Stakeholder-driven Content Moderation of Sensitive Content

Keywords: hate speech, profanity, illicit content, content moderation, social media

Extended Abstract

Current frameworks and methods in NLP focus on hate speech and abusive language, but there is no holistic framework for building contextualized platform and use-case specific models. **In this work, we develop a framework for detecting textual content that violates a social media platform's community guidelines.** Bridging the literature on platform governance and NLP, we build a taxonomy of sensitive content categories grounded in the needs of content moderators. We develop a two-step framework (Figure 1) where we first use distant labeling and heuristics to build 'weak' phase one classifiers. The output of these classifiers are then used to curate a smaller dataset that is manually labeled by human annotators using our taxonomy. Finally, we leverage this smaller dataset to train more precise phase two classifiers.

Conceptualizing the Taxonomy. We use the community guidelines of various social media platforms to ground our taxonomy [2]. Using iterative coding, we refine, merge, and fix 5 broad categories and 8 specific subcategories of sensitive content which are mapped to rules in community guidelines. Our final categories and their definitions are:

1. Illicit Content

   (a) Regulated Goods

      i. **Weapons:** Content that encourages, promotes or glorifies the use of weapons or firearms. Also applicable to content that mentions sales, purchases, or the act of obtaining or trying to obtain firearms or weapons

      ii. **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.

   (b) **Sexually Explicit Content:** pornographic or other types of sexual content

2. Conflictual Language

   (a) **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,

   (b) **Other conflictual language:** Attack based on other categories or without any mention of a category

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

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

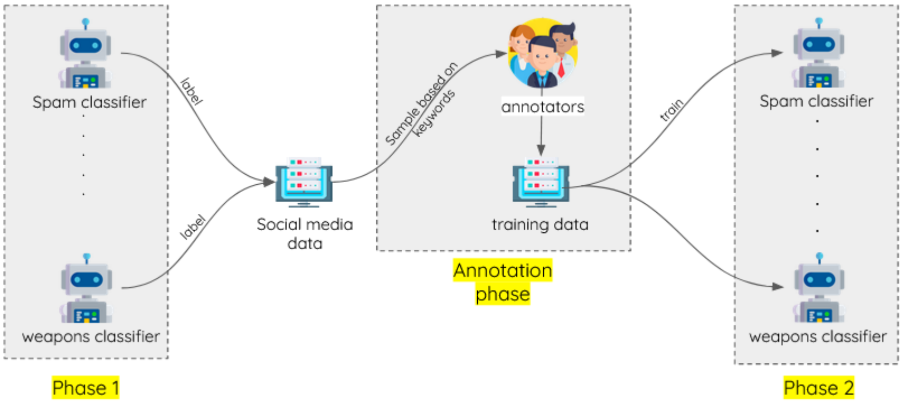

Figure 1: **Overview of our two-step framework** The phase-one classifiers are trained on existing or heuristics-based datasets and are distantly supervised. We use them to label a large social media dataset and then sample from it to smaller dataset. This dataset is carefully labeled by human annotators and we train the phase-two models on this dataset.

5. **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;

Methods. We opt for a two-step procedure for the automated methods for detecting sensitive categories in social media content. The first step entails working with a suite of 'weak' classifiers which might not be very accurate but allow us to sample relevant data with some signal that can be used to bootstrap stronger, more precise models. We sample data that has been labeled with these first-phase weak models for manual annotation using a comprehensive set of guidelines covering the aforementioned categories. This carefully manually labeled data is then used to train stronger phase two classifiers. To build the weak phase one classifiers, we curate heuristic-based datasets for each of our eight classes of sensitive content, either from existing datasets (for the case of 'hate speech') or through distant supervision (using data from self-harm related communities for 'self-harm').

Annotation Phase. Our phase one weak models are designed to capture overall topical information about the different sensitive categories. To state a concrete example, while our phase 1 models can detect drug-related tweets, they cannot differentiate between posts that advertise drug sales versus posts that report about drug sales. To build these more precise models, we leverage large-scale crowdsourcing and human annotation to build a high-quality labeled dataset of 13,065 tweets that is used to train our phase two models.

Future work. For each category of sensitive content, we finetune a binary XLM-T [1] classifier on its respective dataset. In future work, we will evaluate these phase-2 classifiers against existing baselines for detecting sensitive content and use zero-shot Large Language Labeling for more rare and understudied classes like 'sale or solicitation of weapons'.

References

[1] F. Barbieri, L. E. Anke, and J. Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *LREC*, 2022.

[2] M. K. Scheuerman, J. A. Jiang, C. Fiesler, and J. R. Brubaker. A framework of severity for harmful content online. *CSCW*.