HarmPot: An Annotation Framework for Evaluating Offline Harm Potential of Social Media Text

Ritesh Kumar[1,2], Ojaswee Bhalla[1], Madhu Vanthi[1], Shehlat Maknoon Wani[1], Siddharth Singh[2]
[1]Council for Strategic and Defense Research, New Delhi,
[2]UnReaL-TecE-LLP, Agra
unreal.tece@gmail.com

Abstract

In this paper, we discuss the development of an annotation schema to build datasets for evaluating the offline harm potential of social media texts. We define "harm potential" as the potential for an online public post to cause real-world physical harm (i.e., violence). Understanding that real-world violence is often spurred by a web of triggers, often combining several online tactics and pre-existing intersectional fissures in the social milieu, to result in targeted physical violence, we do not focus on any single divisive aspect (i.e., caste, gender, religion, or other identities of the victim and perpetrators) nor do we focus on just hate speech or mis/dis-information. Rather, our understanding of the intersectional causes of such triggers focuses our attempt at measuring the harm potential of online content, irrespective of whether it is hateful or not. In this paper, we discuss the development of a framework/annotation schema that allows annotating the data with different aspects of the text including its socio-political grounding and intent of the speaker (as expressed through mood and modality) that together contribute to it being a trigger for offline harm. We also give a comparative analysis and mapping of our framework with some of the existing frameworks.

Keywords: Offline Harm, Harm Potential, Tagset, HarmPot

1. Background and Rationale

India is a country with a rapidly proliferating social media presence with over 700 million users (including 81% of teens). However, despite massive levels of social media usage, digital media literacy remains low in India. A 2020 survey of a "highly educated online sample" of Indians found that roughly 50% of the fake news presented to them was judged as "accurate" or "very accurate" in their control group (Guess et al., 2020). The wide reach of social media content, the high prevalence of false or misleading information online, and the extreme communalism/groupthink on social media (for example, see Al-Zaman (2019) for a study on communalism in India and Mukherjee (2020) for a discussion on the impact of online groupthink on mob violence in India), coupled with low levels of media discernment skills, have exacerbated long-standing social divisions in India (Froerer, 2019; Banerjee and Ghosh, 2018). India has now become a hotbed for online content spurring real-world physical violence. Online rumours and hate speech leading to physical violence against targeted communities and the subsequent filming of lynching is no longer uncommon in India [1].

Online hate has compounded with pre-existing lines of oppression, incentivizing the publicizing of violence against targeted groups for the sake of gaining public recognition and even praise. Several incidents of lynching have been triggered by misinformation around caste (Staff, 2020; Sajlan, 2021), love-jihad (Muslim men eloping with Hindu women) (NewIndianXpress, 2018), religious desecration etc. There are additional contextual triggers that often cause increased levels of online content and subsequent real-world harm. This includes elections (Deka, 2019), where fake news, rumours, misleading and divisive content are typically spiked for political gains, and global crises like the COVID-19 pandemic (Al-Zaman, 2021), which create a context in which users want "someone to blame", often unjustly.

In the last few years, over 60 datasets of various sizes and kinds, where a wide variety of abusive language has been annotated, have been released publicly (Vidgen and Derczynski, 2020; Poletto et al., 2021). Existing tools such as Hatebase.org, or the Twitter-backed Hate-Lab or a host of other recent studies have focussed on identifying abusive language (Nobata et al., 2016; Waseem et al., 2017), toxic language (Kolhatkar et al., 2020; Kaggle, 2020), aggressive language (Haddad et al., 2019; Kumar et al., 2018b; Bhattacharya et al., 2020), offensive language (Chen et al., 2012; Mubarak et al., 2017; Nascimento et al., 2019; de Pelle and Moreira, 2016; Schäfer and Burtenshaw, 2019; Zampieri et al., 2019a,b, 2020; Kumar et al., 2021; Steinberger et al., 2017), hate speech (several including (Akhtar et al., 2019; Albadi et al., 2018; Alfina et al., 2017; Bohra et al., 2018; Davidson et al., 2017; Malmasi and Zampieri,

[1] In 2018, for example, rumours and accusations of certain individuals being child-lifters, primarily spread on social media led to five additional instances of mob killings.

LREC-COLING 2024, pages 8016–8034
20-25 May, 2024.

2017; Schmidt and Wiegand, 2017; Del Vigna et al., 2017; Fernquist et al., 2019; Ishmam and Sharmin, 2019; Sanguinetti et al., 2018)), threatening language (Hammer, 2017), or narrower, more specific dimensions such as sexism (Waseem, 2016; Waseem and Hovy, 2016)), misogyny, Islamophobia (Chung et al., 2019; Vidgen and Yasseri, 2020), and homophobia (Akhtar et al., 2019). Some datasets include a combination of these such as hate speech and offensive language (Martins et al., 2018; Mathur et al., 2018)), or sexism and aggressive language (Bhattacharya et al., 2020). However, while most of these datasets and frameworks aim to model whether hate or offensive speech has been used or not, there has been no dataset or framework that could directly model the relationship and interdependence of online content and offline incidents of harm and violence.

In this paper, we discuss the development of a framework - HarmPot - that could be used for annotating the text with textual and contextual information such that the annotated dataset could be used for training models that could predict the offline harm potential of online content. In the following sections, we discuss the detailed annotation schema and annotation guidelines, a comparison with the other popular hate speech schema and finally some details of a new dataset annotated with the data.

2. The HarmPot Framework

"Harm Potential" (HarmPot) could be defined as the potential for an online public post to cause offline, real-world physical harm (i.e., violence). Targeted real-world violence is often spurred by a web of triggers, often combining several online tactics and pre-existing intersectional fissures in the social milieu. As such we do not focus on any single divisive aspect (i.e., caste, gender, religion, or other identities of the victim and perpetrators) nor do we focus on just hate speech or mis/dis-information. Rather, we focus on marking the harm potential of online content within a specific set of intersectional, contextual factors, irrespective of whether it is hateful or not. The HarmPot framework is designed to answer the following set of questions for a given text -

***Who** is being talked to, **when**, **how**, **why** and all this results in **what magnitude** of harm potential for the addressee?*

Each of these questions is answered by using a set of parameters, defined in our tagset. We discuss each of these in the following subsections.

2.1. Magnitude of Harm Potential

Depending on what kind of offline harm the text could lead to, we define two broad kinds of harm potential -

- **Physical Harm Potential:** It defines the potential of a text to lead to acts of physical violence such as murder, mob lynching, thrashing and beating, etc.
- **Sexual Harm Potential:** It defines the potential of a text to lead to acts of sexual violence such as rape (or rape threats), molestation, sexual harassment, etc.

Both of these harm potentials are classified on a scale of 0 - 3, defined below.

Value 0: A text will be marked as having '0' harm potential in the following cases:

1. Texts which are a part of the dataset but do not relate to any specific incident of violence or larger narrative campaign.

   E.g., "I felt like a bulldozer trying to catch a butterfly."

2. Texts which are blurbs accompanying links to news reports.

   E.g., "How 'Kashmir Files' added to communal fires in Khargone that ended with bulldozer injustice https://t.co/hQUQ5tsz26"

3. Texts that criticise public figures and not protected identities.

   E.g., "@AtishiAAP @ArvindKejriwal @UN madam kuch bulldozer ka bhi bolna tha aaj desh ki kya haalat hai, woh bolna that". *@AtishiAAP @ArvindKejriwal @UN madam you ought to have spoken about bulldozer as well, you ought to have spoken about the situation in the country today.*

Value 1: [2] A text will be marked as having '1' harm potential, if it is likely to lead to offline harm in very few, specific contexts but more generally is not expected to trigger incidents of offline harm. The most stereotypical instances of such texts include

1. Texts that target communities by using slurs and pejorative terms.
2. Texts that reinforce negative stereotypes regarding a particular community.

[2] The examples of these harm potentials included in the categories discussed below

Value 2: A text that is likely to trigger offline harm in most of the contexts - it is only in very specific contexts that it may not be interpreted as a call to violence - is marked as '2' on the harm potential scale. Some of the most stereotypical instances of such texts include

1. Explicit cases of attack or accusations against communities.
2. Justifying violence or discrimination against communities.

Value 3: Any text that has a high potential of triggering offline harm, irrespective of the context that it occurs in is marked as '3' on the harm potential scale. Instances of such texts include

1. Explicit and clear calls to violence against communities or people.
2. Explicit and clear attempts to instigate violence against communities or people.

The magnitude of harm potential is marked at two levels -

1. **Text Span:** It is marked in conjunction with specific spans of text that are used to refer to specific identities. It refers to the potential of that specific span of text to trigger offline harm/violence against specific identities (refer to Section 2.2 for details).
2. **Document:** It is the overall harm potential of the document - generally it is calculated based on the harm potential of individual spans; however, in cases where none of the spans refers to specific identities then an overall harm potential of the document is independently ascertained.

2.2. Who is being talked to?

This parameter is used to identify the specific types of identities that are 'mentioned/referred' (and not necessarily targeted) in a particular 'span of text'. We discuss the various ontological types of identities that can be potentially targeted in a text. Since this parameter works with the magnitude of harm potential, a text span which simply mentions an identity without targeting it will have '0' harm potential.

Three broad annotation instructions are for this parameter -

1. **Intersectionality:** If more than one identity of the same individual is referred to (viz. Female Dalit or Pakistani Hindu) then the same span is marked with all the identities and the same harm potential is ascribed in all instances - this is how intersectionality is handled in the framework. If different identities are mentioned in different spans then also different spans will carry the same harm potential, considering it to be an instance of intersectionality.
2. **Multiple Identities:** If different identities of different individuals are referred to then they might have different harm potential.
3. **Multiple Spans:** If more than one span refers to the same identity of the same or different persons, each span could have different harm potentials.

The framework itself does not enforce a specific set of identities to be marked. However, for the current project, the following non-exhaustive set of identities have been marked in the dataset. If needed, more, less or different kinds of specific subtypes of these categories may also be marked in the text. For each identity and its sub-category, a set of additional guidelines was used for deciding whether its harm potential is '0' or not, as discussed below - if it's not '0' then the guidelines for marking the magnitude of harm potential are to be used.

Caste: A span is annotated as targeting this identity if there are threats of violence, justification for caste-based discrimination, justification and support for untouchability and criticism of reservation (affirmative action) policy that questions the intellectual capability of these groups.

E.g. A Beemtaaa [3], Bairi Naresh from Telangana has insulted Ayyappa. Then he received treatment by local Police and some people. Only this treatment can fix who insults our Devi (PHM: 2) [4]

Religion: A span is annotated as targeting a member of a religious community if it calls for or justifies violence against them. Propounding or justifying conventional stereotypes associated with the members of such communities or using religious slurs will also have a harm potential greater than '0'. For example, Muslims being called terrorists or jihadis, Muslims and Christians being targeted for alleged forced conversions and Sikhs being called Khalistanis or secessionists.

E.g. Be it Kairana or Beerabhoom, wherever Muslims are in a majority, Hindus get killed or have to face exodus! #lack_of_unity!!" (PHM: 1)

Descent: For our specific case, descent encompasses all identities based on inherited status. This includes ethnicity, race, and place of origin (including linguistic or cultural minorities) of a victim (but not caste given its prevalence in the Indian context).

[3] A slur for Dalits

[4] For all the following examples, we have mentioned the Physical Harm Potential 'PHM' and/or Sexual Harm Potential 'SHM' of the text, as applicable

Spans supporting or justifying attacks based on places of origin are annotated under this category.

E.g. When the Habshis ruled Bengal I guess all the Bengali Hindus escaped by hiding inside Durga Ma's chuut[5] (PHM: 1)

Gender: This label annotates spans attacking gender minorities (LGBTQIA+ community) and women. Spans propounding or justifying conventional stereotypes or using gendered slurs are also marked with non-zero harm potential.

E.g. Bengali girls are greatest wokes of all. Even God Himself can't make them understand the truth of Love Jehad. No sympathy for them. (SHM: 1)

Political Ideology: Political violence including murder, lynching, thrashing, etc of opposing party members or people of different political ideologies happens regularly. Spans calling for or justifying violence, supporting discrimination or furthering stereotypes against the supporters of a political party or ideology are assigned harm potential greater than '0'. However, a criticism of the political ideologies, political leaders, policies, etc are assigned a '0' harm potential.

E.g. Isn't?? It is a dictator ruling! @INCIndia Slated slaves don't question their masters duplicate Gandhi! (PHM: 0)

2.3. When is the discourse happening?

This parameter indicates if a text is posted online about or during a major, public event or happening that might add to its harm potential. The harm potential of the content may increase when posted during or before such sensitive occasions and may lead to real-world violence in the form of mob lynchings and even ethnic cleansing. The major categories that we marked under this parameter are discussed below. This parameter is marked at the text/document level and the same text could take multiple labels. Since generally the dates of the events are already well-known, these labels could be mostly assigned automatically and could be seen as a grouping of multiple dates in a single category.

Riots: In general, violent public disorders are called riots. In India, in the past few years, hate speeches in social media have made a significant contribution to the amplification of violence during the riots. As such posts related to riots at the time of riots (or otherwise) are likely to have higher harm potential than otherwise.

Elections: Elections in India often see violence by supporters of rival political parties, and they are adopted in various themes such as communalism, terrorism allegations, anti-national, systemized threats and disruption of harmony.

[5]Slur for vagina

Pandemic: A huge number of potentially harmful online content were posted during COVID-19, and so, needs to be included as a separate category.

Extremist Attack: An extremist attack on state forces or the public may also lead to online hate against particular communities. The Pulwama suicide attack of February 2019 in India led to widespread hate speech and real-world violence against Kashmiri Muslims throughout India.

Festivals: Religious festivals have recently become flashpoints for communal violence with different sides accusing each other of attacking processions or interfering with rituals. Online hate and disinformation often spikes during these situations.

Group-Specific State Decisions: This context pertains to when the government introduces or implements legislation/decisions affecting a particular community. The government's decisions may be criticised or protested against by the community followed by online and offline attacks by the government's supporters. Recent examples in India have been the Citizenship Amendment Act, Farm Laws and the abrogation of Article 370.

Generic: These refer to the posts related to the incidents that are recurring in nature (like the previous factors) but generally do not have a fixed or pre-determined start or end time (viz. mob lynching on the suspicion of being child-lifters or those related to cow vigilantism in India).

Others: The posts that do not co-occur with any of the above-mentioned contextual factors - seemingly one-off incidents of hate and violence at no specific time - are marked as others.

2.4. How is it being said?

Since we focus on the harm potential of social media content, the methodology developed here is sensitive to the fact that the core objects of study are linguistic events themselves and so it is essential to model the textual features viz its lexical, syntactic and semantic properties that co-occur with the contextual features discussed in the earlier subsections. For the current project, we have defined a set of semantic features (specifically mood and modality) and lexical features (affective expressions) that are marked in the text. Since the other morphosyntactic features could be marked automatically using the earlier existing systems or are implicity learnt by modern transformers-based multilingual models, we have not marked those separately. The labels for this parameter are marked at the span level and generally, but not necessarily, overlap with the spans of the 'who' parameter.

There is an abstract link that can be sketched between the language that a speaker uses to convey harm vis-a-vis how that language is particularly structured to reflect the speaker's intentions and the speaker's evaluation of what they say as possible or necessary. The variation in the speaker's intention and their evaluation of what they say could have a significant impact on the harm potential of what is being said. These could be modelled using the linguistic categories of mood and modality. We discuss the different subtypes of these two categories used for annotation below -

Mood Type: The category of mood is a "grammatical reflection of the speaker's purpose in speaking" (Kroeger, 2005) or an indication of "what the speaker wants to do with the proposition" in a particular discourse context (Bybee, 1985). The grammatical form of the construction changes depending on whether the speaker wants to talk about a situation that has or will actualize in their perspective or whether they want to talk about an event that has not actualized. We annotate three broad kinds of mood types -

- **Realis Mood:** The Realis mood portrays situations as actualized, as having occurred or occurring, knowable through direct perception (Palmer, 2001). Indicative mood (that expresses actions that did take place, are taking place or will take place) is the canonical bearer of realis mood in a language.

  E.g. No Brahmin says that we need the daughters of Rajputs, Baniyas and Dalits! (PHM: 2)

- **Irrealis Mood:** The irrealis portrays situations as purely within the realm of thought, knowable only through imagination (Palmer, 2001). It is used to denote situations or actions that are not known to have happened. Modality-marked constructions, conditionals (that convey dependency of a situation on another situation), counterfactuals (that convey a conditional situation in an alternate reality i.e. a situation that cannot actualise because it is contrary to some fact in the actual world), optatives, hortatives and subjunctives (that express contrary to fact situations) are all grouped under the label of irrealis modality.

  E.g. if you let this fester a moment longer this too will become #ShaheenBagh. (PHM: 1)

- **Neither:** Imperatives (that are used to direct the behaviour of the addressee and get them to act a certain way), interrogatives (that are used to ask some information from the addressee), future-tense marked constructions (that indicate that some event will take place in future as compared to speech time) and negative constructions (that assert that something has not taken place) are marked as 'neither'.

  E.g. Don't expect anything better from an ex Hindu ricebag convert. (PHM: 1)

Illocutionary Mood: Illocutionary mood draws upon Austin and Searle's idea of illocution (what the speaker intends to do via his/her speech) act and encodes speaker intention (in illocution) as a category of mood. There is an expected correlation between a speech act and a sentence type since there is a language-independent tendency for certain illocutionary acts to be mapped onto specific grammatical forms. We have used the following subtypes of illocutionary mood -

- **Declaratives:** Declaratives can be either 'direct' (indicative mood) or 'indirect' (mainly using the rhetorical forms). A rhetorical question is an indirect speech act (a mismatch between the sentence type and the intended force) which involves the use of the interrogative form for some purpose other than asking questions (Kroeger, 2005). It can be used to indirectly assert something and thereby is an indirect declarative.

  E.g. Nagas eat dogs ,they are of Han decent. (PHM: 1)

- **Interrogatives:** These are primarily the question sentences.

- **Imperatives:** It could be in the form of a command, a request, advice, a plea, permission, an offer or an invitation.

  E.g. Boycott all sickular actors, politicians and social activists. (PHM: 1)

- **Admonitive:** Admonitives are the warnings that a speaker issues to the addressee(s).

  E.g. Once Hindus take up swords against rice bags, the bag distributors will quickly find their way to Africa. (PHM: 1)

- **Prohibitive:** Prohibitives curtail the addressee's actions and stop them from engaging with some situation or action.

  E.g Do not comment on personal matters, Mr Reporter. (PHM: 0)

- **Hortative:** It is used for softened commands or exhortations and so shares properties with imperatives (Puskás, 2018). It is often used with first-person inclusive reference ('let us...').

  E.g. Provided the ADC bill moved by the hill committee is within the constitution valley brothers should not take the hill brothers otherwise let us learn how to co exist peacefully let us get rid of bais attitude. (PHM: 0)

- **Optative:** Sentences in an optative mood express a wish or a desire of the speaker that some situation be brought about.

  E.g. If riots happen or there is khalistani activity (in any state and if links are traced back to

Punjab) then the central government will take over the control of Punjab. (PHM: 1)

- **Imprecative:** It indicates that the speaker wishes for an unfavourable proposition to come about.

  E.g. They must be beaten more mercilessly, what is done is not enough , they are not students but roadside goondas in disguise. (PHM: 2)

- **Exclamative:** These are primarily the exclamation sentences.

Modality: Modality can be viewed as speaker modification of a state of affairs concerning how the basic event/situation is construed by the speaker. Modalities come in two flavours - whether the speech event is 'possible' or 'necessary' given the particular set of conditions. We have used the following modalities for marking the text spans -

- **Epistemic:** It deals with "an estimation, typically but not necessarily by the speaker, of the chances or the likelihood that the state of affairs expressed in the clause applied in the world" (Nuyts and van der Auwera, 2016). It is marked for a sentence if that sentence concerns the speaker's knowing or believing that the state of affairs described in the sentence is possibly (possibility or dubitative) or certainly (necessity) true.

  E.g. #ShaheenBagh #AntiCAAprotest #JamiaViolence r all part of Gazwa-E-Hind. Such massive pan-India violent protest is impossible without foreign support. (PHM: 1)

- **Deontic:** Deontic modality can be defined as "an indication of the degree of moral desirability of the state of affairs expressed in the utterance, typically but not necessarily on behalf of the speaker" (Nuyts and van der Auwera, 2016). The conception of morality includes "societal norms as well as personal ethical criteria of the person responsible for deontic assessment" (Nuyts and van der Auwera, 2016). This modality involves an evaluation of the state of affairs that ranges from absolute moral necessity to moral acceptability.

  E.g. Secularism has no importance in the face of rigorous jihad and bigotry. Apart from this, in front of bio weapons, everyone is forced to be helpless, it has also been proved that religion did not work before them. Now is the time to think about them all. (PHM: 0)

- **Dynamic:** Dynamic modality is concerned with (a) an ability or a capacity ascribed to the participants of the action/ situation; (b) a need/necessity imposed on the participant by external circumstances that lie beyond their control; (c) a possibility/potential or necessity/inevitability inherent in the state of affairs described in the sentence and not related to the participants in that state of affairs.

  E.g. If 'fear of Bulldozer' is the new norm through which law n order can be maintained, then I whole heartedly welcome it. (PHM: 1)

- **Teleological:** Teleological modality concerns "what means are possible or necessary for achieving a particular goal" (Von Fintel, 2006).

  E.g. Only this treatment can fix who insults our Devi (PHM: 1)

In addition to these, the presence of affective expressions is marked if there is some word in the text that conveys the speaker's evaluative attitude or some emotional state towards some part of the information being conveyed by the sentence.

2.5. Why is it being said?

This parameter analyses the discursive role of the text placed within its context and checks for the reason or rationale behind posting the text. It is a direct induction of the 'discursive roles' by Kumar et al. (2022a) in this framework. We discuss the five categories and their relationship to the magnitude of harm potential here -

Attack: This label is used when any comment/post poses an attack on any individual or group based on any of their identities. Not all attacks are accompanied by a positive harm potential. For example, criticisms, which are not likely to trigger real-world harm against them are tagged as 'attack' with harm potential '0'.

E.g Let #ShaheenBagh rot.... All anti-nationals to rot - is their God watching or busy in hal.... ?? https://t.co/BBazCKFx9c (PHM: 2)

Defend: This label is used when any comment/post defends or counter-attacks a previous comment/post. Again not all instances of defend have '0' harm potential - in instances where the defense of the perceived 'victim' has the possibility of triggering real-world harm against the attacker, they are marked with non-zero harm potential.

Abet: This label is used when any comment/post lends support and/or encourages an aggressive act which has a harm potential.

E.g. If 'fear of Bulldozer' is the new norm through which law n order can be maintained, then I whole heartedly welcome it. (PHM: 1)

Instigate: This label is used when any comment/post encourages someone to perform an aggressive act. The comment itself may or may not be aggressive but the purpose must be to instigate an act that is potentially harmful in the real world. Instigation happens before the event and its purpose is to trigger or provoke a harmful act unlike abet which occurs during or after the harmful act and its purpose is to praise, support, and/or encourage that act as well as other such acts in the future.

E.g. @desimojito all the nations of the world are fighting #coronavirus and India is fighting the illiterates and morons of #peaceful_religion. All of them should be burnt alive; it will save the country from #biojihad #CoronaJihad along with #coronavirus (PHM: 3)

Counterspeech: Texts that diffuse the potentially harmful situation will be tagged as counterspeech. Just as influential speakers can make violence seem acceptable and necessary, they can also favourably influence discourse through counterspeech.

E.g. @ANI BUT Shaheen Bagh has not moved an INCH. Road is reopened, can you open your eyes NOW to look for the reason why it was shut? (PHM: 0)

3. Data Collection and Annotation

The framework discussed in Section 2 is developed over several stages and iterations. In order to test the reliability and validity of the framework, we collected a dataset from different social media platforms and annotated those using the framework. As a first step towards data collection, we focussed on a few incidents of physical harm (riots, lynchings etc.) that had a link to online disinformation and hate campaigns from 2016 – 2022 [6]. Finding relevant government-published data related to hate crimes was a challenge as the Indian government stopped collecting data on hate crimes in 2017. Therefore, we decided to use databases from non-governmental organisations like Documentation of the Oppressed (DOTO). This database consisted of a list of over 1,100 incidents of offline hate crimes and violence since 2016. Out of these, we sampled a little over 150 crimes since they had a link to social media discourse. We extracted social media data related to these incidents from different social media platforms viz Twitter, YouTube, Facebook, Telegram and WhatsApp. We also ensured that we collected data both from before and after the incident separately. This approach ensured that we got data that had links to offline harm incidents so they could be considered as potential triggers for the offline harm incident; at the same time, we also got data that might be triggers for a related post-event incident. We collected a total of over 574,000 datapoints in Hindi and English using this methodology (Figure 1) [7].

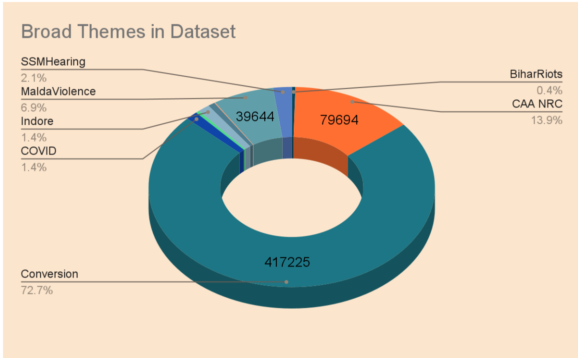

Figure 1: Broad Themes in Dataset

3.1. Inter-Annotator Agreement

Approximately 5 - 10 data points were selected from approximately 50 incidents for running the inter-annotator agreement experiments. Each of these was annotated by 3 annotators and Krippendorff's Kappa was calculated for the magnitude of harm potential. The first round of experiments with around 500 data points gave a rather dismal Kappa of 0.25. Following this, we made certain changes to the tagset such as merging different categories to reduce overlap across categories (for example, race, ethnicity and nationality under 'Who' were combined into a single category of 'Descent') and introducing new categories to better classify different kinds of categories (for example, several new categories under mood and modality were added for a better analysis). We also made changes to the annotation guidelines for clarity. These changes led to significant improvements in the alpha - the second round of experiments gave a final value of 0.53.

While the Kappa value still remained low, for a highly subjective task such as predicting the magnitude of harm potential as reasonably good. We started conducting focus group discussions to understand the reason behind disagreements. As it has been argued earlier as well (for example, Kumar et al. (2022b) and also the Perspectivist Data Manifesto [8]), most of these disagreements seemed reasonable. As such we decided not to push for further agreement - instead, we will be making the disaggregated annotations by different annotators publicly available.

[6]This time period was decided given the introduction of low-cost data and smartphones by Reliance Jio in 2016 which led to a manifold increase in per-capita data usage

[7]The complete dataset, along with the hate crimes they were associated with is available here - https://github.com/unrealtecellp/harmpot

[8]http://pdai.info/

3.2. Data Annotation

We have annotated a total dataset of approximately 2,000 data points (taking around 15 - 20 data points from over 100 incidents) to demonstrate the validity of the presented framework. We made use of an online app - LiFE App (Singh et al., 2022) - for data annotation since it allowed us to annotate the data simultaneously at the document and the span level. Each data point was annotated by 3 annotators working independently.

4. HarmPot and Other Frameworks: A Comparison

As we mentioned earlier, most of the existing frameworks attempt to only model hateful, aggressive, offensive (or one of the other similar flavours) speech but do not attempt to predict the potential of the text to trigger offline harm incidents. However, such language usage is expected to have some correlation with offline harm. Moreover, prior studies have also pointed out the need to flesh out the interrelationship between different frameworks so as to ensure interoperability and cross-use of datasets annotated with different hate speech frameworks (Poletto et al., 2021; Kumar et al., 2022b). In order to understand this relationship, we carried out a comparative study between our framework and three of the other popular frameworks. We took 500 texts annotated with each of these different frameworks, annotated those with the HarmPot framework and carried out a comparative study. The results of these are discussed in the following subsections.

4.1. HarmPot and HASOC

Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC) is a series of workshops/shared tasks that have been held since 2019 and that makes data available for Indo-European languages, marked with hate and offensive labels (Modha et al., 2019; Mandl et al., 2020, 2021; Amjad et al., 2021; Mandl et al., 2022). The first level of the schema distinguishes between Hate and Offensive (HOF) and Not hate and offensive (NOT). The second level of the schema classified HOF into three classes - Hate, Offensive and Profane. In the 2022 and 2023 editions, the second level of the schema was a multiclass annotation indicating Standalone Hate (hate by itself), Contextual Hate (hate in the context of its parent) and Non-hate (not hate by itself). In the 2023 edition, the task of identifying spans of hate was also introduced in the HateNorm track. We took a total of 500 texts each from the 2019 and 2023 editions of the task and annotated those using the HarmPot framework. to understand their interrelationship.

The study showed that most of the NOT texts had '0' harm potential but the vice-versa was not necessarily true. On the other hand, probably because of the broad definition of HOF (which includes texts with swear words and profanity), unexpectedly, at least some of the offensive texts carried '1' and even '0' harm potential. At the second level, the mapping becomes clearer as most of the 'Profane' texts are marked with '1' or '0' harm potential, most of the 'hate' texts were marked as '3' harm potential (or in some cases '2' as well). The offensive texts were marked with '2', '1' and even '0' harm potential. Table 1 illustrates the mapping between the two. In our comparison of the spans being marked using the HarmPot framework and those marked in the HateNorm task, we did not find many exact overlap of the spans selected in the two datasets. In most instances, the spans marked in the HateNorm task were a substring of those marked using the HarmPot framework and the total number of spans was also less in the HateNorm task - this is mainly because we do not mark hate spans, rather it's the identity spans that are marked in our framework. Level 2 of 2022 and 2023 editions, which mark whether it is contextual or standalone hate do not have a direct relationship to any of the levels in HarmPot - the main reason being that our definition of 'context' is more rooted in how it is defined in discourse analysis and pragmatics as different socio-cultural factors affecting the interpretation of the text (and not just parent / previous text in the thread).

4.2. HarmPot and OLID

The Offensive Language Identification Dataset (OLID) contains a collection of over 14k annotated English tweets using a three-level annotation framework. Level A distinguishes between Offensive and Non-offensive texts. At Level B offensive texts are further classified into targeted and untargeted insults. Level C categorises targeted insults into Individual, Group and Other targets (Zampieri et al., 2019a). It's a comparatively coarse-grained tagset but unlike the HASOC tagset, it addresses the two questions of 'who' is being targeted and 'magnitude' of the attack.

For level A, the results of the comparative study were similar to the HASOC dataset. The rest of the two levels in the OLID framework relate to the 'who' parameter in HarmPot but work at different axes - while OLID marks whether an individual or a group is being attacked, HarmPot looks at the specific identities irrespective of it being that of an individual or a group (more appropriately they represent the identity of an individual as a member of a group). All the texts with spans mentioning one of the identities and carrying a harm potential greater than '0' were marked as 'Targeted Insult'

HASOC Framework			
	Level A	Level B	Span
HarmPot Harm Potential	HOF $\epsilon\{0,1,2,3\}$ NOT $\epsilon\{0,1\}$	Offensive $\epsilon\{0,1,2\}$ Hate $\epsilon\{2,3\}$ Profane $\epsilon\{0,1\}$	–
HarmPot 'Who'	–	–	$HASOC \subset HarmPot$ HASOC $\epsilon\{1,2,3\}$
OLID Framework			
	Level A	Level B	Level C
HarmPot Harm Potential	OFF $\epsilon\{0,1,2,3\}$ NOT $\epsilon\{0,1\}$	TIN $\epsilon\{\{1,2,3\}\cup$ {Caste, Religion...} }	{IND,GRP} ϵ {Caste, Religion...}
ComMA Framework			
	Aggression	Aggression Intensity	Threat and Bias
HarmPot Harm Potential	OAG $\epsilon\{0,1,2,3\}$ CAG $\epsilon\{0,1,2\}$ NOT $\epsilon\{0,1\}$	{PTH, STH} $\epsilon\{1,2,3\}$ NtAG, CuAG $\epsilon\{0,1,2\}$	–
HarmPot 'Who'	–	–	$ComMA \subset HarmPot$ ComMA $\epsilon\{1,2,3\}$
HarmPot 'Why'	ComMA $\epsilon\{Attack, Defend, Instigate, Abet, Counterspeech\}$		

Table 1: Mapping of HarmPot and Other Frameworks

at OLID's Level B. Any text without mention of any of the identities was mostly marked 'Untargeted' - however, the magnitude of harm potential for such texts varied. This follows from the fact that the use of profane or unacceptable language may not necessarily trigger offline harm - one such instance could be friendly banter, which will have '0' harm potential. Texts marked as 'Individual' or 'Group' at Level C in the OLID dataset were marked for one of the identities in HarmPot. However, those marked as 'Other' were not marked for identities (although the number of such texts was very small in the overall dataset). Also, the mapping of these categories to harm potential is quite unpredictable. The tentative mapping of OLID framework to HarmPot is summarised in Table 1.

4.3. HarmPot and ComMA

Lastly we also conducted a comparative study between the HarmPot and ComMA framework (Kumar et al., 2018b, 2022b). The top level of the framework distinguishes between overtly, covertly and non-aggressive texts. At the second level, the aggression intensity of the aggressive texts - physical threat, sexual threat, non-threatening aggression and curse/abuse - are marked. Parallel to this, bias and threats of four kinds - religious, caste/class, gender and racial/ethnic - are marked. It also marks the discursive roles - attack, defend, counterspeech, abet and instigate and gaslighting - of the text. These discursive roles are already borrowed and incorporated in the HarmPot framework. Besides this, there are several parallels between the ComMA and HarmPot frameworks and also since social or physical 'harm' is inherent to the idea of aggression, we expected a good mapping between the notion of verbal aggression and the harm potential of a text.

The study showed that most of the non-aggressive texts (NAG) are at Level 0 but the vice-versa is not necessarily true. Also, most of the 'covertly aggressive' (CAG) texts are categorised with level '1' harm potential. At the second level, physical and sexual threats were mostly marked as having '2' or '3' harm potential while non-threatening aggression is mostly marked as '1'. As in the earlier instances, some of the curse/abuse texts were also marked with '0' harm potential. At the level of threat and bias, even though religious, caste and gender bias have direct parallels in HarmPot since we are marking all mentions of these identities and not just biased or threatening ones (unlike the ComMA dataset), the instances of such spans were higher in our case. However, threats generally carried a harm potential of '2' or '3' while bias carried a harm potential of '1' or '2'. Some of the comments marked as non-biased in the ComMA dataset also carried a harm potential of '1' or even '2'. Moreover, the ComMA dataset marked the biases at the document level while HarmPot marks these at the span level.

5. Conclusion

In this paper, we have presented a new framework that could be used for annotating social media text with its potential for triggering offline harm. The framework incorporates contextual information such as the identity of the victim (as mentioned/referred to in the text), the broad socio-political situation in which the post is situated and the role that the text assumes in the discourse. We have also proposed using mood and modality as relevant categories for marking the speaker's intention, intended goal and their own evaluation of whether what they are saying is 'necessary' or 'possible'. These semantic categories have been rarely utilised in NLP but they could prove to be extremely useful in the identification of subjective phenomena like harm potential. We have annotated a total dataset of 4,000 texts - 2,000 related to the possible triggers of offline harm incidents and another 2,000 from datasets available for aggressive and hateful language identification. We use these 2,000 to carry out a comparative study of HarmPot with three popular frameworks and establish that a one-to-one mapping between these frameworks is not possible mainly because HarmPot does not mark hateful language; rather offline harm potential of the text. It shows some correlation between hateful language and its harm potential but neither entails the other. We are currently annotating some more data and also conducting experiments for the automatic identification of harm potential to understand the practical efficacy of the framework.

6. Ethical Considerations

The nature of the task - the creation of datasets with high harm potential and its annotation - in itself raises several ethical issues of bias and psychological impact on the annotators working with the data. In order to reduce the impact of working with such data, we took 3 steps - (a) a 'maximum' limit of 200 texts per week was set for the annotators - the annotators were barred from going through more than this number of texts in a week; (b) we had made arrangements for psychological counselling of the annotators working on the data; (c) a compulsory weekly 'venting out' meeting was organised to enable annotators to talk to each other and other members of the project that allowed them to talk about, discuss and (hopefully) figure out the ridiculousness of the data that they were going through. We made a very conscious decision not to use crowdsourcing or even third-party annotators for data annotation and collection to ensure that these mechanisms are put in place.

In order to minimise the bias in the annotations and also make different perspectives on the data public, we have decided to release the disaggregated dataset with the annotations of all the annotators (with their disagreements). We were very conscious not to push for an agreement where it was not possible. Moreover, our in-house annotators were from mutually distinct socio-political, religious, cultural, and educational backgrounds, providing an innate cancelling out of any one type of bias overpowering the data analysis and interpretation - we have tried to annotate the data in such a way as to reflect different perspectives on the data (and not propound a single, homogeneous view).

7. Limitations

One of the primary limitations of the framework and the dataset is the lack of multimodal information being included in it. A large number of hateful and abusive language used on social media, with a high potential for harm, is expected to be accompanied by visuals including images and video. We are working on expanding the dataset to include multimodal data and see how well the framework adapts to that and also what kind of modifications would be needed for handling those cases. The second limitation is the pipeline-based workflow that the framework enforces, which has a greater chance of error propagation - if, for example, the system makes an error in recognising mood and modality, that might ultimately lead to an error in the prediction of harm potential itself. This is a general limitation of the hierarchical frameworks.

8. Bibliographical References

1969. *Gopal Vinayak Godse vs. The Union of India and Ors_1969*.

A. Agha. 2007. *Language and Social Relations*. Cambridge: Cambridge University Press.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *Proceedings of the international conference of the Italian association for artificial intelligence*, pages 588–603.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. In *Computer Science and Information Technology 2019*, pages 83–100.

Md Sayeed Al-Zaman. 2019. Digital disinformation and communalism in bangladesh. *China Media Research*, 15(2):68–76.

Md. Sayeed Al-Zaman. 2021. Covid-19-related social media fake news in india. *Journalism and Media*, 2(1):100–114.

Azalden Alakrot, Liam Murray, and Nikola S. Nikolov. 2018. Towards accurate detection of offensive language in online communication in arabic. *Procedia Computer Science*, 142:315–320. Arabic Computational Linguistics.

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *Proceedings of the 2018 IEEE/ACM international conference on advances in social networks analysis and mining*, page 69–76.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *Proceedings of 2017 international conference on advanced computer science and information systems (ICACSIS), IEEE*.

Maaz Amjad, Alisa Zhila, Grigori Sidorov, Andrey Labunets, Sabur Butt, Hamza Imam Amjad, Oxana Vitman, and Alexander Gelbukh. 2021. Overview of abusive and threatening language detection in urdu at fire 2021. In *Proceedings of the 12th forum for information retrieval evaluation (FIRE)*, pages 744–762, New York, USA. Association for Computing Machinery.

Amena Akter Aporna, Istinub Azad, Nibraj Safwan Amlan, Md Humaion Kabir Mehedi, Mohammed Julfikar Ali Mahbub, and Annajiat Alim Rasel. 2022. Classifying offensive speech of bangla text and analysis using explainable ai. In *Advances in Computing and Data Sciences*, pages 133–144, Cham. Springer International Publishing.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets.

Supurna Banerjee and Nandini Ghosh. 2018. *Caste and Gender in Contemporary India: Power, Privilege and Politics*. Taylor & Francis.

Nayan Banik and Md. Hasan Hafizur Rahman. 2019. Toxicity detection on bengali social media comments using supervised models. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)*, pages 1–5.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of Hindi-English code-mixed social media text for hate speech detection. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, pages 36–41, New Orleans, Louisiana, USA. Association for Computational Linguistics.

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

Joan L Bybee. 1985. *Morphology: A Study of the Relation between Meaning and Form*. John Benjamins Publishing Company.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

Puja Chakraborty and Md. Hanif Seddiqui. 2019. Threat and abusive language detection on social media in bengali language. In *2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT)*, pages 1–6.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jonathan Culpeper. 2011. *Impoliteness: Using Language to Cause Offence*. Cambridge: Cambridge University Press.

Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md. Nur Hossain. 2021. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591.

Anne Boyle David. 2015. *Descriptive Grammar of Bangla:*. DE GRUYTER.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the eleventh international conference on web and social media, AAAI*, page 512–515.

R. de Pelle and V. P. Moreira. 2016. Offensive comments in the brazilian web: A dataset and baseline results. In *Proceedings of the fifth Brazilian workshop on social network analysis and mining (BraSNAM 2016)*, page 510–519.

Kaushik Deka. 2019. Fake news: What you read in election season.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian conference on cybersecurity (ITASEC17), CEUR.org*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. pages 29–30.

Prabhas k Dutta. 2018. 16 lynchings in 2 months. is social media the new serial killer?

V. D'Orazio, M. Kenwick, M. Lane, G. Palmer, and Reitter D. 2016. Crowdsourcing the measurement of interstate conflict. *PloS one*, 11(6):e0156527.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Gino Eelen. 2001. *A Critique of Politeness Theories*, volume 1. Routledge.

Shahnoor C. Eshan and Mohammad S. Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6.

Johan Fernquist, Oskar Lindholm, Lisa Kaati, and Nazar Akrami. 2019. A study on the feasibility to detect hate speech in swedish. In *2019 IEEE international conference on big data (Big Data), 2019, IEEE*, pages 4724–4729.

Paula Fortuna. 2017. Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Peggy Froerer. 2019. *Religious division and social conflict: the emergence of Hindu nationalism in rural India*. Routledge.

Andrew M Guess, Michael Lerner, Benjamin Lyons, Jacob M Montgomery, Brendan Nyhan, Jason Reifler, and Neelanjan Sircar. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences*, 117(27):15536–15545.

Hatem Haddad, Hala Mulki, and Asma Oueslati. 2019. T-hsab: A tunisian hate speech and abusive dataset. In *7th international conference on Arabic language processing*, pages 251–263.

Hugo Hammer. 2017. Automatic detection of hateful comments in online discussion. In *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, volume 188, pages 164–173.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Md Gulzar Hussain and Tamim Al Mahmud. 2019. A technique for perceiving abusive bangla comments. *GREEN UNIVERSITY OF BANGLADESH JOURNAL OF SCIENCE AND ENGINEERING*, 04(01).

Alvi Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *18th IEEE international conference on machine learning and applications, ICMLA 2019*, pages 555–560, Boca Raton, FL, USA.

Tanvirul Islam, Nadim Ahmed, and Subhenur Latif. 2021. An evolutionary approach to comparative analysis of detecting bangla abusive text. *Bulletin of Electrical Engineering and Informatics*, 10:2163–2169.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666, Florence, Italy. Association for Computational Linguistics.

Kaggle. 2020. Jigsaw multilingual toxic comment classification.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Md. Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Sagor Sarker, Mehadi Hasan Menon, Kabir Hossain, Md. Azam Hossain, and Stefan Decker. 2021. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Md. Rezaul Karim, Bharathi Raja Chakravarthi, John P. McCrae, and Michael Cochez. 2020. Classification benchmarks for under-resourced bengali language based on multichannel convolutional-lstm network. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 390–399.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4.

Paul R Kroeger. 2005. *Analyzing grammar: An introduction*. Cambridge University Press.

Ritesh Kumar, Bornini Lahiri, and Atul Ojha. 2021. Aggressive and offensive language identification in hindi, bangla, and english: A comparative study. *SN Computer Science*, 2.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2018a. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ritesh Kumar, Atul Kr. Ojha, Shervin Malmasi, and Marcos Zampieri. 2020. Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, Akanksha Bansal, and Atul Kr. Ojha. 2022a. The ComMA dataset v0.2: Annotating aggression and bias in multilingual social media discourse. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4149–4161, Marseille, France. European Language Resources Association.

Ritesh Kumar, Shyam Ratan, Siddharth Singh, Enakshi Nandi, Laishram Niranjana Devi, Akash Bhagat, Yogesh Dawer, bornini lahiri, Akanksha Bansal, and Atul Kr. Ojha. 2022b. The comma dataset v0.2: Annotating aggression and bias in multilingual social media discourse. In *Proceedings of the Language Resources and Evaluation Conference*, pages 4149–4161, Marseille, France. European Language Resources Association.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018b. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Shervin Malmasi and Marcos Zampieri. 2017. Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria. INCOMA Ltd.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Shrey Satapara, Tithi Patel, and Hiren Madhu. 2022. Overview of the hasoc subtrack at fire 2022: Identification of conversational hate-speech in hindi-english code-mixed and german language. In *Proceedings of the 13th forum for information retrieval evaluation (FIRE)*, pages 475–488, New York, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Jaiswal, D Nandini, D Patel, P Majumder, and J Schäfer. 2020. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th forum for information retrieval evaluation (FIRE)*, page 29–32, New York, USA. Association for Computing Machinery.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the hasoc subtrack at fire 2021: Hatespeech and offensive content identification in english and indo-aryan languages. In *Proceedings of the 12th forum for information retrieval evaluation (FIRE)*, pages 1–19, New York, USA. Association for Computing Machinery.

Ricardo Martins, Marco Gomes, João Almeida, Paulo Novais, and Pedro Henriques. 2018. Hate speech classification in social media using emotional analysis. In *Proceedings of the 2018 Brazilian conference on intelligent systems, BRACIS 2018*, pages 61–66.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Omar Shahabudin McDoom. 2014. Predicting violence within genocide: A model of elite competition and ethnic segregation from rwanda. *Political Geography*, 42:34–45.

Sandip Modha, Thomas Mandl, Prasenjit Majumder, and Daksh Patel. 2019. Overview of the hasoc track at fire 2020: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 10th forum for information retrieval evaluation (FIRE)*, pages 167–190, New York, USA. Association for Computing Machinery.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Rahul Mukherjee. 2020. Mobile witnessing on whatsapp: Vigilante virality and the anatomy of mob lynching. *South Asian popular culture*, 18(1):79–101.

Gabriel Nascimento, Flavio Carvalho, Alexandre Cunha, Carlos Viana, and Gustavo Paiva Guedes. 2019. Hate speech detection using brazilian imageboards. In *Proceedings of the 25th Brazillian symposium on multimedia and the web, WebMedia 2019*, pages 325–328.

Online Desk NewIndianXpress. 2018. "no regret for hacking afrazul to death": Shambhulal regar releases another video, this time from rajasthan prison.

Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web (WWW'16)*, pages 145–153. International World Wide Web Conferences Steering Committee.

Jan Nuyts and Johan van der Auwera. 2016. *The Oxford Handbook of Modality and Mood*. Oxford University Press.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Frank Robert Palmer. 2001. *Mood and modality*. Cambridge university press.

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

Genoveva Puskás. 2018. To wish or not to wish: Modality and (metalinguistic) negation. *Glossa: a journal of general linguistics*, 3(1).

Tharindu Ranasinghe and Marcos Zampieri. 2021. An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, 12(8).

Sumaiya Salim Ritu, Joysurya Mondal, Md. Moinu Mia, and Ahmed Al Marouf. 2021. Bangla abusive language detection using machine learning on radio message gateway. In *2021 6th International Conference on Communication and Electronics Systems (ICCES)*, pages 1725–1729.

Nauros Romim, Mosahed Ahmed, Md Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2021a. HS-BAN: A benchmark dataset of social media comments for hate speech detection in bangla. *CoRR*, abs/2112.01902.

Nauros Romim, Mosahed Ahmed, Md. Saiful Islam, Arnab Sen Sharma, Hriteshwar Talukder, and Mohammad Ruhul Amin. 2022. Bd-shs: A benchmark dataset for learning to detect online bangla hate speech in different social contexts.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md. Saiful Islam. 2021b. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468, Singapore. Springer Singapore.

Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Marcos Zampieri, and Preslav Nakov. 2021. Solid: A large-scale semi-supervised dataset for offensive language identification. pages 915–928.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *NLP4CMC III: 3rd workshop on natural language processing for computer-mediated communication.*

Devanshu Sajlan. 2021. Hate speech against dalits on social media: Would a penny sparrow be prosecuted in india for online hate speech? *CASTE / A Global Journal on Social Exclusion*, 2(1):77–96.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 2798–2895, Miyazaki, Japan. European Language Resources Association (ELRA).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Salim Sazzed. 2021a. Abusive content detection in transliterated Bengali-English social media corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 125–130, Online. Association for Computational Linguistics.

Salim Sazzed. 2021b. Identifying vulgarity in bengali social media textual content. *PeerJ Comput Sci.*

Johannes Schäfer and Ben Burtenshaw. 2019. Offence in dialogues: A corpus-based study. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1085–1093, Varna, Bulgaria. INCOMA Ltd.

Natalie Schluter and Željko Agić. 2017. Empirically sampling Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 117–122, Gothenburg, Sweden. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Zeeshan Shaikh. 2020. Palghar lynching: A recap of what happened.

Omar Sharif and Mohammed Moshiul Hoque. 2021. Identification and classification of textual aggression in social media: Resource creation and evaluation. In *Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pages 9–20, Cham. Springer International Publishing.

Omar Sharif and Mohammed Moshiul Hoque. 2022. Tackling cyber-aggression: Identification and fine-grained categorization of aggressive texts on social media using weighted ensemble of transformers. *Neurocomputing*, 490:462–481.

Bhushan Sharma and KA Geetha. 2021. Casteing gender: Intersectional oppression of dalit women. *Journal of International Women's Studies*, 22(10):0–7.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of NLP crowdsourcing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3758–3769.

Alexandra A Siegel and Vivienne Badaan. 2020. # no2sectarianism: Experimental approaches to reducing sectarian hate speech online. *American Political Science Review*, 114(3):837–855.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Siddharth Singh, Ritesh Kumar, Shyam Ratan, and Sonal Sinha. 2022. Towards a unified tool for the management of data and technologies in field linguistics and computational linguistics - LiFE. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 90–94, Marseille, France. European Language Resources Association.

Scroll Staff. 2020. Una case: Victims ask president to deport them to country where they will not face discrimination.

Richard Stansfield and Kirk R Williams. 2014. Predicting family violence recidivism using the dvsi-r: Integrating survival analysis and perpetrator characteristics. *Criminal Justice and Behavior*, 41(2):163–180.

Josef Steinberger, Tomáš Brychcín, Tomáš Hercig, and Peter Krejzl. 2017. Cross-lingual flames detection in news discussions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 694–700, Varna, Bulgaria. INCOMA Ltd.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Bertram Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15:e0243300.

Bertram Vidgen and Taha Yasseri. 2020. Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17:66–78.

Kai Von Fintel. 2006. Modality and language.

Shuohuan Wang, Jiaxiang Liu, Xuan Ouyang, and Yu Sun. 2020. Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455, Barcelona (online). International Committee for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142. Association for Computational Linguistics (ACL).

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Sebastian Weingartner and Lea Stahel. 2019. Online aggression from a sociological perspective: An integrative view on determinants and possible countermeasures. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 181–187, Florence, Italy. Association for Computational Linguistics.

Eran Zaidise, Daphna Canetti-Nisim, and Ami Pedahzur. 2007. Politics of god or politics of man? the role of religion and deprivation in predicting support for political violence in israel. *Political Studies*, 55(3):499–521.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A. Data Statement

A.1. Header

Dataset Title: HarmPot Dataset

Dataset Curator(s):

- **Ojaswee Bhalla**, Council for Strategic and Defense Research, New Delhi
- **Madhu Vanthi**, Council for Strategic and Defense Research, New Delhi
- **Shehlat Maknoon Wani**, Council for Strategic and Defense Research, New Delhi

Dataset Version: Version 0.1, September 30, 2023

Dataset Citation: NA

Data Statement Authors:

- **Ritesh Kumar**, Council for Strategic and Defense Research, New Delhi

Data Statement Citation and DOI: NA

Links to versions of this data statement in other languages: NA

A.2. Executive Summary

The HarmPot dataset is collected to develop a computational system that could automatically identify the potential of social media content to trigger offline incidents of harm and violence (such as lynching, murder, etc). The dataset currently contains approximately 2,000 manually annotated data instances in Hindi and (Indian) English and approximately 20,000 data instances automatically mapped from the ComMA dataset in Bangla, Hindi, Meitei and (Indian) English. The automatically mapped instances are marked only for a few parameters specified in the HarmPot framework.

A.3. Curation Rationale

This dataset was created with the ultimate goal of developing a system that is able to identify and mark the potential of social media content to cause real-world, offline harm (such as murder, rape, etc). "Harm Potential" (HarmPot) could be defined as the potential for an online public post to cause offline, real-world physical harm (i.e., violence). Targeted real-world violence is often spurred by a web of triggers, often combining several online tactics and pre-existing intersectional fissures in the social milieu. The HarmPot framework is designed to answer the following set of questions for a given text -

Who is being talked to, when, how, why and all this results in what magnitude of harm potential for the addressee?

Given this task, we followed an event-driven methodology to collect the data - we first identified the events where offline harm events have been triggered by some social media content (using a database of harm events in India, called the DOTO database) and then automatically crawled the data related to that event from different social media platforms viz Twitter, Facebook, YouTube, Telegram and WhatsApp. The dataset has been manually annotated by multiple annotators in order to identify the linguistic and pragmatic features that characterize harm potential and is meant to answer the following questions -

- **Who is being talked to**: This marks the category of the different identities mentioned or referred to in the data instance, viz., caste, religion, descent, gender and political ideology.
- **When**: It marks the broader event within which the text is situated and includes events such as riots, elections, pandemic, extremist attack, festivals, group specific state decisions, generic and others.
- **How**: It marks the linguistic devices being used by the speakers and includes four broad

categories of mood type, illocutionary mood, modality and presence of affective expressions.

- **Why**: It marks the discursive role of the specific data instance viz., attack, defend, abet, instigate and counterspeech.
- **Magnitude of Harm Potential**: Physical and Sexual Harm potential for each identity ('who') and that of overall document is marked on a scale of 0 (no harm potential) to 3 (harm potential in all contexts).

Depending on the data source, a single post, tweet or comment is taken as one document for annotation. The annotations of 'who' and 'how' include marking the span of texts with the appropriate categories, while other aspects are marked at the document level. In order to represent this complex structure of the annotations, the dataset and its annotations has been stored in a MongoDB database and is accessed and edited using a custom WebApp build for this purpose. The complete dataset is made publicly available in a JSON format. Some specific parts of the dataset has also been made available in CSV and other tabular formats.

A.4. Documentation for Source Datasets

A part of the current dataset - approximately 20,000 data instances - have been automatically mapped from the ComMA dataset, available here `https://github.com/unrealtecellp/ComMA`. These instances are marked only for two aspects - (a) the overall harm potential; (b) 'who' is being addressed. Moreover, there is no span-level annotation in these instances.

A.5. Language Varieties

The languages included in this dataset, listed with their respective BCP-47 language tags, include:

- **mni-IN: Meitei** as spoken by the Meitei community in Manipur, India.
- **bn-IN and bn-BD: Bangla** (and its varieties) as spoken in India and Bangladesh.
- **he-IN: Hindi** (and its varieties) as spoken in various parts of India.
- **en-IN: English** (and its varieties) as spoken in India, otherwise known as Indian English.

Since this dataset has been exclusively collected from online sources, the users writing the comments are assumed to be multilingual and may be based in any part of the world, not just in the places these languages are primarily spoken in. However, the language varieties used in the dataset are primarily those mentioned in the list above.

A.6. Speaker Demographic

This dataset has been sourced exclusively from the internet, hence the speaker demographic of the dataset cannot be identified beyond the language they speak. It is assumed that the speakers could be of any age, gender, sexual orientation, educational background, nationality, caste, class, religion, race, tribe, or ethnicity.

The speakers are probably multilingual as well, with the language they post in being one of the many they would know or be fluent in. It is a safe assumption to make that many of these comments are made by Indians (specifically people who have Meitei, Bangla, and Hindi as their first or primary language) and Bangladeshis given the nature and reach of the topics selected, but this assumption is not backed by any data or statistical findings.

A.7. Annotator Demographic

The annotation scheme and guidelines for this dataset has been developed by Dr Ritesh Kumar, the principal investigator of the HarmPot Project, fellow and lead of the Division of Artifical Intelligence and Linguistics at the Council for Strategic and Defense Research and a co-founder and CEO of the UnReaL-TecE LLP. He was assisted by the research associates of the project and the annotators of this dataset, who have been listed below. Further, these annotators have manually identified the appropriate events and content to work on, crawled the data, and then annotated and analysed the processed data in their respective languages.

- A 32-year-old Punjabi Hindu woman from Chandigarh. She has a PhD in Linguistics, speaks Pubnjabj, Hindi, and English, and her ideological leanings are non-right.
- A 29-year-old Kashmiri Muslim man. He was a journalist before joining the project. His research looks at technology policy and digital governance, and he has extensively covered the Kashmir conflict, internet shutdowns and online hate speech in India.
- A 30-year-old Tamil Hindu woman. She has an UG in computer science and a postgraduate degree in geopolitics and international relations before joining the project.

A.8. Speech Situation and Text Characteristics

This dataset comprises of online comments written by users of various social media platforms. The comments collected range from 2016 to 2022, and form part of an extensive and intensive social media discourse.

- **Time and place of linguistic activity** - Online
- **Date(s) of data collection** - August 2022 - September 2023
- **Modality** - Written
- **Scripted/edited vs. spontaneous** - Spontaneous
- **Synchronous vs. asynchronous interaction** - Asynchronous (online comments)
- **Speakers' intended audience** - Other users of the respective social media platforms and channels
- **Genre** - Social media
- **Topic** - Content related to the incidents of offline harm that were linked to activity on social media. It includes incidents such as murder, lynching, etc
- **Non-linguistic context** - Incidents of offline harm.
- **Additional details about the cultural context** - The data instances are mainly posted by the users of social media who live in a sociopolitically polarised eneivronment.

A.9. Preprocessing and Data Formatting

The preprocessing of the raw data involves deleting all duplicates of a data instance, deleting data instances with urls and texts with less than three words, and removing data instances which occur in languages apart from Hindi and English.

A.10. Capture Quality

The data is collected using an event-driven methodology, which means that we have first identified the incidents of offline harm and then crawled data related to those incidents. This implies that the dataset represents data instances related to only these incidents. This also implies that we do not have a class-balanced dataset. The incidents themselves are not balanced in the sense that certain kinds of incidents (for example, communal incidents) are much higher in number than others (for example, incidents related to the descent).

A.11. Limitations

Adequate representation of different kinds of incidents and sufficient representation of all the tags could not be ensured in this version of the data.

A.12. Metadata

The relevant links to the metadata for this dataset have been provided below:

License: *CC BY-NC-SA 4.0*

Annotation Guidelines: *https://docs.google.com/document/d/1x-ZMMWlw5ajtykCMKY1SySaadNsywwdrMrpMPVPDdN4/edit?usp=sharing*

Annotation Process: *Manual annotation and automatic mapping from an existing dataset*

Dataset Quality Metrics: *Krippendorff's Alpha for IAA. We are also releasing a disaggregated dataset available.*

Errata: *NA*

A.13. Disclosures and Ethical Review

This dataset has been partially funded by Logically Inc and UnReaL-TecE LLP.