Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4149–4161
Marseille, 20-25 June 2022

The ComMA Dataset V0.2: Annotating Aggression and Bias in Multilingual Social Media Discourse

Ritesh Kumar[1], Shyam Ratan[1], Siddharth Singh[1], Enakshi Nandi[2], Laishram Niranjana Devi[2], Akash Bhagat[3], Yogesh Dawer[1] Bornini Lahiri[3], Akanksha Bansal[2], Atul Kr. Ojha[4,2]
[1]Dr Bhimrao Ambedkar University, [2]Panlingua Language Processing LLP, [3]Indian Institute of Technology-Kharagpur
[4]DSI, National University of Ireland Galway
New Delhi, Agra, Kharagpur, Galway
comma.kmi@gmail.com

Abstract

In this paper, we discuss the development of a multilingual dataset annotated with a hierarchical, fine-grained tagset marking different types of aggression and the "context" in which they occur. The context, here, is defined by the conversational thread in which a specific comment occurs and also the "type" of discursive role that the comment is performing with respect to the previous comment. The initial dataset, being discussed here consists of a total 59,152 annotated comments in four languages - Meitei, Bangla, Hindi, and Indian English - collected from various social media platforms such as YouTube, Facebook, Twitter and Telegram. As is usual on social media websites, a large number of these comments are multilingual, mostly code-mixed with English. The paper gives a detailed description of the tagset being used for annotation and also the process of developing a multi-label, fine-grained tagset that has been used for marking comments with aggression and bias of various kinds including sexism (called gender bias in the tagset), religious intolerance (called communal bias in the tagset), class/caste bias and ethnic/racial bias. We also define and discuss the tags that have been used for marking the different discursive role being performed through the comments, such as attack, defend, etc. Finally we present a basic statistical analysis of the dataset. The dataset is being incrementally made publicly available on the project website .

Keywords: aggression, bias, Meitei, Bangla, Hindi, Tagset

1. Introduction

Aggression, bias, polarisation and hate are now commonplace phenomena on all kinds of social media platforms. And so are the research efforts to automatically identify these and process them in some meaningful way so as to reduce their harmful impact on social communication and society, in general. Two recent systematic surveys (Vidgen and Derczynski, 2020) and (Poletto et al., 2021) have shown that over 60 datasets annotated with different aspects of hateful and abusive speech have been developed in various languages of the world in just around half a decade.

The wide range of interrelated phenomena that has been used for annotating the datasets include broad dimensions such as 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., 2018); (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), (Zampieri et al., 2019b); (Zampieri et al., 2020)), 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, 2017); (Schmidt and Wiegand, 2017)), 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), homophobia (Akhtar et al., 2019), etc. Some of the 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). In fact, (Jurgens et al., 2019) has rightly recommended for the abusive language researchers to broaden their scope so as to also study more subtle (such as condescension) as well as more serious forms of abuse (such as doxxing) and also posit the research within the broader framework of justice.

In this paper, we discuss the development of dataset annotated with different kinds of aggressive language and bias in four languages viz. Meitei, Bangla, Hindi and English. We have used a modified version of the tagset discussed in (Kumar et al., 2018). These modifications in the tagset are made considering the need to reduce the complexity of the earlier tagset and also in accordance with the needs of comprehensively annotating the available dataset, based on the observation of trends and patterns of discursive behaviour in each of these languages. For instance, the category of caste/class bias was included upon observing the nature of the comments in Bangla that were directed against the beggar-turned-singer Ranu Mondol, whose overnight success and subsequent change in attitude led to a deluge of comments directed not just at her gender but also at her lower class and caste identities. Similarly, the category for ethnic/racial bias was included upon observing the online animosity in the interactions between the Meitei and Kuki tribes in the Meitei data.

We will discuss these decisions in more detail in the following sections of the paper.

2. Building and annotating an extensive dataset: Methodology

(Kumar et al., 2018) proposed a detailed classification and tagset for marking aggression and bias, which included the distinction between overt and covert aggression as well as the target-based classification such as misogynistic, communal, geographical, sexual, etc. While the tagset was quite detailed, it posed two problems - (a) there were too many tags to be comprehended and classified manually by annotators with an appropriate degree of precision; (b) it clubbed together non-mutually exclusive categories at the same level (for example, curse/abuse and non-threatening aggression were at the same level) - while this was handled somewhat by allowing for multi-label annotation, in principle, it was a non-rigorous scheme and posed several problems at the time of annotation.

In order to present the classification in a more principled way and also to practically ease the task of annotation, the scheme has been restructured and also extended to include those aspects which were left out in the earlier version of the tagset. This modified tagset now includes a gradation of the intensity of aggression in a comment (physical threat, sexual threat, non-threatening aggression, curse/abuse) at every level, the discursive role of the given aggressive comment also include three new/modified roles (counterspeech, abet/instigate and gaslighting), and two new categories are added to tag biased speech targeted at individuals or social groups on the basis of their caste/class and ethnic/racial identities given in Table 1.

This dataset was built over multiple stages of the project, in concordance with the stages of development of the tagset. In the initial stages, the data was collected from online news websites and were tagged on the basis of two parameters: aggression and misogyny. In the second stage, this tagset was expanded to include communal bias, and the term "misogyny" was altered to "gender bias" so as to take into account bias directed at all genders (including transgenders) as well as people with different sexual orientations. The data was collected from the popular social media apps such as YouTube, Telegram, Facebook and Twitter and included data in English, Hindi, Bangla, and Meitei.

The data, collected from YouTube, Twitter and Telegram, is sampled based on identifying specific videos and channels that attracted a great deal of hate speech and aggressive speech in the comments section. A large number of these comments were directed at women and minority religious groups, especially Muslims. The process of sampling these videos and channels involved typing keywords and hashtags related to controversial socio-political, religious, or cultural events in the recent and not-so-recent past in any of the four languages mentioned above. The keywords used for searching the

	Aggression	
Code	**Aggression Level**	**TAG**
1.1	Overtly Aggressive	OAG
1.2	Covertly Aggressive	CAG
1.3	Non Aggressive	NAG
1.4	Unclear	UNC
	Aggression Intensity Level	
Code	**Attribute**	**TAG**
1A.1.1	Physical Threat	PTH
1A.1.2	Sexual Threat	STH
1A.1.3	Non-threatening Aggression	NtAG
1A.1.4	Curse/Abuse Aggression	CuAG
	Discursive Role	
Code	**Attribute**	**TAG**
1B.1.1	Attack	ATK
1B.1.2	Defend	DFN
1B.1.3	Counterspeech	CNS
1B.1.4	Abet and Instigate	AIN
1B.1.5	Gaslighting	GSL
	The Gender Bias	
Code	**Attribute**	**TAG**
2.1	Gendered Comments	GEN
2.2	Gendered Threats	GENT
2.3	Non-Gendered Comments	NGEN
2.4	Unclear	UNC
	The Religious Bias	
Code	**Attribute**	**TAG**
3.1	Communal Comments	COM
3.2	Communal Threats	COMT
3.3	Non-Communal Comments	NCOM
3.4	Unclear	UNC
	The Caste / Class Bias	
Code	**Attribute**	**TAG**
4.1	Casteist/Classist Comments	CAS
4.2	Casteist/Classist Threats	CAST
4.3	Non-Casteist/Classist Comments	NCAS
4.4	Unclear	UNC
	The Ethnicity / Racial Bias	
Code	**Attribute**	**TAG**
5.1	Ethnic/Racial Comments	ETH
5.2	Ethnic/Racial Threats	ETHT
5.3	Non-Ethnic/Racial Comments	NETH
5.4	Unclear	UNC

Table 1: The ComMA Project Tagset

videos on YouTube and the amount of raw data (number of comments) they yielded are included in Table 2. This dataset was then manually annotated by minimally two, and in most cases, three, annotators using a method named the 'Discursive Method of Annotation'. It has been demonstrated in several pragmatic and social science studies that the judgment of speakers on socio-pragmatic phenomena like aggression or bias (or even hate speech) is a function of:

- **Contextual factors**, or more specifically the discursive experiences of the speaker, including what

Sources: Videos/Channels	Meitei	
	Search Keywords	**Number of Comments**
Twitter	#koubru, #lawaimacha	33, 1
	#manipurdaCAB	2
	#manipurdaILP	19
	#meiteimuslim	4
YouTube	triple talaq manung hutna	390
	minister bishorjit nupi	348
	ccpur nupi, paktabi diana	138, 107
	non-manipuri, pangal nupi	77, 320
	utlou case, Naga Accord	206, 304
	manipur da CAB	1118
	Manipurdagi Meitei Furup mutpa	59
	Potti Kappe, Koubru Conflict	59, 39
	Pangal gi identity	668
	Muslim macha meitei na hatpa	509
	Momoco gi thoudok	66
	Rani Sharma, Brinda Kanano	164, 87
	M.U Normalcy, ADC Bill	577, 115
Sources: Videos/Channels	**Bangla**	
	Search Keywords	**Number of Comments**
Twitter	#SaveBangladeshiHindus	257
	#TripurarJonnoTrinamool	452
	#kutta, #khanki	42, 19
	#magi, #sala, #hijra	2, 66, 22
	#shuor, #harami	17, 62
YouTube	Police Files, Khela hobe	936, 1147
	Bengal Ram Navami	687
	Nusrat Jahan baby	964
	Ranu Mondol, Pori moni	457, 329
Sources: Videos/Channels	**Hindi & English**	
	Search Keywords	**Number of Comments**
Youtube	BJP	848
	Ripped Jeans	903
	Nehru-Gandhi	947
Telegram	Hindutva	2461
	Love jihad	249
	Feminism	21
Twitter	#MuslimVirus	1422
	#BengalBurning	651

Table 2: Sources of Raw Data, Keywords and Number of Comments

kind of discourses the speaker has been a part of, with whom, how many times, under what circumstances, and multiple other factors (see (Agha, 2006) and his theory of enregisterment for some details on this). As such in this methods, at the time of annotation the speakers continuously discuss their decisions with each other, whereby they modify each other's discursive experience(s) and possibly arrive at a space of mutual (dis)agreement. It is to be noted that these discussions did NOT involve deciding on a specific tag, rather, they focussed on what the commenters were doing in the comments. The tags were assigned by everyone independent of each other without discussing it with each other.

- **Co-textual factors** such as what else is included in the text as well as in its immediate context.

At this point, the three co-textual factors included aggression, communal bias, and gender bias. While these are individually evaluated, they are also evaluated in the presence of each other during the process of annotation.

The socio-political position and ideological leanings of the annotators working on the dataset also play a significant role in determining how the data is analysed and tagged. While attempts are made after extensive discussions within the team of annotators to draw clear

guidelines and definitions, supported by relevant examples from each language, for each tag and subtag so as to bring personal differences of opinion to a minimum, human differences between the socio-political, cultural, and ideological contexts of the annotators still manage to draw out some degree of difference in tags between the annotations of those working on the same dataset. The details of annotators who worked on annotating the datasets is included in the Data Statement attached in Appendix I of the paper for a better understanding of the decisions made by them.

2.1. Rationale behind the different levels of the Tagset

A major point of discussion that emerged from the annotation process was to do with the aggression tagset, which was divided into OAG (overtly aggressive), CAG (covertly aggressive), and NAG (non-aggressive speech). The distinction between CAG and OAG are based purely on the kind of language that one is using for expressing aggression - this distinction was important to make in order to better understand the performance of the classifiers and also for a better linguistic understanding of aggression, which would not be possible in a binary distinction. However, these sub-tags did not capture the range and intensity of aggression that was on display in the comments - from use of only curse words to threats of a physical and sexual nature. This necessitated the introduction of a subtag under aggression titled Aggression Intensity Level (with tags for physical threat (PTH), sexual threat (STH), non-threatening aggression (NtAG), and curse/abuse based aggression (CuAG)), which would only be marked if the comment was tagged OAG or CAG.

Along with this, another optional subtag was added under Aggression, titled Discursive Effects, which marks the broad function of a given comment in the discourse - attack (ATK), defend (DFN), counterspeech (CNS), abet/instigate (AIN), or gaslight (GSL). This particular subtag is used in a nuanced manner to help us identify the nature or discursive function of comments that can be found on a thread, thus, giving us an analytical tool to distinguish comments on the basis of what they are intended to do. The annotators are given access to the complete thread of comments in the same sequence as they occur on video or as they occur in a Telegram, thereby, giving them a peek into how these are read and possibly perceived by the readers of those comments. However, in case of tweets, where this kind of conversational information was not available in the way we collected the data, the discursive roles were minimally marked. Using this, the discursive effects of a conversation in a thread are marked accordingly. It is to be noted that the annotators are discouraged from tagging comments as DFN, CNS, AIN, and GSL unless they feature in a thread, and are in response to a previous comment on the same thread - this is because it would become impossible to classify these without an understanding of what the comments are responding to. This particular subtag thus helps us distinguish the ways in which aggressive speech plays out when a commenter is engaging in a dialogue with another commenter in real time versus the kind of one-way conversations that characterize independent comments.

This tagset is thus being developed during the course of annotating the raw data such that each is contributing to the development of the other. The new tagset is being used to annotate newer data, with the annotators conscious to highlight any shortcomings and flaws in the tagset that can be improved upon in the course of annotating the data, so that we can build a tagset that can be most optimal at identifying various forms of aggression and bias in social media interactions with the least margin for error.

3. The ComMA Tagset

The complete tagset is given in Table 1 and their definitions and examples are discussed in the following subsections. The full annotation guidelines is available on the project website [1] and for the lack of space could not be included here.

3.1. Aggression

Aggression is classified on the basis of three broad levels, which have been discussed below with suitable examples. It is to be noted that we are annotating the way aggression is expressed in language but NOT the intensity of aggression expressed via the text at this level. There is a common tag for all categories, which is unclear (UNC) included with other tags in Table 1. The three labels used for annotation - OAG, CAG and NAG - and their definitions remain same as in (Kumar et al., 2018) and are not being reproduced here. An example to demonstrate how each of the aggression tags have been assigned have been presented below:

1. **Overtly Aggressive (OAG)**

   Bangla: Sob khanki magi dol eder lojja nai

   Translate: Bunch of sluts, have they no shame?

2. **Covertly Aggressive (CAG)**

   Meitei: Kanglase maringna hanna leiramme haiye adu d eikhoigi church ama saba tare..

   Translate: Kangla is said to be occupied by Maring community so we should build a church.

3. **Non Aggressive (NAG)**

   Bangla: Jayta mon chay saytai kora vhalo.

   Translate: It is best to listen to one's heart.

3.2. Aggression Intensity Level

This level marks the intensity of aggression in the current post/comment. This level is marked only for aggressive posts/comments.

[1] https://sites.google.com/view/comma-ctrans/resources

3.2.1. Physical Threat (PTH)

A post/comment is potentially physically aggressive (verbal aggression transforming into physical aggression) when it directly threatens to physically harm, hit or even kill someone (an individual or the community). An example of physical threat is:

1. **Hindi:** Muh kala hai dogle ka dil bhi kala gaddar hai mujhe tum dikh jaye sala juta marunga dogala deshdrohi

   Translation: This hypocrite has lost his face, his heart is also bad, as soon as I shall see you moron, I will hit you with a shoe, you hypocritical anti-national.

3.2.2. Sexual Threat (STH)

Any post/comment that contains words expressing sexual coercion and assault is marked as sexual threat.

1. **Bangla:** Tok chudi

   Translation: Let me fuck you.

3.2.3. Non-threatening Aggression (NtAG)

Any act of aggression targeted at personal attributes like one's intelligence, physical features, various identities, or anything else but not containing a threat is classified as non-threatening aggression. Examples include:

1. **Hindi:** Bhut kam jankari h babu tmko.general caste ki aabadi 15 percent v kam h. general cat ka MATLAB hi h open category.

   Translation: You know very little, boy. The population of general caste is less than 15 percent. The meaning of general category is open category.

3.2.4. Curse/Abuse Aggression (CuAG)

Any comment containing an act of aggression that involves cursing or abusing the victim is tagged as curse/abuse aggression.

1. **Hindi:** Mujahid Irfan meri kaonsi pol khol Di iss ghade kanhaiya ne

   Translation: Mujahid Irfan what has this ass Kanhaiya revealed about me?

2. **Meitei:** Manipur se kasuba kasubi gi makon natte khngbra nama napana nahei tamhandana warktra jatlo

   Translation: Manipur is not a place for prostitutes and gigolos. Did your parents gave you proper education?

3.3. Discursive Role

This category refers to the role of the current post/comment in the ongoing discourse.

3.3.1. Attack (ATK)

Any comment/post which attacks a previous comment/post is tagged as attack. It can only be tagged for an aggressive comment.

1. **Meitei:** @suan neihsial Adunadi ta asangba pairaga adum leitaba

   Translation: @suan neihsial For that reason you will always stay forever holding your spear.

3.3.2. Defend (DFN)

Any comment/post which defends or counter-attacks a previous comment/post is tagged as defend. The previous comment/post must be an attack and the current one should be in support of the victim. It could be both aggressive as well as non-aggressive, but it is imperative that it be in the same thread as the previous comment. Examples include:

1. **Hindi:** Kitna dukhi hai bhai tu, lagta hai teri pool kholdi kanhaiya ne. Agar tu jo ilzam uspe laga raha hai wo sach hai toh wo kiyon jail me nahi hai. (Covertly aggressive)

   Translation: How sad you are bro. It seems that Kanhaiya has shown your true face. If the accusation that you are levelling on him is true then why is he not in prison.

3.3.3. Counterspeech (CNS)

Counterspeech is any direct response to hateful or harmful speech which seeks to undermine it. Counterspeech is always non-aggressive

1. **Bangla:** Hindu sobai amra ek ekhane Brahman Namasudra abar ki? Amra sobai Sanatani etai amader porichay ho

   Translation: We are all Hindus; what is this Brahmin Namasudra? We are all Sanatani, this is our identity.

3.3.4. Abet/Instigate (AIN)

Any comment/post which supports or encourages a previous (aggressive) comment and instigates an individual or group to perform an aggressive act is tagged as abet/instigate.

1. **Hindi:** Great sachchai likha aapne

   Translation: You have written a great truth. (In response to hate speech).

3.3.5. Gaslighting (GSL)

Any comment/post that seeks to minimize the trauma or distort the memory of a trauma faced by another person (usually mentioned in the previous post/comment) is tagged as gaslighting.

1. **Bangla:** Tar cheye bhalo garur mangsho khao tomra. Tomra napak jinish bokkhon korcho. Tomra opobitro. Tomaderke pobitro korar cheshta korteche. Huzoor tomader bhalo chacche.

Translation: It's better that you guys have cow's meat (than cow's urine). You people are consuming impure things. You are impure people. (The Maulvi) is trying to make you pure. Huzoor wants the best for you.

The intention of this comment is to manipulate the readers into believing that what the Maulvi, Huzoor, is asking for is for the good of the people, when it is actually a call to declare a Muslim sect, the Kadiyanis, as un-Muslims due to their "un-Muslim" beliefs and practices.

3.4. The four Biases: Gender, Religious, Caste/Class and Race/Ethnicity

The four biases included in the tagset annotates comments based on different targets of aggression viz. gendered stereotypes, traditional gender roles of the speaker or addressee, biased references to one's sexual orientation, real/presumed religious affiliation/identity/beliefs of the victim, the caste or class identity of the victim and discriminates on the basis of it and ethnic or racial or tribal identity, culture, language, skin colour, physical features, place of origin, nationality, etc.

Each of these are broadly classified into whether a given comment simply expresses a bias or also poses some threat to the victim or shows no bias. Presented below are examples of some of these tags:

1. **Gender Bias (GEN)**

   Bangla: Rate ktw magir

   Translation: What is the rate of this slut?

2. **Communal Threat (COMT)**

   Hindi: Sabhi hindi bhaiyo ek hokar hum sb en katuo ko apni takt dika do

   Translation: All Hind(u) brothers let's come together and show these Muslims our strength

3. **Non-Caste/Class Bias (NCAS)**

   Bangla: Aapnar shobdo choyoni aapnar jogyotar porichoy dicche

   Translation: Your choice of words shows who you are

4. **Racial Bias (ETH)**

   Meitei: Kanglup Khudingmak hourakpham loina lei. Aduna Kuki singi hourakfam kaidano amta haibirak o.

   Translation: Every community has its own origin. So, please tell us the origin of Kuki.

4. Inter-annotator agreement

In order to measure the IAA scores, the first phase of the annotations was done by 9 annotators - 5 for Meitei, 3 for Hindi and 2 for Bangla. The Meitei, Bangla and Hindi has 140, 275 and 230 comments respectively. Comments are taken from YouTube and Twitter. The annotation is done using the aggression, gender bias and religious bias tagset. The Krippendorff's Alpha is used to measure inter-annotator agreement for the Meitei, Bangla and Hindi, given in Table 3. We did not find any gendered comment in Meitei and communal comment in Bangla, thus the Table 3 does not have any IAA value for them. Overall from the Table 3 we can say that the IAA is very less in first phase.

Language	Meitei	Bangla	Hindi
Aggression	0.28	0.41	0.42
Gender	-	0.26	0.28
Religious	0.38	-	0.41

Table 3: Inter-annotator agreement Phase 1

In the second phase of the experiments, the guidelines were further enriched via discussions.

Language	Meitei	Bangla	Hindi
Aggression	0.51	0.66	0.93
Aggression Intensity	0.49	0.69	0.77
Discursive Role	0.66	0.74	0.75
Gender	0.56	0.81	0.80
Religious	0.27	0.80	0.82

Table 4: Inter-annotator agreement Phase 2

The results of Phase 2 of the inter-annotator agreement is shown in Table 4. It was conducted on 205 comments for Meitei, 208 and 209 comments for Bangla and Hindi respectively. There were 3 annotators in each language. There is significant improvement in Phase 2 of IAA when compared to Phase 1 and given the subjective nature of the task, we decided to move ahead with this.

5. Dataset

The dataset in its present form contains a total of 59,152 comments in four languages[2] - Meitei, Bangla, Hindi and English - each annotated by at least 3 annotators. The numbers reported in the paper are based on a majority voting aggregation of the dataset for the lack of space to report the numbers for each of the annotators. However, the dataset is being released with disaggregated labels by all the annotators instead of aggregating the annotations using majority voting or other methods. Language-wise distribution of the dataset given in Table 5 and Figure 1 is precisely 25.4% Meitei,

[2]Mn-Meitei, Ba-Bangla, Hi-Hindi, En-English and Cm-Code-Mix

26.6% Bangla, 27.1% Hindi, 17.0% English including 3.9% code-mix.

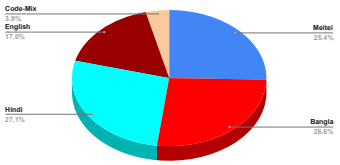

Figure 1: Languages in the Dataset

In Figure 2 the distribution of overt, covert and non-aggression in Meitei is 51.18%, 27.35% and 21.47%; in Bangla 47.74%, 14.98% and 37.28%; and in Hindi 61.54%, 15.48%, and 22.98%; in English 29.43%, 12.92% and 57.66%; in Code-Mix 51.02%, 20.69% and 28.29% respectively. In Meitei 27.35% comments are covertly aggressive while in Meitei, Bangla, and Hindi more than 45% comments are overtly aggressive.

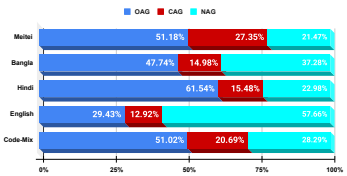

Figure 2: Aggression in the Dataset

In case of biases, overall 16.84% portion of the comments are gendered and language-wise percentage in Meitei and Code-Mix is 6.05% and 8.50% respectively, 32.92% for Bangla; 6.18% for English and 19.02% for Hindi. Total communal comments are more than 9.36% and language-wise communal comments distribution in Meitei, Bangla, Hindi, English and Code-Mix are 2.57%, 7.42%, 18.11%, 9.11% and 6.98% respectively, On broader scale Bangla has most number of gendered comments, whereas Hindi and English has most of communal comments. The other two biases are rather underrepresented in the dataset and do not form a significant part of it.

Co-occurrence graphs of Aggression with Misogyny, Communal, Caste/Class and Ethnicity/Racial categories given in Figure 3, Figure 4, Figure 5, Figure 6 show that most of the communal comments are overtly or covertly aggressive in all languages. Similar pattern could be seen for gendered comments.

In the wider sense, we can say that if a comment is gendered or communal, possibility is that the comment has some kind of aggression.

When we do a union of the 10 most frequent words taken from comments which are marked aggressive/gendered/communal or combination of these and 10 most frequent words displayed in Figures 7 and 8 in the dataset we get these words - muslimvirus, muslimsspreadingcorona, muslim / muslims, meitei, desh, hindu, nupi, pangal, BJP, corona and maa. The words like 'muslim', 'hindu'[3] , 'corona', 'bjp'[4] are mostly use

	Aggression			
	TOTAL	**OAG**	**CAG**	**NAG**
Mn	**15,001**	7,677	4,103	3,221
Ba	**15,758**	7,523	2,360	5,875
Hi	**16,032**	9,866	2,482	3,684
En	**10,056**	2,959	1,299	5,798
Cm	**2,305**	1,176	477	652
	Aggression Intensity Level				
	TOTAL	**PTH**	**STH**	**NtAG**	**CuAG**
Mn	**15,001**	395	15	6,717	4,653
Ba	**15,758**	391	293	5,838	3,361
Hi	**16,032**	403	262	7,055	4,628
En	**10,056**	57	10	3,190	1,001
Cm	**2,305**	32	9	1,056	556
	Discursive Role					
	TOTAL	**ATK**	**DFN**	**CNS**	**AIN**	**GSL**
Mn	**15,001**	1,123	30	33	562	-
Ba	**15,758**	9,295	210	119	55	8
Hi	**16,032**	1,284	29	61	382	-
En	**10,056**	1,004	28	51	243	-
Cm	**2,305**	226	16	21	36	-
	Gendered			
	TOTAL	**GEN**	**NGEN**	**GENT**
Mn	**15,001**	908	14,021	72
Ba	**15,758**	5,188	10,426	144
Hi	**16,032**	3,049	12,770	213
En	**10,056**	621	9,432	3
Cm	**2,305**	196	2,100	9
	Communal			
	TOTAL	**COM**	**NCOM**	**COMT**
Mn	**15,001**	385	14,608	8
Ba	**15,758**	1,169	14,536	53
Hi	**16,032**	2,904	13,000	128
En	**10,056**	916	9,120	20
Cm	**2,305**	161	2,132	12
	Caste / Class			
	TOTAL	**CAS**	**NCAS**	**CAST**
Mn	**15,001**	3	14,998	-
Ba	**15,758**	442	15,309	7
Hi	**16,032**	55	15,977	-
En	**10,056**	69	9,985	2
Cm	**2,305**	12	2,293	-
	Ethnicity / Racial			
	TOTAL	**ETH**	**NETH**	**ETHT**
Mn	**15,001**	868	14,132	1
Ba	**15,758**	235	15,519	4
Hi	**16,032**	42	15,989	1
En	**10,056**	417	9,636	3
Cm	**2,305**	106	2,198	1

Table 5: The ComMA Dataset

[3]Major religious group in India

[4]BJP is a Hindu Nationalist Organisation currently head-

Figure 3: Co-occurrence heatmap of Aggression with Gender Dataset

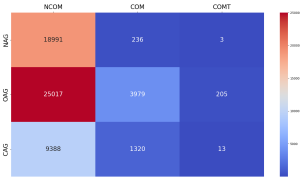

Figure 4: Co-occurrence heatmap of Aggression with Communal Dataset

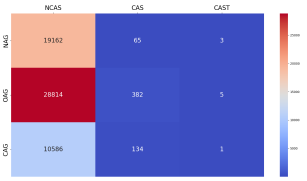

Figure 5: Co-occurrence heatmap of Aggression with Caste/Class Dataset

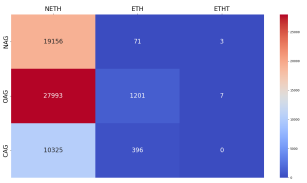

Figure 6: Co-occurrence heatmap of Aggression with Ethnicity/Raical Dataset

in the context of religion and politics were present in communal comments. The words like 'maa'(mother), 'nupi'(girl) were used in the comments which are misogynistic and gendered. A majority of comments in which the above mentioned words occurred, also turn out to be aggressive. Most gendered/communal comments are aggressive in the dataset.

When we look at the intersection of the words which are most frequent overall and also in aggressive and biased comments of various kinds, we start getting a better picture of the dataset. If we take the word 'muslims', it has occurred 184 times in the dataset out of which it has occurred 153, about 83% times in aggressive/gendered/communal comments. If we take the word 'hindu', it has occurred 169 times in the dataset out of which it has occurred 75, about 44% times in aggressive/gendered/communal comments. Similarly, the word 'nupi'(girl) has occurred 50% of the times in aggressive/gendered/communal comments. The words like 'bjp' and 'pangal' which do not have any count in the **'Frequency Agg./Gen./Com.'** column, they mostly occur in non-aggressive/non-gendered/non-communal comments. The words like 'maa'(mother) and 'corona', which do not have any count in the **'Frequency'** column , they mostly occur in aggressive/gendered/communal comments.

The analysis shows the possibilities of bias in the dataset itself and possibly in the models trained on this dataset - we are currently working on handling this bias in the dataset such that these 'trigger' words have a more balanced distribution across different classes.

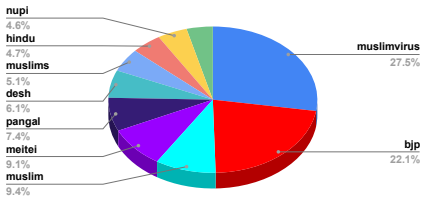

Figure 7: Top 10 Most Frequent Words In The Dataset

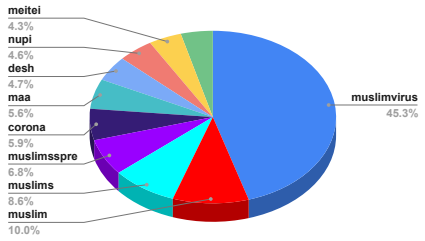

Figure 8: Top 10 Frequent Words In The Dataset For Comments Which Are Either Aggressive/Communal/Gendered Or Combination Of These

6. The Way Ahead

In this paper we have presented a multilingual dataset annotated with different levels of annotation and bias. Currently the dataset consists of over 59,152 data points in Meitei, Bangla, Hindi, English and Code-Mix, collected from various sources including YouTube comments, Facebook, Twitter and Telegram.

We are currently working on augmenting the dataset with more data points from more Indian languages from different language families and also from more sources. We are also collecting and annotating multimodal data including memes and videos across different languages.

ing the central government of India

7. Acknowledgments

This research is funded by Facebook Research under its Content Policy Research Initiative.

We would like to thank Mohit Raj, Shiladitya Bhattacharya, Afrida Aainun Murshida, Sanju Pukhrambam, Diana Thingujam and Sonal Sinha for helping us out with annotations at different points in the project.

8. Bibliographical References

Agha, A. (2006). *Language and Social Relations*. Cambridge: Cambridge University Press.

Akhtar, S., Basile, V., and Patti, V., (2019). *A New Measure of Polarization in the Annotation of Hate Speech*, pages 588–603. 11.

Albadi, N., Kurdi, M., and Mishra, S. (2018). Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. page 69–76, 08.

Alfina, I., Mulia, R., Fanany, M. I., and Ekanata, Y. (2017). Hate speech detection in the indonesian language: A dataset and preliminary study. 10.

Bhattacharya, S., Singh, S., Kumar, R., Bansal, A., Bhagat, A., Dawer, Y., Lahiri, B., and Ojha, A. K. (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, May. European Language Resources Association (ELRA).

Bohra, A., Vijay, D., Singh, V., Akhtar, S. S., and Shrivastava, M. (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, June. Association for Computational Linguistics.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012). Detecting offensive language in social media to protect adolescent online safety. pages 71–80, 09.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (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, July. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. 03.

de Pelle, R. and Moreira, V. P. (2016). Offensive comments in the brazilian web: A dataset and baseline results. page 510–519.

Haddad, H., Mulki, H., and Oueslati, A., (2019). *T-HSAB: A Tunisian Hate Speech and Abusive Dataset*, pages 251–263. 10.

Hammer, H. (2017). Automatic detection of hateful comments in online discussion. volume 188, pages 164–173, 01.

Jurgens, D., Hemphill, L., and Chandrasekharan, E. (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, July. Association for Computational Linguistics.

Kaggle. (2020). Jigsaw multilingual toxic comment classification.

Kolhatkar, V., Wu, H., Cavasso, L., Francis, E., Shukla, K., and Taboada, M. (2020). The sfu opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4, 06.

Kumar, R., Reganti, A. N., Bhatia, A., and Maheshwari, T. (2018). 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, May. European Language Resources Association (ELRA).

Malmasi, S. and Zampieri, M. (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, September. INCOMA Ltd.

Martins, R., Gomes, M., Almeida, J., Novais, P., and Henriques, P. (2018). Hate speech classification in social media using emotional analysis. pages 61–66, 10.

Mathur, P., Shah, R., Sawhney, R., and Mahata, D. (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, July. Association for Computational Linguistics.

Mubarak, H., Darwish, K., and Magdy, W. (2017). Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Nascimento, G., Carvalho, F., Cunha, A., Viana, C., and Paiva Guedes, G. (2019). Hate speech detection using brazilian imageboards. pages 325–328, 10.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. pages 145–153, 04.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523.

Schäfer, J. and Burtenshaw, B. (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, September. INCOMA Ltd.

Schmidt, A. and Wiegand, M. (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, April. Association for Computational Linguistics.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15:e0243300, 12.

Vidgen, B. and Yasseri, T. (2020). Detecting weak and strong islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17:66–78, 01.

Waseem, Z. and Hovy, D. (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, June. Association for Computational Linguistics.

Waseem, Z., Davidson, T., Warmsley, D., and Weber, I. (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, August. Association for Computational Linguistics.

Waseem, Z. (2016). Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. pages 138–142, 01.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (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, June. Association for Computational Linguistics.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (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, June. Association for Computational Linguistics.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (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), December. International Committee for Computational Linguistics.

A. Data Statement

A.1. Header

Dataset Title: ComMA Dataset v0.2

Dataset Curator(s):

- **Akash Bhagat**, Indian Institute of Technology-Kharagpur
- **Enakshi Nandi**, Panlingua Language Processing LLP, New Delhi
- **Laishram Niranjana Devi**, Panlingua Language Processing LLP, New Delhi
- **Mohit Raj**, Panlingua Language Processing LLP, New Delhi
- **Shyam Ratan**, Dr. Bhimrao Ambedkar University, Agra
- **Siddharth Singh**, Dr. Bhimrao Ambedkar University, Agra
- **Yogesh Dawer**, Dr. Bhimrao Ambedkar University, Agra

Dataset Version: Version 0.2, 2nd October 2021

Dataset Citation: NA

Data Statement Authors:

- **Enakshi Nandi**, Panlingua Language Processing LLP, New Delhi
- **Laishram Niranjana Devi**, Panlingua Language Processing LLP, New Delhi
- **Shyam Ratan**, Dr. Bhimrao Ambedkar University

Data Statement Version: 1, 17th November 2021

Data Statement Citation and DOI: NA

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

A.2. Executive Summary

The objective of working on this dataset is to identify and tag aggression and various kinds of bias (gender, communal, caste/class, ethnic/racial) in social media discourse. To that end, this dataset has been compiled by collecting over 15,000 comments from YouTube, Facebook, Twitter and Telegram in Meitei, Bangla, Hindi and English, with around 5000 comments in Meitei, Bangla and Hindi each, which were all mixed with English data. The data was collected from videos and posts that were politically, socially, sexually, religiously, racially, or otherwise polarized or controversial in nature, so as to elicit a wide and extensive range of hateful, aggressive, gendered, communal, casteist, classist, and racist speech data for our dataset.

A.3. Curation Rationale

This dataset was created with the ultimate goal of developing a system that is able to identify and tag aggression, gender bias, communal bias, caste/class bias, and ethnic/racial bias in social media discourse. To that end, this dataset has been manually annotated by multiple annotators in order to identify the linguistic and pragmatic features that characterize aggression, gender bias, communal bias, caste/class bias, and ethnic/racial bias in the comments on posts, videos, and articles posted on social media sites such as YouTube, Facebook, Twitter, and Telegram.

The specific social media posts and articles whose comments we collected were selected manually, and then crawled with the help of their respective web crawlers. This selection process was contingent on many factors, the chief of which was the need to collect as many aggressive, gender biased, communal, casteist, classist, and racist comments as possible to create a robust dataset. To that end, we focused on identifying controversial posts of a politically, socially, sexually, communally, and racially charged nature that have elicited a significant number of the kind of comments described above. We have then followed similar suggested posts, videos, or articles on the platform to collect more data of a similar or comparable nature. The second factor was language: the comments needed to be in Meitei, Bangla, and Hindi for the most part, with English comments included because they are ubiquitous in the context of Indian social media.

The dataset was organized in the form of a spreadsheet, with each comment identified by a unique comment code that would help the annotators distinguish an independent comment from comments featuring in a thread, and each comment posted under an article or video constituting one data instance, regardless of its length, language, or content. In other words, a data instance can be a single letter or an essay-length comment, can be written in a single language or a combination of languages, and can contain text (in any script), numerals, and emojis individually or all in one comment. The data instance is annotated taking the entire comment as one single, compact unit.

A.4. Documentation for Source Datasets

This dataset has been developed from a source dataset that marked only aggressive speech collected from public Facebook and Twitter pages, and a subsequent source dataset that developed the tagset to include speech with aggression and gender bias, collected from Facebook, Twitter, and YouTube.

The links for the research papers published and the workshops conducted on the respective source datasets are listed below:

1. `http://arxiv.org/abs/1803.09402`
2. Trac - 1[5], `https://aclanthology.org/W18-4401`
3. `https://arxiv.org/abs/2003.07428`
4. Trac - 2[6], `https://aclanthology.org/2020.trac-1.1`

The current dataset was built on the foundation laid down by these source datasets, and has added several new, finely-grained tags, including two primary tags marking caste/class bias and ethnic/racial bias, and two secondary tags that mark the discursive roles and discursive effects of (overtly and covertly) aggressive speech.

A.5. Language Varieties

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

- **code unavailable: 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.

[5]First Workshop on Trolling, Aggression and Cyberbullying

[6]Second Workshop on Trolling, Aggression and Cyberbullying

A.7. Annotator Demographic

The annotation scheme and guidelines for this dataset has been developed by Dr Ritesh Kumar, the principal investigator of the ComMA Project and a faculty at the Centre of Transdisciplinary Studies, Dr. Bhimrao Ambedkar University, Agra, India. He was assisted by the co-PIs of the project - Dr. Bornini Lahiri, Assistant Professor at IIT-Kharagpur and Dr. Atul Kr. Ojha and Akanksha Bansal, co-founders of Panlingua Language Processing LLP - and annotators of this dataset, who have been listed below. Further, these annotators have manually identified the appropriate posts and videos to work on, crawled the data, and then annotated and analysed the processed data in their respective languages.

- A 31-year-old Bengali Muslim woman working from Gangtok, Sikkim. She has a PhD in English, speaks Bangla, Hindi, and English, and her ideological leanings are centrist. She is annotating the Bangla data.
- A 29-year-old Bengali Hindu man working from Malda, West Bengal. He has an MA in Linguistics, and speaks Bangla, English, Hindi and Bhojpuri. He is annotating the Bangla data.
- A 33-year-old Bengali Hindu woman working from Kalyani, West Bengal. She has a PhD in Linguistics, speaks English, Hindi, Bangla, and Sylheti, and her ideological leanings are leftist. She is annotating the Bangla data.
- A 30-year-old Meitei Hindu woman working from Imphal, Manipur. She is pursuing a PhD in Linguistics, speaks English, Hindi, and Meitei, and her ideological leanings are centrist. She is annotating the Meitei data.
- A 28-year-old North Indian Hindu man working from Patna, Bihar. He is pursuing a PhD in Linguistics, speaks English, Hindi, Magahi, and Bhojpuri, and his ideological leanings are centrist. He is annotating the English and Hindi data.
- A 25-year-old North Indian Hindu man working from Agra, Uttar Pradesh. He is pursuing an M.Phil in Linguistics, and speaks Braj, Hindi and English. He is annotating the English and Hindi data.
- A 27-year-old North Indian Hindu man working from Agra, Uttar Pradesh. He is pursuing an M.Sc. in Computational Linguistics, speaks Hindi, Bhojpuri,and English and his ideological leanings are centrist. He is annotating the English and Hindi data.
- A 32-year-old Punjabi Hindu man working from Agra, Uttar Pradesh. He is an MA in Journalism and in Linguistics, speaks English, Hindi, and Punjabi, and his ideological leanings are leftist. He is annotating the English and Hindi data.

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 2012 to 2021 (and continuing), and form part of an extensive and intensive social media discourse.

- **Time and place of linguistic activity** - Online
- **Date(s) of data collection** - April to September 2021
- **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** - Socially or politically polarizing or controversial topics
- **Non-linguistic context** - The videos which provide the context for the comments generated
- **Additional details about the cultural context** - The sociopolitical climate and cultural context in which the commenters live have a huge influence on the nature, tone, and ideological underpinnnings of the comments they write on social media

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 Meitei, Bangla, Hindi, and English. In the Telegram data, all translations of texts have been deleted manually. The data instances are listed without the names of the commenters, but when someone has replied to a previous comment by tagging them with the '@' symbol, that information is available to the annotator within the text itself.

Next, the processed data is arranged on a Google spreadsheet, columns are made with the relevant headings and tags (using the option for data validation), and copies of the spreadsheet are shared amongst the annotators working on a particular language so they can annotate the files individually and without consultation with each other. This is to ensure that no annotator is influenced in their annotation by the ideas of another. However, at no stage in the process are the annotators anonymous to each other or anyone else in the team.

A.10. Capture Quality

As with any other dataset, we have faced quality issues in data capture. The primary of these is the difficulty in finding every kind of data in every language. For instance, in Bangla, it is very difficult to find racist or communal data, because most conversations that we have come across in social media platforms that are of a communal or racist nature and involves Bangla speakers occur in English. Similar challenges have been faced in Meitei with regard to casteist data, and in Hindi with regard to ethnically and racially biased data. These discrepancies can be explained when the social, political, and cultural contexts of each of these language and speech communities is taken into consideration, which are significantly different from each other.

A.11. Limitations

Following the point in the previous section, another limitation in the data is the dearth of comments that can be tagged by the discursive effects of counterspeech, abet and instigate, and gaslighting. In contrast, the discursive effect of attack is very well-represented, not so closely followed by defend. All of these factors combine to make it challenging for the dataset in each language to be equally representative of each of the primary tags, thus making it difficult for the researchers to embark on intensive comparative analyses of the characteristics of each of these phenomena across all of the languages being analysed.

This tagset also does not allow us the option to distinguish a personal attack from an identity based one, to mark national/regional or political bias, and to distinguish sexual harassment from aggression, sexual threat, and gender bias. These are shortcomings that will have to be addressed and resolved in subsequent versions of the tagset.

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**:*

1. *http://arxiv.org/abs/1803.09402*
2. *https://arxiv.org/abs/2003.07428*
3. *https://drive.google.com/file/d/1ZUZxDaYIfotVur-cJF30cfqIY1dSzJ6K/view?usp=sharing*

***Annotation Process**: Manual annotation*

***Dataset Quality Metrics**: Krippendorff's Alpha for IAA*

***Errata**: NA*

A.13. Disclosures and Ethical Review

This dataset has been funded by Facebook Research under Content Policy Research Initiative Phase 2.