2.2), and out-of-the-box content moderation systems (Section 4.2.3). All the systems are clearly not fully comparable, but our dataset can serve as the basis for establishing this basic ground comparison.

4.2.1 Fine-tuning

We evaluate three distinct models tailored for various applications, including general-purpose and those specialized for social media, each differing in size for our fine-tuning experiments. The large version of **RoBERTa** (Liu, 2019) is tested in order to assess the performance of smaller, non specialised, masked language models on our dataset. Timelm, **tlm**, (Loureiro et al., 2022), a RoBERta based model trained on a large X corpus of 154 million tweets is also evaluated to assess the performance of specialised models on social media. The two models are fine-tuned using the implementations provided by Hugging Face (Wolf et al., 2020) and optimising hyper parameters (learning rate, training epochs, warm-up steps) is conducted using Ray Tune (Liaw et al., 2018)[6]. Finally, the 8 billion version of Llama-3, **Llama3-8b**, (AI@Meta, 2024) is also fine-tuned on our dataset by utilising quantisation and PEFT (Liu et al., 2021; Mangrulkar et al., 2022) explore the capabilities of more recent and larger-scale models.

4.2.2 Zero- and Few-shot

In order to assess the zero/few-shot capabilities of large language models in our dataset, we compare four models of different sizes and architectures.

Llama3: The 8 and 70 billion instruct versions of Llama3 are tested. These models are designed to follow user instructions more effectively, allowing us to assess how well they adapt in settings where training data is limited or not available.

chat-gpt-3.5-turbo (chat-gpt): from OpenaAI, [7] an encoder/decoder model with approximately 175 billion parameters (Brown et al., 2020).

gpt-4o: the currently latest model from OpenAI which significantly outperforms its predecessor.

Assessing the performance in zero- and few-shot settings, helps us to exlore the capabilities and limitations of these large language models for sensitive content detection.

4.2.3 Out of the box Systems

The need for detecting sensitive or harmful content has led to several companies to develop their own models, which are made publicly available. In order to highlight the relevance of existing models for this task, we selected three popular specialised systems.

[6] Details of the models used can be found in Appendix C.
[7] https://openai.com/chatgpt

Training	Model	Binary	multi-label	Conflictual	Profaninty	Sex	Drugs	Self-harm	Spam	Not sens.
fine-tuned	RoBERTa	82.4	64.7	60.6	88.9	81.6	52.3	34.3	52.0	83.3
	tlm	84.4	67.7	59.6	88.8	84.3	48.9	50.6	59.1	82.4
	llama3-8b	**85.6**	**69.8**	61.7	**90.6**	**85.8**	53.9	50.6	61.2	**85.1**
Zeroshot	llama3-8b	75.0	52.2	53.5	69.8	70.0	39.2	35.5	21.6	75.6
	llama3-70b	76.5	57.4	54.5	79.4	74.3	55.0	32.8	42.1	63.5
	chat-gpt	60.0	63.2	49.0	60.0	71.0	57.0	41.0	37.0	69.0
	gpt-4o	75.7	64.9	62.2	82.9	84.0	**64.9**	**53.2**	26.2	81.1
Fewshot	llama3-8b	74.9	53.2	43.9	73.3	74.8	49.0	18.5	43.0	70.1
	llama3-70b	79.2	63.0	62.2	82.8	78.9	61.5	32.8	53.9	69.1
	chat-gpt	71.0	64.0	59.0	84.0	83.0	52.0	27.0	48.0	72.0
	gpt-4o	83.3	67.9	63.4	85.7	81.7	61.1	41.9	**64.8**	76.9
Out of the box Systems	llama-g	55.0	-	16.1	-	75.5	-	43.6	-	68.2
	openai-m	72.0	-	63.1	-	73.0	-	46.3	-	75.9
	Perspective	70.0	-	**64.0**	89.0	81.0	-	-	53.0	44.0

Table 2: Macro F1 scores for fine-tuned and zero-/few-shot models are reported in both binary and multi-label settings. Additionally, the F1 scores for each class in the multi-label setting are provided. For out-of-the-box systems, we report the F1 scores in the binary setting and, when available, the F1 scores achieved in each class.

Google's Perspective API (Perspective) (Google, 2023) is a tool developed to detect and score various attributes of text, such as toxicity, and acts as a baseline performance of a production-ready API. In total Perspective provides scores for 16 different categories but in our use case we focus only on 12 of them that fit our taxonomy better[8].

OpenAI's moderation API (openai-m) is an endpoint tailored for content moderation. It classifies content into 18 potentially sensitive categories, 15 of which we map to our own taxonomy.

Meta-Llama-Guard-2-8B (llama-g) (Inan et al., 2023) is a specialised version of LLama-3 that aims to classify content based on a safety risk taxonomy of 11 harm categories (Vidgen et al., 2024)[8]. For our use case we consider only 5 of the categories that correspond better to the taxonomy used in X-Sensitive. Specifically we consider: "Hate" for *Conflictual*, "Suicide & Self-Harm" for *self-harm*; and "Sexual Content", "Sex-Related Crimes", and "Child Sexual Exploitation" for *sexual explicit content*.

4.3 Evaluation Metrics

Given the critical nature of the task and the importance of accurately identifying and recalling all potentially harmful content, we utilise F1 scores to evaluate our models. We assign equal weight to each label and report the macro-F1 score in both binary and multi-label settings. The F1 scores for individual labels are also considered in the multi-label scenario. This approach helps us gain a deeper insight into the challenges posed by the dataset.

5 Results

The scores for both binary and multi-label scenarios, across all models tested in the fine-tuning and zero-/few-shot settings, are presented in Table 2. In general, fine-tuning leads to clear improvement for all models, which reinforces the importance of our dataset not only to evaluate models, but to build specialised models based based on it. *llama3-8b* performs best overall, with macro-f1 scores of 85.6 in the binary setting, and 69.8 in the more challenging and fine-grained multi-label setting.

5.1 Fine-tuned Systems

All the fine-tuned models demonstrate high performance with *RoBERTa* as the least effective, achieving macro-F1 scores of 82.4 in the binary setting and 64.7 in the multi-label setting. The specialised training corpus of *tlm* appears to enhance its performance, as it consistently surpasses *RoBERTa* of the same architecture in both settings. Moreover, the fine-tuned version of the larger and more recent *llama3-8b* model achieves the best overall results in both settings, with macro-F1 scores of 85.6 in the binary setting and 69.8 in the multi-label setting, notably achieved without any hyper-parameter tuning, unlike the other models. Overall, the fine-tuned models tend to struggle the most with the least represented classes, such as *Drugs*, *Self-harm*, and *Spam*. Interestingly, despite comprising 17.3% of the total entries, the models under perform in the *Conflictual* category, while they exhibit better performance in the less prevalent *Sexual Explicit*

[8]Detailed taxonomy can be found in Appendix C

Content class, which accounts for only 9.7% of tweets. This disparity may indicate the models' difficulties in identifying subtler features within the *Conflictual* category.

5.2 Zero/Few-shot

Zero-shot. When tested without any contextual information, the models display varying degrees of effectiveness. Notably, the 70b version of llama3, *llama3-70b*, outperforms its smaller counterpart and competes with OpenAI's models, achieving the highest macro F1 score of 76.5 in the binary setting. In the multi-label scenario, the *gpt-4o* model excels, achieving a macro-F1 score of 64.9. Generally, the zero-shot models do not reach the performance levels of their fine-tuned counterparts, with the notable exception of *gpt-4o*, which surpasses only *RoBERTa* in the multi-label setting.

Few-shot. In the few-shot setting, *llama3-8b* exhibits performance comparable to its zero-shot execution, illustrating the constraints of smaller models. This limitation is further highlighted by the performance gains observed in the rest of the models, *llama3-70b*, *chat-gpt*, and *gpt-4o*, which show average increases of 7.6 and 3.2 points in macro-F1 for the binary and multi-label settings, respectively. This underscores the effectiveness of in-context learning in larger models. Overall, *gpt-4o* achieves the best performance, competing with the fine-tuned models and notably outperforming *llama3-8b* in specific categories, *Conflictual*, *Drugs*, and *Spam*.

5.3 Out of the box

When evaluating the performance of "out-of-the-box" models, we find that they generally fail to achieve high scores (Table 2). Notable exceptions occur in specific categories such as *Profanity* and *Sexual Explicit Content*, where the *Perspective* and *llama-g* systems excel. The best performing out-of-the-box model, *openai-m* API, achieves a macro-F1 score of 72% in the binary setting, demonstrating greater robustness in detecting non-sensitive content (F1: 75.9%) compared to its peers. Despite this, its overall performance remains the lowest among the models tested, except for *chat-gpt* in zero- and few-shot scenarios. It is important to note that these scores may be influenced by the fact that these systems do not utilise the same taxonomy as X-Sensitive, which can impact their performance.

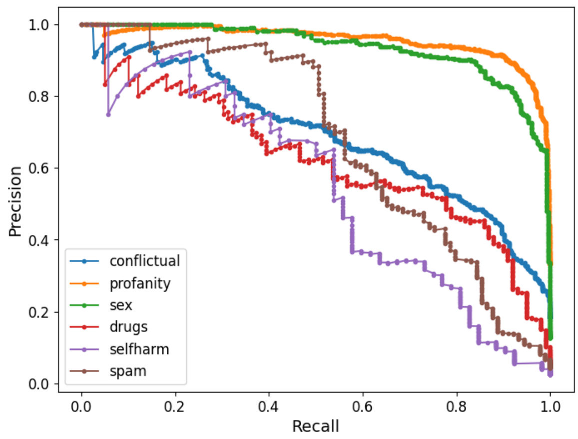

Figure 2: Precision-Recall curve for the fine-tuned *llama3-8b* in the multi-label setting.

6 Error Analysis

Aiming to understand better the dataset and the challenges that the models face on identifying sensitive content we consider the best performing model, the fine-tuned *llama3-8b*, and try to understand better its performance.

In the binary setting the model displays a strong performance and achieves high precision and recall values, 85.5%, 85.7% respectively, signifying its ability to effectively identify true positive cases with a relatively low number of false positives (Figure 4, Appendix C.4). In contrast, for the multi-label setting, the model seems to struggle with several categories as seen in Figure 2. Despite the strong performance of *llama-8b* in classifying Profanity and Sex labels, the model struggles with the Conflictual, Drugs, Self-harm, and Spam categories. As recall increases, precision for these categories drops significantly. This poses a particular challenge for health-related categories like Drugs and Self-harm, where high recall is critical, as missing cases could have serious consequences.

7 Conclusions

In this paper, we presented a complete research approach into sensitive content moderation in social media. Going beyond hate speech, we focus on categories that need to consistently be monitored in social media, let it be to filter to adult users or to remove from the platform, among others. We construct a multi-label dataset using six categories. The results show that LMs fine-tuned on our datasets are generally robust, although there are some categories where they are less precise, and hence these models are probably to be used as a support for human moderators. Nonetheless, the fact that these models perform at a high accuracy represents a useful tool to filter the most relevant messages for each category.

8 Limitations

In this paper, we introduce a valuable new resource expected to benefit a wide range of researchers and industry professionals. However, it is important to acknowledge several limitations. Firstly, the dataset is limited in size, which may restrict the generalisability and robustness of the models trained on it. Additionally, it exclusively contains English-language content due to budget constraints, potentially overlooking the nuances and challenges present in other languages.

The methodology used for aggregating the data in our dataset (Section 3.3) may also be subject to differing opinions. To facilitate transparency and further research, we plan to release all the collected annotations along with the dataset version used in our experiments. Moreover, the dataset was curated based on a specific selection of keywords, which might introduce biases and limit the diversity of the content. Another limitation is that the dataset is derived from only one social media platform, which may not fully represent the variety of sensitive content found across different platforms and contexts.

Finally, while we conduct an in-depth analysis using the results of six different models, there is significant room for improvement in terms of analysis and model development. This includes, investigating the performance of models of different architectures and optimising the prompts used[9].

9 Ethics Statement

We recognize the significance of the ACL Code of Ethics and are dedicated to adhering to its guidelines in our proposed task. Since our task involves user-generated content, we ensure user privacy by replacing each user mention in the texts with a placeholder, recognising the importance of anonymity, especially taking into account the potential for harm towards people expressing self-harming tendencies.

We also ensure fair treatment of the annotators who labelled the dataset by: 1) compensating them fairly at an average rate of 12$ per hour, and 2) not sharing or storing personal identification information. As annotator demographics play an important role in the perception of toxicity, following Prabhakaran et al. (2021), we release the data[10], disaggregated by individual annotator labels, while making sure that the demographic information is course enough to prevent deanonymization of the crowd-workers.

Lastly, recognise the sensitive and potentially dangerous nature of the dataset. However, we believe it is crucial to address and combat such behaviours. X-Sensitive will be shared under the CC BY-NC 4.0 Deed (Attribution-NonCommercial 4.0 International) following best practices in sharing social media-based data collections (Fiesler and Proferes, 2018; Assenmacher et al., 2020).

10 Acknowledgments

A big thank you to Debora Nozza for helpful discussions and feedback on the initial part of this project. Dimosthenis Antypas and Jose Camacho-Collados are supported by a UKRI Future Leaders Fellowship.

References

AI@Meta. 2024. Llama 3 model card.

Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, et al. 2023. Detecting harmful content on online platforms: what platforms need vs. where research efforts go. *ACM Computing Surveys*, 56(3):1–17.

Dennis Assenmacher, Indira Sen, Leon Fröhling, and Claudia Wagner. 2020. The end of the rehydration era the problem of sharing harmful twitter research data.

Gonzalo Molpeceres Barrientos, Rocío Alaiz-Rodríguez, Víctor González-Castro, and Andrew C Parnell. 2020. Machine learning techniques for the detection of inappropriate erotic content in text. *International Journal of Computational Intelligence Systems*, 13(1):591–603.

Priyam Basu, Tiasa Singha Roy, Soham Tiwari, and Saksham Mehta. 2021. Cyberpolice: Classification of cyber sexual harassment. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 701–714. Springer.

[9]The prompts used in our experiments can be found in Appendix C.3

[10]`https://huggingface.co/datasets/cardiffnlp/x_sensitive/blob/main/all_annotations.json`

George Beknazar-Yuzbashev, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Cody Buntain and Jennifer Golbeck. 2015. This is your twitter on drugs: Any questions? In *Proceedings of the 24th international conference on World Wide Web*, pages 777–782.

José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.

Robyn Caplan. 2018. Content or context moderation?

Stevie Chancellor, Michael L Birnbaum, Eric D Caine, Vincent MB Silenzio, and Munmun De Choudhury. 2019. A taxonomy of ethical tensions in inferring mental health states from social media. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 79–88.

Stevie Chancellor, Zhiyuan Lin, and Munmun De Choudhury. 2016a. " this post will just get taken down" characterizing removed pro-eating disorder social media content. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1157–1162.

Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016b. Quantifying and predicting mental illness severity in online pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1171–1184.

Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016c. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2111–2123.

Stevie Chancellor, Jessica Annette Pater, Trustin Clear, Eric Gilbert, and Munmun De Choudhury. 2016d. # thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In *Proceedings of the 19th ACM conference on computer-supported cooperative work & social computing*, pages 1201–1213.

Stevie Chancellor, Steven A Sumner, Corinne David-Ferdon, Tahirah Ahmad, and Munmun De Choudhury. 2021. Suicide risk and protective factors in online support forum posts: annotation scheme development and validation study. *JMIR mental health*, 8(11):e24471.

Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, fight back! detection of social media disclosures of sexual harassment. In *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Student research workshop*, pages 136–146.

Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, 10:1178222618792860.

Tao Ding, Arpita Roy, Zhiyuan Chen, Qian Zhu, and Shimei Pan. 2016. Analyzing and retrieving illicit drug-related posts from social media. In *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1555–1560. IEEE.

Jenny Fan and Amy X Zhang. 2020. Digital juries: A civics-oriented approach to platform governance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–14.

Casey Fiesler and Nicholas Proferes. 2018. "participant" perceptions of twitter research ethics. *Social Media+ Society*, 4(1):2056305118763366.

Eve Fleisig, Su Lin Blodgett, Dan Klein, and Zeerak Talat. 2024. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. *arXiv preprint arXiv:2405.05860*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

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

Google. 2023. Perspective api. Retrieved from https://www.perspectiveapi.com/.

Mitchell L Gordon, Michelle S Lam, Joon Sung Park, Kayur Patel, Jeff Hancock, Tatsunori Hashimoto, and Michael S Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1):2053951719897945.

Maanak Gupta, CharanKumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access*.

Naveed Hussain, Hamid Turab Mirza, Ghulam Rasool, Ibrar Hussain, and Mohammad Kaleem. 2019. Spam review detection techniques: A systematic literature review. *Applied Sciences*, 9(5):987.

Hakan Inan, Kartikeya Upasani, Jianfeng Chi, Rashi Rungta, Krithika Iyer, Yuning Mao, Michael Tontchev, Qing Hu, Brian Fuller, Davide Testuggine, et al. 2023. Llama guard: Llm-based input-output safeguard for human-ai conversations. *arXiv preprint arXiv:2312.06674*.

Jialun'Aaron' Jiang, Skyler Middler, Jed R Brubaker, and Casey Fiesler. 2020. Characterizing community guidelines on social media platforms. In *Companion Publication of the 2020 Conference on Computer Supported Cooperative Work and Social Computing*, pages 287–291.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

PM Lavanya and E Sasikala. 2022. Auto capture on drug text detection in social media through nlp from the heterogeneous data. *Measurement: Sensors*, 24:100550.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Luiz Henrique Quevedo Lima, Adriana Silvina Pagano, and Ana Paula Couto da Silva. 2024. Toxic content detection in online social networks: a new dataset from brazilian reddit communities. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 472–482.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Meena Devii Muralikumar, Yun Shan Yang, and David W McDonald. 2023. A human-centered evaluation of a toxicity detection api: Testing transferability and unpacking latent attributes. *ACM Transactions on Social Computing*, 6(1-2):1–38.

Ayodeji Oseni, Nour Moustafa, Helge Janicke, Peng Liu, Zahir Tari, and Athanasios Vasilakos. 2021. Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*.

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Vinodkumar Prabhakaran, Aida Mostafazadeh Davani, and Mark Diaz. 2021. On releasing annotator-level labels and information in datasets. *arXiv preprint arXiv:2110.05699*.

Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.

Kyle Robinson and Vijay Mago. 2022. Birds of prey: identifying lexical irregularities in spam on twitter. *Wireless Networks*, 28(3):1189–1196.

Ramit Sawhney, Prachi Manchanda, Puneet Mathur, Rajiv Shah, and Raj Singh. 2018. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 167–175.

Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *CSCW*.

Sean S Simpson, Nikki Adams, Claudia M Brugman, and Thomas J Conners. 2018. Detecting novel and emerging drug terms using natural language processing: a social media corpus study. *JMIR public health and surveillance*, 4(1):e7726.

Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of suicide ideation in social media forums using deep learning. *Algorithms*, 13(1):7.

Vankayala Tejaswini, Korra Sathya Babu, and Bibhudatta Sahoo. 2024. Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(1):1–20.

Samantha Thomson, Emily Kluftinger, and Jocelyn Wentland. 2018. Are you fluent in sexual emoji?: Exploring the use of emoji in romantic and sexual contexts. *The Canadian Journal of Human Sexuality*, 27(3):226–234.

Ana-Sabina Uban and Paolo Rosso. 2020. Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In *CEUR workshop proceedings*, volume 2696, pages 1–12. Sun SITE Central Europe.

Sahana Udupa, Antonis Maronikolakis, and Axel Wisiorek. 2023. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society*, 10(1):20539517231172424.

Qamar Un Nisa and Rafi Muhammad. 2021. Towards transfer learning using bert for early detection of self-harm of social media users. *Proceedings of the Working Notes of CLEF*, pages 21–4.

Bertie Vidgen, Adarsh Agrawal, Ahmed M Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar, Lora Aroyo, Trupti Bavalatti, Borhane Blili-Hamelin, et al. 2024. Introducing v0. 5 of the ai safety benchmark from mlcommons. *arXiv preprint arXiv:2404.12241*.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Data Collection

Table 3 displays the sources used for finding keywords used to collect tweets.

Category	Sources
Self-harm	(Chancellor et al., 2016b) (Chancellor et al., 2016d) (Chancellor et al., 2016a) (Chancellor et al., 2016c) Ngrams from the Reddit training corpus
Conflictual, profane, sexually explicit	IDEA-NTHU-Taiwan LDNOOBW
Drugs	talktofrank
Spam	(Founta et al., 2018) #f4f, #l4l, #follow4follow, #like4like

Table 3: Sources for seed words used for expanding our keyword lists.

Table 4 displays the total number of keywords and seed words used in our data collection.

Category	#seed-words	#keywords
Conflictual, profane, sexually explicit	203	1322
spam	76	767
drugs	118	519
self-harm	56	734

Table 4: Overall summary of the keyword list.

B Annotation

B.1 Guidelines

Figure 3 displays the guidelines provide to each coder for the annotation task.

B.2 Annotator Agreement

Table 5 displays the inter annotator agreement based on Krippendorff's alpha between different gender groups of coders. Similarly, Table 6 displays agreement scores between different codes age groups.

The objective of this experiment is to:

- Label a textual social media post (tweet, reddit post, snap, etc.,) with corresponding sensitive content categories. **Sensitive content is defined as any type of content that violates a social media platform's community guidelines.** Multiple categories can apply to a single post.
- Annotators are also asked to provide rationales, i.e., highlight which parts of the post led them to their decision. Rationales may be contiguous or disjoint spans of text.
- Finally, we also ask annotators to flag posts where they require additional context to make a judgment. Additional context includes - but is not limited to - poster background, preceding conversation, and content metadata.

Annotators should consider these content category options:

- **Hate speech.** Attack based on protected category like race, color, caste, ethnicity, national origin, religion, sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight or pregnancy status
- **Other conflictual language.** Attack based on other categories or without any mention of a category
- **Profanity.** Language containing slurs and profanity even if they are not directed towards a specific entity.
- **Sexually explicit content.** pornographic or other types of sexual content
- **Self-harm.** Posts depicting, promoting, or glorifying violence or harm against oneself such as eating disorders or suicide.
- **Spam.** irrelevant content that is unsolicited; or content that aims to drive traffic or attention from a conversation on the platform to promote, websites, products, services, or initiatives outside the platform; such as increasing the number of followers or selling a product in an insistent way.
- **Drugs.** Content that encourages, promotes or glorifies the use of regulated drugs. Also applicable to content that mentions sales, purchases, or the act of obtaining or trying to obtain regulated drugs. This category does not apply to cases which are merely talking about drugs (check examples below).

Annotators have the option to select more than one category for a post. Annotators may also mark posts for which additional context is required. If annotators have selected any content categories, we also ask them to highlight which parts of the post (words, sentences, emojis, etc.,) led to their judgment.

Figure 3: Guidelines provided to annotators.

Tables 7 and 8 show the percentage of tweets labelled as each class by age and gender groups of coders, respectively.

C Models

C.1 Resources

In total we estimate 112 hours used for the training of *RoBERTa, tlm* and *llama3-8b* models using a NVIDIA GeForce RTX 4090 GPU and 90 hours for inferences with the *llama3-8b* and *llama3-70b* models using an NVIDIA Quadro RTX 8000 GPU. Table 9 provides details for the models used in our experiments.

C.2 Taxonomies

Taxonomy used by Google's perspective API:

1. TOXICITY
2. SEVERE_TOXICITY
3. IDENTITY_ATTACK
4. INSULT
5. PROFANITY
6. SEXUALLY_EXPLICIT
7. THREAT
8. FLIRTATION
9. ATTACK_ON_AUTHOR
10. ATTACK_ON_COMMENTER
11. INCOHERENT
12. INFLAMMATORY
13. LIKELY_TO_REJECT
14. OBSCENE
15. SPAM
16. UNSUBSTANTIAL

In our experiments we utilise the following class mapping to the X-Sensitive taxonomy: "TOXICITY": Conflictual, "PROFANITY": Profanity, "SEXUALLY_EXPLICIT": Sexual Explicit Content, "SPAM": Spam.

MLCommons taxonomy used in Meta-LLama-guard:

1: Violent Crimes

2: Non-Violent Crimes

3: Sex-Related Crimes

4: Child Sexual Exploitation

5: Specialized Advice

6: Privacy

Gender	Multi	Bin
Man	0.49	0.57
Woman	0.49	0.55
Non-binary	0.47	0.82

Table 5: Krippendorff's alpha within each gender group of coders.

Age	Multi	Bin
18-25	0.51	0.57
26-39	0.51	0.59
40-65	0.47	0.52
over 65	0.47	0.54

Table 6: Krippendorff's alpha within each age group of coders.

Age	conflictual	profanity	sex	drugs	self-harm	spam	AVG	Coders
18-25	15	28	9	3	3	4	10	77
26-39	13	28	9	3	3	4	10	269
40-65	15	25	8	3	2	3	9	166
over 65	10	27	5	3	1	2	8	10

Table 7: Percentage of tweets labelled as each class for each age bracket of coders.

Gender	conflictual	profanity	sex	drugs	self-harm	spam	AVG	Coders
Woman	14	27	8	3	3	3	10	268
Non-Binary	20	28	10	2	3	3	11	12
Man	14	26	9	3	3	4	10	241

Table 8: Percentage of tweets labelled as each class for each gender of coders.

Model	Parameters
RoBERTa	355M
tlm	355M
llama3-8b	8B
llama3-70b	70B
chat-gpt	175B (approximate)

Table 9: Number of Parameters in different language models used.

7: Intellectual Property

8: Indiscriminate Weapons

9: Hate

10: Suicide & Self-Harm

11: Sexual Content

Openai's moderation API taxonomy:

1: harassment

2: harassment_threatening

3: hate

4: hate_threatening

5: self_harm

6: self_harm_instructions

7: self_harm_intent

8: sexual

9: sexual_minors

10: violence

11: violence_graphic

12: self-harm

13: sexual/minors

14: hate/threatening

15: violence/graphic

16: self-harm/intent

17: self-harm/instructions

18: harassment/threatening

In our experiments we utilise the following class mapping to the X-Sensitive taxonomy:

```
{
'harassment': 'Conflictual',
'harassment_threatening': 'Conflictual',
'hate': 'Conflictual',
'hate_threatening': 'Conflictual',
'self_harm': 'Self-Harm',
'self_harm_instructions': 'Self-Harm',
'self_harm_intent': 'Self-Harm',
'sexual': 'Sexual Explicit Content',
'sexual_minors': 'Sexual Explicit Content',
'self-harm': 'Self-Harm',
'sexual/minors': 'Sexual Explicit Content',
'hate/threatening': 'Conflictual',
'self-harm/intent': 'Self-Harm',
'self-harm/instructions': 'Self-Harm',
'harassment/threatening': 'Conflictual'
}
```

C.3 Prompts

Below are the prompt templates used in the zero- and few-shot settings.

Multi-label

Classify the following sentence in one or more of the categories listed below: 1. self-harm, 2. drugs, 3. conflictual language, 4. spam, 5. profane or bad words, 6 sexually explicit. Answer with the following format: 'ANSWER= 2, 3' if the sentence contains sensitive content or 'ANSWER= 0' if the sentence does not contain sensitive content.

Sentence: "{{ message }}"

Binary

Below there is a sentence. I would like to know if it contains sensitive content. We understand the sentence contains sensitive content if it treats topics like self-harm, drugs, if it contains conflictual language, spam, expresses profanity or it is sexually explicit. Does this tweet contain any kind of sensitive content? Please answer with the following format: 'ANSWER= 1' if the tweet contains sensitive content or 'ANSWER= 0' if the sentence does not contain sensitive content.

Sentence: "{{ message }}"

C.4 Extended Results

Figure 4 displays the Precision-Recall curve for the fine-tuned *llama3-8b* model in the binary setting.

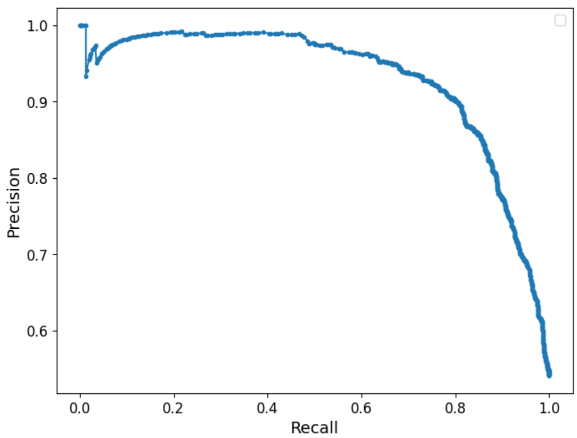

Figure 4: Precision-Recall curve for the fine-tuned *llama3-8b* model in the binary setting.