Detecting Harmful Content on Online Platforms
What Platforms Need vs. Where Research Efforts Go
Arora, Arnav; Nakov, Preslav; Hardalov, Momchil; Sarwar, Sheikh Muhammad; Nayak, Vibha; Dinkov, Yoan; Zlatkova, Dimitrina; Dent, Kyle; Bhatawdekar, Ameya; Bouchard, Guillaume; Augenstein, Isabelle

Published in:
ACM Computing Surveys

DOI:
10.1145/3603399

Publication date:
2023

Document version
Peer reviewed version

Document license:
Unspecified

Citation for published version (APA):
DETECTING HARMFUL CONTENT ON ONLINE PLATFORMS: WHAT PLATFORMS NEED VS. WHERE RESEARCH EFFORTS GO

Abd Alfatah Almutairi
Qatar AI Institute
Copenhagen, Denmark
almutairi@qatarai.org

Preslav Nakov
Checkstep Research;
Mohamed bin Zayed University of Artificial Intelligence
Abu Dhabi, UAE

Momchil Hardalov
Checkstep Research;
Sofia University “St. Kliment Ohridski”
Sofia, Bulgaria

Sheikh Muhammad Sarwar
Checkstep Research;
University of Massachusetts, Amherst
USA

Vibha Nayak
Checkstep
London, UK

Yoan Dinkov
Checkstep Research
Sofia, Bulgaria

Dimitrina Zlatkova
Checkstep Research
Sofia, Bulgaria

Kyle Dent
Checkstep
London, UK

Ameya Bhatawdekar
Checkstep; Microsoft
Doimukh, Arunachal Pradesh

Guillaume Bouchard
Checkstep
London, UK

Isabelle Augenstein
Checkstep Research;
University of Copenhagen
Copenhagen, Denmark

ABSTRACT

The proliferation of harmful content on online platforms is a major societal problem, which comes in many different forms including hate speech, offensive language, bullying and harassment, misinformation, spam, violence, graphic content, sexual abuse, self harm, and many other. Online platforms seek to moderate such content to limit societal harm, to comply with legislation, and to create a more inclusive environment for their users. Researchers have developed different methods for automatically detecting harmful content, often focusing on specific sub-problems or on narrow communities, as what is considered harmful often depends on the platform and on the context. We argue that there is currently a dichotomy between what types of harmful content online platforms seek to curb, and what research efforts there are to automatically detect such content. We thus survey existing methods as well as content moderation policies by online platforms in this light and we suggest directions for future work.

1 INTRODUCTION

Online harms is a serious and growing problem, and targeted groups and individuals have suffered it for years. There are various types of harms, ranging from clearly illegal activities (e.g., child sexual abuse, human trafficking, and terrorist propaganda) to more subtle ones, such as abusive language and spam, which are not always illegal, but nevertheless harmful. As harmful content is frequent online — e.g., when surveyed for a 2019 report by the UK government, 23% of the 12–15-year-olds stated to have observed it within the last year [Barker & Jurasz 2019] — it is of particular concern to online communities, governments, and social media platforms. With this in mind, we present the first computing survey that relates computational solutions to harmful content detection to online platform policies, with focus on analysing to what extent existing research efforts are suitable to address the types of harms online platforms aim to curb.
While combating harm is a high priority, preserving individuals’ rights to free expression is also vital, which makes content moderation particularly difficult; yet, some form of moderation is clearly needed. Platform providers face very challenging technical and logistical problems in limiting harmful content, while at the same time wanting to allow a level of free speech that would enable rich and productive online interactions. Aiming to strike the right balance, many platforms institute guidelines and policies to specify what content is considered inappropriate. As manual filtering is hard to scale, and can even cause post-traumatic stress disorder in human annotators (mod, 2021), there has been significant research effort to develop tools and technologies to automate or assist in the process.

There have been several surveys of computational methods to detect and to address online harm. Schmidt & Wiegand (2017) and Fortuna & Nunes (2018) surveyed automated hate speech detection methods, but focused primarily on the features shown to be most effective in classification systems. Hardalov et al. (2022) and Guo et al. (2022) surveyed automated methods to address mis- and disinformation. Salawu et al. (2017) provided an extensive survey on detecting cyberbullying, and Vidgen & Derczynski (2021) worked on cataloging abusive language training data.

There are some studies closely related to ours, that focus on the content guidelines of online platforms. Gillespie (2018) qualitatively studied several policies in the content guidelines of more than 60 platforms. They stated how content guidelines are the platforms’ “most deliberate and carefully crafted statement of principles” outlining what content is not permitted on the platform and why. They also suggested to look at inspecting guidelines of a variety of platforms in order to understand the overlap and the differences between them, as we do in this study. Jiang et al. (2020) analysed the relative focus that platforms place on types of abuse to moderate. They iteratively coded and analysed the community guidelines of eleven major social media platforms for coverage of 66 different types of policies. Through their analysis, they outlined the imbalance in coverage of policies, speculating that “platforms may have chosen to focus on and made rules regarding the types of misbehaviour that is most rampant on their platform, or made explicit the rules that are most reflective of their values”. Our work builds on their study by categorising and computationally analysing the content guidelines of a larger set of platforms.

More concretely, we aim to bridge the gap between work on harmful content detection and platform solution requirements by surveying online platforms’ content policies, and relating them to proposed approaches by the research community. We further quantify the extent of this disconnect, and identify under-explored directions by analysing the platform requirements based on the keyword usage in their T&C clauses, and comparing those to what is currently available in the literature (both in published papers and in preprints). Nonetheless, no prior work has juxtaposed these policies with research directions and assessed the alignment between what platforms need vs. what technology has to offer. Hereby, we believe that such a survey is urgently needed.

2 Requirements of Online Platforms

Harmful content on online platforms can take on different forms. A 2019 white paper on online harms by the UK government considered the following sub-problems: online bullying and abuse, videos and images of children suffering sexual abuse, cyber-flashing, “pile-on” harassment, propaganda by terrorist groups, disinformation and misinformation Barker & Jurasz (2019). The paper noted that some of these are illegal, whereas others are lawful but potentially harmful.

Online platforms is a broad term, representing various categories of digital content providers such as social media, online marketplaces, online video communities, dating websites and apps, support communities and forums, and online gaming communities. Each of them is governed by its own content moderation policies and follows its own definition of harmful content. Below, we explore how online platforms define harmful content, and what content moderation policies they put in place accordingly by analysing their Terms and Conditions. Even though these T&Cs change over time, as platforms adapt to emerging issues such as the COVID-19 infodemic Alam et al. (2021), Nakov et al. (2022), we analyse policies that we downloaded at a specific timestamp.

Table lists the 42 online platforms that we study. We perform this selection with diversity of domain, content moderation needs of the platform as well as diversity of its user-base in mind in

1Our offline Terms and Conditions archive are added in the Supplementary Material
order to have a representative sample. Moreover, we only include platforms with publicly available content guidelines. Guidelines for what is acceptable or not on a platform is included either as part of their T&Cs document or as a separate community guidelines document. As such, for our analysis, we choose the document based on where the platform lists this information. Note that we exclude “alternative” platforms such as Parler, Bitchute, Gab, and Koo, which are often specifically designed and commonly used as a substitute to platforms like Twitter in order to circumvent their moderation; these are already covered in designated studies (Buckley & Schafer, 2022). We categorise the 42 platforms based on the domain in which they operate, since the domain defines the type of content they need to moderate and is thus reflected in the respective T&Cs. For instance, Reddit is a generic forum where any topic can be discussed, and thus it needs a wider range of policies compared to a specific forum such as Motley Fool, where the focus is on investing. The Mixed category includes platforms that provide services (and hence, T&Cs spanning multiple categories, e.g., Google has policies that spread across its different services including Meet, Mail, Help Communities, etc., each of which requires different expansiveness in its policy coverage.

### 2.1 Qualitative Analysis of the Terms and Conditions of Big Tech Platforms

We a conducted qualitative analysis of the terms and conditions of big social media and mixed platforms, commonly known as Big Tech. Big Tech has stringent content moderation policies as well as the most advanced technology to detect harmful content. We summarise this analysis in Table 2 we observe a lot of overlap, but also many differences.

<table>
<thead>
<tr>
<th>Category</th>
<th>Platforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dating</td>
<td>Bumble</td>
</tr>
<tr>
<td>Generic Forum</td>
<td>Quora, Reddit, Disqus, BG Mamma,</td>
</tr>
<tr>
<td>Specific forum - Gaming</td>
<td>Discord, Something Awful, Substack, Clubhouse</td>
</tr>
<tr>
<td>Specific forum - Finance</td>
<td>Instr, Money Saving Expert, Finimize, Public,</td>
</tr>
<tr>
<td>Specific forum - Health</td>
<td>Mumsnet, Student Doctor Network, Patient, Doctissimo, Flo Health, Strava</td>
</tr>
<tr>
<td>Specific forum - Other</td>
<td>Fiveable, Airbnb, Blind, The Student Room, Shutterstock</td>
</tr>
<tr>
<td>Online Marketplace</td>
<td>Amazon, Depop, NTWRK, Rarible</td>
</tr>
<tr>
<td>Social Media</td>
<td>Facebook, YouTube, Twitter, Girl Tribe, TikTok</td>
</tr>
<tr>
<td>Mixed</td>
<td>Google, Spotify, Apple</td>
</tr>
</tbody>
</table>

Table 1: Classification of the 42 online platforms we study.

<table>
<thead>
<tr>
<th>Policy Clause</th>
<th>Facebook</th>
<th>Twitter</th>
<th>Google</th>
<th>Apple</th>
<th>Amazon</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violence</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Dangerous org/people</td>
<td>✓</td>
<td>✓</td>
<td>Maps, Gmail, Meet*</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Glorifying crime</td>
<td>✓</td>
<td>✓</td>
<td>Maps, Gmail, Meet*</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Illegal goods</td>
<td>✓</td>
<td>✓</td>
<td>Maps, Google Chat and Hangout, Drive, Meet*</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Self-harm</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Child sexual abuse</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Sexual Abuse (Adults)</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Animal abuse</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Human trafficking</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Bullying and harassment</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Revenge porn</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Hate Speech</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Graphic content</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Nudity and pornography</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Sexual Solicitation</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Spam</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Impersonation</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Misinformation</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
<tr>
<td>Medical Advice</td>
<td>✓</td>
<td>✓</td>
<td>☑</td>
<td>☑</td>
<td>☑</td>
</tr>
</tbody>
</table>

Table 2: Summary of the terms and conditions of Big Tech. We use the following notation: ✓ – explicitly mentioned in the policy; ✗ – not mentioned in the policy; ☑ – implicitly mentioned in the policy; ? – broadly mentioned in the policy under a more generic clause; * mentioned in additional service-specific policy; † – the same policy applies to Instagram.

Facebook covers most policy clauses related to harmful language, though, e.g., the coverage of medical advice is described quite broadly, and it is thus unclear whether users are free to share medical advice in their posts. Google’s terms of service cover basic guidelines on acceptable conduct. The
more specific clauses regarding hate speech, bullying, and harassment are in service-specific policies, as shown in Table 2. Amazon and Apple offer very different services compared to Facebook, Twitter, and Google, which is reflected in the clauses covered in their terms of service.

Apple’s policies very broadly mention dangerous organisations, glorifying crime, illegal goods, child sexual abuse, sexual abuse (adults), animal abuse, and human trafficking under illegal acts. Violence falls under threatening and intimidating posts. Their policy does not mention medical advice, spam, sexual solicitation, revenge porn, graphic content, and self-harm. For Amazon, dangerous organisations and people, glorifying crime, illegal goods, child sexual abuse, sexual abuse (adults), animal abuse, and human trafficking are mentioned under illegal acts, which leaves room for ambiguity. Revenge porn, graphic content, nudity and pornography are also broadly mentioned in clauses pertaining to obscenity. However, there are no policies in place that directly address medical advice, misinformation, sexual solicitation, and self-harm.

2.2 Quantitative Analysis of all Platforms’ Terms and Conditions

In order to understand to what degree these patterns exist more widely, we further conducted a quantitative analysis of the T&Cs across the 42 platforms that we consider in our study (see Table 1).

2.2.1 Methodology

We quantified the presence of specific categories in the T&Cs by searching for policy-related keywords and by counting the number of occurrences we found. We compiled the keyword lists semi-automatically by looking for words and phrases commonly mentioned in the T&Cs when discussing a given topic. For instance, the keyword list for hate speech includes racist, ageist, disablist, and homophobic, among others. A caveat to note here is that the number of search keywords differs across the policy clauses, which could affect the number of matches we found. Moreover, the T&Cs are often long, wide-ranging, and spread across several documents. Thus, in order to perform a more comprehensive search, we scraped all pages hyperlinked in the main T&C document. We further converted them to lowercase and we removed emails, numbers, URLs, and punctuation.

2.2.2 Results

By platform category We outline the policy presence across all platforms in Figure 1. Confirming what we observed in the qualitative analysis, we found that social media platforms tended to have the most expansive policies, with all topics being present in the T&Cs of Facebook and Twitter, while TikTok and YouTube covered all but sexual solicitation.

The outlier here is Girl Tribe by MissMalini, which is a much smaller platform exclusively for women. The expansiveness compared to other categories could be attributed to the additional level of public scrutiny that social media companies are subject to.

While on average, generic fora cover more categories compared to specific fora, there is considerable variance within the specific fora with Gaming having a higher average policy coverage compared to generic fora, whereas fora pertaining to Finance have the least coverage among all categories. We attribute this to the increased attention that Gaming fora have received in the media as well as by regulatory authorities, and to accusations that they are not taking enough action to moderate harmful content (News, 2021).

By policy Interesting patterns can be observed in a per-policy analysis of the T&C documents. Figure 2a plots the number of platforms that cover each policy, which gives a sense of relative requirement for moderation for these policies across the platforms. We can see that Hate speech and graphic content policies are included in the T&Cs of all platforms except for Gatsby, followed by violence, illegal content and spam, respectively. Sexual Solicitation is the least covered, with only some Social Media and Gaming fora listing it. Self-harm is covered by all Social Media platforms, but it is not explicitly stated in half of the Health fora and in the Dating platform Bumble’s policies.

---

2Gatsby does not cover any of the policies taken into account in this work.

3Yet, this kind of content is illegal in certain parts of the world and can have wide-ranging consequences; many platforms had to scramble to remove sexual solicitation and sexual content from their websites, most notably Craigslist removed their entire personals section to deal with the issue.
where it ought to be important. A similar pattern can be observed for *Sexual Abuse* and *Human Trafficking*, which have low coverage in *Dating* and *Health* fora.

Figure 2b shows the number of times each topic is mentioned across all platform T&Cs. While multiple mentions do not perfectly capture how important a particular topic is, it does give us a sense of how much focus a platform is paying to a particular policy clause across several spread out pages of its T&Cs. Similarly to Figure 2a, *Violence*, *Graphic Content* and *Hate Speech* are the most mentioned topics, demonstrating the significant attention paid to them as well as the importance of detecting policy violations regarding these topics. *Sexual Solicitation* and *Political Propaganda* are the topics with the fewest mentions. For *Political Propaganda*, this can be partly attributed to its narrower focus, e.g., *Gaming* and *Finance* fora are unlikely to host large amounts of political content. *Child Sexual Abuse* is in the same range of mentions as *Misinformation*, *Bullying and Harassment* and *Spam*, while being covered by fewer platforms.

---

(a) number of platform T&Cs out of 42 mentioning each topic

(b) number of mentions of each topic across all platform T&Cs

Figure 2: Topic mentions across the platform Terms and Conditions (T&Cs).
3 Automatic Harmful Content Detection

Below, we first analyse the publications in arXiv and to what extent they cover the above-described topics of concern in the T&Cs, and then we discuss some specific research, with focus on offensive language.

3.1 Analysing Publications in arXiv

In order to estimate how much research attention has been paid to each T&C topic, we looked for mentions of policy-related keywords in arXiv within the cs subject classification for 2015–2021, taking the number of results as a proxy for the amount of work conducted pertaining to that topic. We show the results in Figure 3. We can see that the number of papers on Illegal Content is consistently higher than for the other topics. We further see that work on Hate Speech and Misinformation has seen a steep growth since 2017, which reflects the attention these topics have attracted in recent years. Spam is another topic with a consistently high volume of publications. In contrast, topics such as Sexual Solicitation, Child Sexual Abuse, and Human Trafficking have seen very little research. We attribute this to the difficulty of getting access to datasets and performing research, due to the sensitive nature of these topics.

Figure 3: Number of papers found in arXiv for the different topics over time.

Work on harmful content detection has addressed several aspects of the above issues, but we are unaware of any work that covers them comprehensively. Vidgen & Derczynski (2021) attempted to categorise the available datasets with focus on abusive language detection, analysing a total of 64 datasets. They found that around half of the datasets covered English only, with some datasets covering European languages, Hindi and Indonesian, and six datasets covering Arabic. The primary source of the data was Twitter, but there were also datasets using content from other social networks. Overall, the size of most datasets was under 50,000 instances, and under 5,000 instances for half of the datasets.

3.2 Approaches

There has been a lot of research effort aimed at detecting specific types of offensive content, e.g., hate speech, offensive language, cyberbullying, and cyber-aggression. Below, we briefly describe some relevant tasks, datasets, and approaches.

Hate Speech Detection This is by far the most studied harmful language detection task Waseem & Hovy (2016), Kwok & Wang (2013), Burnap & Williams (2015), Ousidhoum et al. (2019), Chung et al. (2019), Sarwar & Murdock (2022). A recent shared task on the topic is HateEval Basile et al. (2019) for English and Spanish. The problem was also studied from a multimodal perspective,
e.g., [Sabat et al., 2019] developed a collection of 5,020 memes for hate speech detection. More recently, the Hateful Memes Challenge by Facebook introduced a dataset consisting of more than 10K memes, annotated as hateful or non-hateful [Kiela et al., 2020]: the memes were generated artificially, so that they resemble real memes shared on social media, along with “benign confounders.” More recent work studied harmful memes [Pramanick et al., 2021a;b; Sharma et al., 2022] and propagandistic memes [Dimitrov et al., 2021a;b].

Offensive Language Detection

One of the most widely used datasets is the one by [Davidson et al., 2017], which contains over 24,000 English tweets labelled as non-offensive, hate speech, and profanity. There have been several shared tasks with associated datasets that focused specifically on offensive language identification, often featuring multiple languages: OffensEval 2019–2020 [Zampieri et al., 2019b; 2020] for English, Arabic, Danish, Greek, and Turkish, GermEval 2018 [Wiegand et al., 2018] for German, HASOC 2019 [Mandl et al., 2019] for English, German, and Hindi, TRAC 2018–2020 for English, Bengali, and Hindi [Fortuna et al., 2018; Kumar et al., 2020]. Offensive language was also studied from a multimodal perspective, e.g., [Mittos et al., 2020] analysed memes shared on 4chan as offensive vs. non-offensive.

Aggression Detection

The TRAC shared task on Aggression Identification [Kumar et al., 2018] provided participants with a dataset containing 15,000 annotated Facebook posts and comments in English and Hindi for training and validation. For testing, two different sets, one from Facebook and one from Twitter, were used. The goal was to discriminate between three classes: non-aggressive, covertly aggressive, and overtly aggressive.

Toxic Comment Detection

The Toxic Comment Classification Challenge [Jigsaw, 2018] was a Kaggle competition, which provided participants with almost 160K comments from Wikipedia organised in six classes: toxic, severe toxic, obscene, threat, insult, and identity hate. The dataset was also used outside of the competition [Georgakopoulos et al., 2018], including as additional training material for the TRAC shared task [Fortuna et al., 2018]. It task was later extended to multiple languages [Jigsaw Multilingual, 2020] offering 8,000 Italian, Spanish, and Turkish comments. Recently, [Juuti et al., 2020] presented a systematic study of data augmentation techniques in combination with state-of-the-art pre-trained Transformer models for toxic language detection. A related Kaggle challenge features Detecting Insults in Social Commentary. Other datasets include Wikipedia Detox [Wulczyn et al., 2017], and the dataset by [Davidson et al., 2017].

Cyberbullying Detection

[Xu et al., 2012] used sentiment analysis and topic models to identify relevant topics. [Dadvar et al., 2013] relied on user-related features such as the frequency of profanity in previous messages, and [Rosa et al., 2019] presented a systematic review of automatic cyberbullying detection research.

Abusive Language Detection

There have also been datasets that cover various types of abusive language. [Founta et al., 2018] tackled hate and abusive speech on Twitter, introducing a dataset of 100K tweets. [Glavas et al., 2020] targeted hate speech, aggression, and attacks in three different domains: Fox News (from GAO), Twitter/Facebook (from TRAC), and Wikipedia (from WUL). In addition to English, it further offered parallel examples in Albanian, Croatian, German, Russian, and Turkish. However, the dataset was small, containing only 999 examples. Among the popular approaches for abusive language detection are cross-lingual embeddings [Ranasinghe & Zampieri, 2020], cross-lingual neighbourhood transformer representations [Sarwar et al., 2022], and deep learning [Founta et al., 2019]. As abusive language has many aspects, there have been several proposals for multi-level taxonomies covering different types of abuse. This enables understanding how different types and targets of abusive language relate to each other, and informs efforts to detect them. For example, offensive messages targeting a group are likely hate speech, while when the target is an individual, this is likely cyberbullying. [Waseem et al., 2017] presented a taxonomy that differentiates between (abusive) language directed towards a specific individual or entity, or towards a generalised group, as well as between explicit and implicit abusive content. [Wiegand et al., 2018] extended this idea to German tweets: they developed a model to detect offensive vs. non-offensive tweets, and further sub-classified the former as profanity, insult, or abuse. [Zampieri et al., 2021a;b].

https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification
(2019a) developed a very popular three-level taxonomy, which considers both the type and the target of offence. The taxonomy served as the basis of the OLID dataset of 14,000 English tweets, which was used in two shared tasks (OffensEval) at SemEval in 2019 Zampieri et al. (2019b) and in 2020 Zampieri et al. (2020). For the latter, an additional large-scale dataset was developed, consisting of nine million English tweets labelled in a semi-supervised fashion Rosenthal et al. (2021). This new dataset enabled sizeable performance gains, especially at the lower levels of the taxonomy. The taxonomy was also adopted for Arabic Mubarak et al. (2021), Danish Sigurbjergsson & Derczynski (2020), Greek Pitenis et al. (2020), and Turkish Çöltekin (2020).

Troll detection has been addressed using semantic analysis Cambria et al. (2010), domain-adapted sentiment analysis Seah et al. (2015), various lexico-syntactic features about user writing style and structure Chen et al. (2012); Mihaylov et al. (2015a), Mihaylov & Nakov (2016), as well as using graph-based approaches Ortega et al. (2012); Kumar et al. (2014). There have also been studies on general troll behaviour Herring et al. (2002); Buckels et al. (2014); Mihaylov et al. (2018); Nakov et al. (2017) and role Mihaylov et al. (2015b); Atanasov et al. (2019), cyberbullying Galán-García et al. (2016); Sarna & Bhatia (2017); Wong et al. (2018); Sezer et al. (2015), as well as on linking fake troll profiles to real users Galán-García et al. (2016). Some studies related to cyberbullying have already been applied in real settings in order to detect and to stop cyberbullying in elementary schools using a supervised machine learning algorithm that links fake profiles to real ones on the same social medium Galán-García et al. (2016).

Identification of malicious accounts in social networks is another important research direction. This includes detecting spam accounts Almaatouq et al. (2016); McCord & Chuah (2011), fake accounts Fire et al. (2014); Cresci et al. (2015), compromised accounts and phishing accounts Adewole et al. (2017). Fake profile detection has also been studied in the context of cyberbullying Galán-García et al. (2016).

Web spam detection was addressed text classification Sebastiani (2002), e.g., using spam keyword spotting Dave et al. (2003), lexical affinity of words to spam content Hu & Liu (2004), frequency of punctuation, and word co-occurrence Li et al. (2006).

**Methods** The most common way to address the above tasks is to use pre-trained transformers (Vaswani et al., 2017): typically BERT Devlin et al. (2019), but also RoBERTa Liu et al. (2019), ALBERT Lan et al. (2020), and GPT-2 Radford et al. (2019). In a multi-lingual setup, also mBERT Devlin et al. (2019) and XLM-RoBERTa Conneau et al. (2020) have proved useful. CNNs Fukushima (1980), RNNs Rumelhart et al. (1986), and GRUs Cho et al. (2014), including ELMo Peters et al. (2018). Older models such as SVMs are sometimes also used, typically as part of ensembles. Moreover, lexic such as HurtLex Bassignana et al. (2018) and Hatebase are also popular. For multi-modal detection of these harms, popular approaches include Visual BERT, ViLBERT, VLP, UNITER, LXMERT, VILLA, ERNIE-Vil, Oscar and various Transformers Li et al. (2019); Su et al. (2020); Zhou et al. (2020); Tan & Bansal (2019); Gan et al. (2020); Yu et al. (2021); Li et al. (2020); Lippe et al. (2020); Zhu (2020); Muennighoff (2020); Zhang et al. (2020); Kumar & Nandakumar (2022).

4 LESSONS LEARNED AND MAJOR CHALLENGES

Despite the stringent policies in place, the amount of harmful content online keeps growing. Research is complicated as online platforms have different understanding of harmful content, and platform’s Terms and Conditions (T&Cs) try to imagine every possible outcome of particular statements, while trying to be transparent about the areas they fall short in. The reality is that some clauses in the T&Cs are hard to automate and require human attention and understanding. Below, we describe some of these challenges in more detail. The list is not exhaustive, but it illustrates the main challenges faced by online platforms and researchers, as new issues regularly appear, policies evolve, and new detection technologies are constantly being developed.

**Mismatch in Goals** Table 3 shows a comparison of the number of papers on arXiv that mention a specific type of harmful content and the corresponding number of platforms that mention that type.

in their T&Cs. We can see that graphic content and violence are relatively less researched in spite of being covered by large number of platforms. On the other hand, misinformation, illegal content, etc. have attracted a lot of research. As is readily apparent when comparing T&Cs documents of online platforms (Section 2) to harmful content tasks tackled in research (Section 3), there is a clear disconnect between what types of content platforms seek to moderate, and what current harmful content detection research offers, as demonstrated by the high variance in the ratio column in Table 3.

Lack of participatory design. While this work focuses on platform policies and academic literature around content moderation, it is the users of a platform who are at the receiving end of decisions based on these policies and algorithms. Hence, a people-centric and participatory process model for AI-assisted content moderation is crucial (Udupa et al., 2022). Following a one-size-fits-all approach does not work when norms, values, and what is considered harmful differs across communities (Chandrasekharan et al., 2018) and cultures (Jiang et al., 2021).

Policy challenges. It is hard to define harmful content policy that is congruent with the harm that such content inflicts. For example, Facebook’s hate speech policy is tiered, where Tier 1 aims to define the kind of hate speech that is considered most harmful. The policy covers content targeting a person or a group of people based on their protected characteristics such as race, ethnicity, national origin, disability, gender, etc. According to this policy, certain offensive statements are considered more harmful than others. Moreover, a number of policies are left under-specified or are ambiguous in what content is allowed on the platform. The policies are also fluid, change based on the circumstances, and are affected by events in the real world (Vengattil & Culliford, 2022), which makes them challenging for researchers to follow.

Context. An obvious challenge that comes with the need to moderate unlawful content is that platforms are operating in multiple countries with a wide spectrum of legal standards. Similarly, the same content considered in different contexts can be either perfectly benign or can be perceived as harmful. Such context also changes with time, e.g., certain slur words are appropriated by their target groups as reclaimed speech. Such historical, legal, and cultural context is not explicitly encoded into machine learning models at present and makes it impossible for them to accurately detect and for platforms to moderate this content.

Dataset Availability As outlined earlier, data for sensitive policies such as sexual solicitation and child sexual abuse are hard to collect and share. However, even for categories such as misinformation and bullying and harassment, access to data is primarily at the discretion of the platforms. While there have been efforts to increase research access, these remain insufficient. Moreover, since violating content is often taken down, researchers enter a timed race with platforms in their efforts to collect content before it is removed or only have access to the content that platforms do not remove, making it incredibly hard for them to develop systems capable of detecting or assessing the potential harms caused by such content. Hence, there is a need for platforms to develop privacy-preserving ways of sharing datasets.

Content is Interlinked. People use text, images, videos, audio files, GIFs, emojis — all together — to communicate. While each modality by itself may be seemingly benign, when combined, they might make content that violates the T&Cs. Thus, there is a need for holistic understanding of the content.

Dataset Specificity Most publicly available datasets are specific to the type of harm, to the targeted group, or to other fine-grained distinctions, as harmful content can be highly context-sensitive. Further, in the hierarchical annotation schemes introduced by these datasets, lower levels of the taxonomy contain a subset of the instances in the higher levels, and thus there are fewer instances in the categories in each subsequent level. This diversity has led to the creation of many datasets, most of which are small in size, and much of the research on harmful content detection is fragmented, with very few studies viewing the problem from a holistic perspective.
Table 3: Relative research focus in terms of number of research papers in arXiv that study a given type of harmful content and the number of online platforms that aim to moderate such content, as well as a ratio thereof.

<table>
<thead>
<tr>
<th>Topic</th>
<th>arXiv Papers</th>
<th>Online Platforms</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>sexual solicitation</td>
<td>1</td>
<td>5</td>
<td>0.20</td>
</tr>
<tr>
<td>child sexual abuse</td>
<td>12</td>
<td>26</td>
<td>0.46</td>
</tr>
<tr>
<td>sexual abuse</td>
<td>24</td>
<td>24</td>
<td>1.00</td>
</tr>
<tr>
<td>graphic content</td>
<td>79</td>
<td>41</td>
<td>1.93</td>
</tr>
<tr>
<td>human trafficking</td>
<td>53</td>
<td>15</td>
<td>3.53</td>
</tr>
<tr>
<td>self harm</td>
<td>102</td>
<td>22</td>
<td>4.64</td>
</tr>
<tr>
<td>dangerous organisations</td>
<td>142</td>
<td>29</td>
<td>4.90</td>
</tr>
<tr>
<td>violence</td>
<td>255</td>
<td>40</td>
<td>6.38</td>
</tr>
<tr>
<td>bullying and harassment</td>
<td>276</td>
<td>37</td>
<td>7.46</td>
</tr>
<tr>
<td>impersonation</td>
<td>192</td>
<td>19</td>
<td>10.11</td>
</tr>
<tr>
<td>spam</td>
<td>434</td>
<td>39</td>
<td>11.13</td>
</tr>
<tr>
<td>political propaganda</td>
<td>209</td>
<td>11</td>
<td>19.00</td>
</tr>
<tr>
<td>hate speech</td>
<td>1089</td>
<td>41</td>
<td>26.56</td>
</tr>
<tr>
<td>illegal content</td>
<td>1113</td>
<td>40</td>
<td>27.82</td>
</tr>
<tr>
<td>misinformation</td>
<td>1029</td>
<td>35</td>
<td>29.40</td>
</tr>
</tbody>
</table>

5 CONCLUSION

We discussed recent methods for harmful content detection in the context of the content policies of online platforms. We inspected the content guidelines of 42 platforms and compared the needs highlighted there to available research. We found a substantial imbalance in the relative focus to different policy clauses in research vs. by platforms. For instance, misinformation and political propaganda are among the clauses with a high research-papers-to-platform-coverage ratio, while sexual solicitation and graphic content received limited attention, despite being critical for a number of platforms. While academic research should not aim to follow industrial needs, shedding light on this relative focus can aid in identifying major challenges and possible remedies that harmful content detection faces going forward. We argued that current research does not produce the tools for automating content moderation on online platforms that matched their needs. We further outlined some major challenges that need to be addressed.

REFERENCES

Facebook moderators call for better mental health support, end to NDAs, 2021. Section: Social Media.


Nicole Buckley and Joseph S. Schafer. ‘Censorship-free’ platforms: Evaluating content moderation policies and practices of alternative social media. For(e)Dialogue, 4(Vol 4, Issue 1), 2022.


Erik Cambria, Praphul Chandra, Avinash Sharma, and Amir Hussain. Do not feel the trolls. In SDow, Shanghai, China, 2010.


BBC News. TikTok and Twitch face fines under new Ofcom rules, 2021.


Munsif Vengattil and Elizabeth Culliford. Facebook allows war posts urging violence against Russian invaders, 2022.


