AI4DIGNITY

Collaborative AI Counters Hate

NLP ❤ Anthropology

TABLE OF CONTENTS

01 ABOUT

ABOUT

- AI4Dignity is a collaboration between anthropologists, fact-checkers and NLP researchers to tackle extreme speech against marginalized communities around the globe
- Worked with fact-checkers from Brazil, Germany, India and Kenya for the data collection/annotation process
- Fact-checkers were chosen because of their accredited expertise and familiarity with local communities
- The end goal was the curation of an extreme speech dataset, labeled with granularities and targets of extreme speech

ABOUT

- *A quick note on definition divergence*
- In other works, we see terminology such as “hate speech”, “offensive language”, etc.
- Here, we use the term “extreme speech” to denote language that crosses the line between civil and uncivil speech
- Our definition and taxonomy of extreme speech was guided by both anthropologists and the annotators themselves

02 MOTIVATION

MOTIVATION (CURRENT ISSUES)

- “Usual suspects” of NLP challenges in data curation: human error during annotation, subjectivity, varying definitions
- Annotators are oftentimes white and/or male. For example, in Founta et al. 66% of annotators are male and in Sap et al. 82% are white
- These groups are usually not the target of extreme speech, so they may miss/misclassify extreme speech
- This could lead to propagation of harm against marginalized communities. For example, AAE has been annotated as extreme speech more often than other dialects of English (Kim et al.)

MOTIVATION (PLAN)

- With our plan of action we wanted to ensure quality of data while at the same time not losing track of the bigger picture
- Involved anthropologists and their expertise to build definitions and help with the theoretical background of the project
- Focus on marginalized communities around the world (in Brazil, Germany, India and Kenya)
- Sit down with annotators from each country, discuss the needs and challenges faced by their respective communities
- Hopefully data will be more representative of the hatred these vulnerable communities receive

03 DATA

DATA (GENERAL)

- We collect data for four languages: German, Hindi, Swahili, Brazilian Portuguese. In some countries, English is also used, so we included English examples in our dataset as well.
- In total, there are 20k examples (around 5k from each country)
- Data comes from social media, direct messages, forum posts, etc. Annotators were given the freedom to choose the sources and venues most fitting for their community.
- Each text example is short, usually at most a paragraph long.

"Luos are the only community whose thinking and decision making revolves around one man/family, the Ogingas"

DATA (LANGUAGES)

	Brazil	Germany	India	Kenya
Native	5245	4945	2610	400
Native & English	0	69	1172	2098
English	0	6	1034	2721

Table 1: Distribution of languages per country.

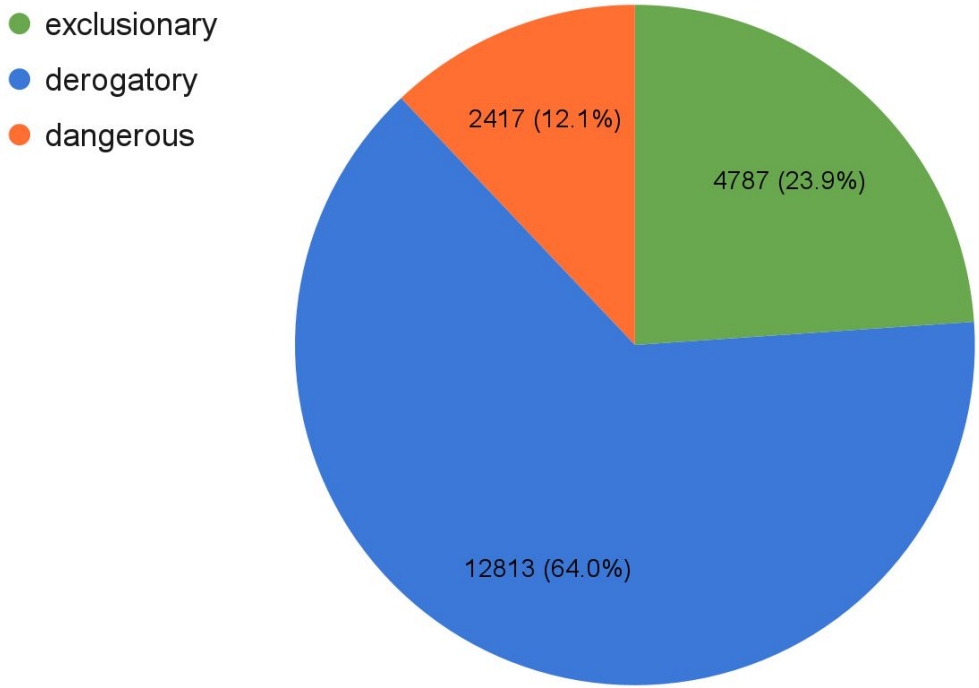
DATA (LABELS)

- Only extreme speech examples were collected. There are no “negative”/“neutral” examples.
- Three granularities of extreme speech were covered alongside eight main protected target groups.
- To capture the difference between extreme speech against protected groups and institutions of power, another option was included for four more targets (politicians, legacy media, the state and civil society advocates for inclusive societies).

DATA (EXTREME SPEECH GRANULARITIES)

- Derogatory Extreme Speech: swearwords and offensive or disrespectful language. This is considered “acceptable” extreme speech and should not be filtered out.
- Exclusionary Extreme Speech: targets a vulnerable group/s, singling them out. Unacceptable, and *should* be filtered out.
- Dangerous Extreme Speech: like the exclusionary label, but also incites violence or poses a threat.

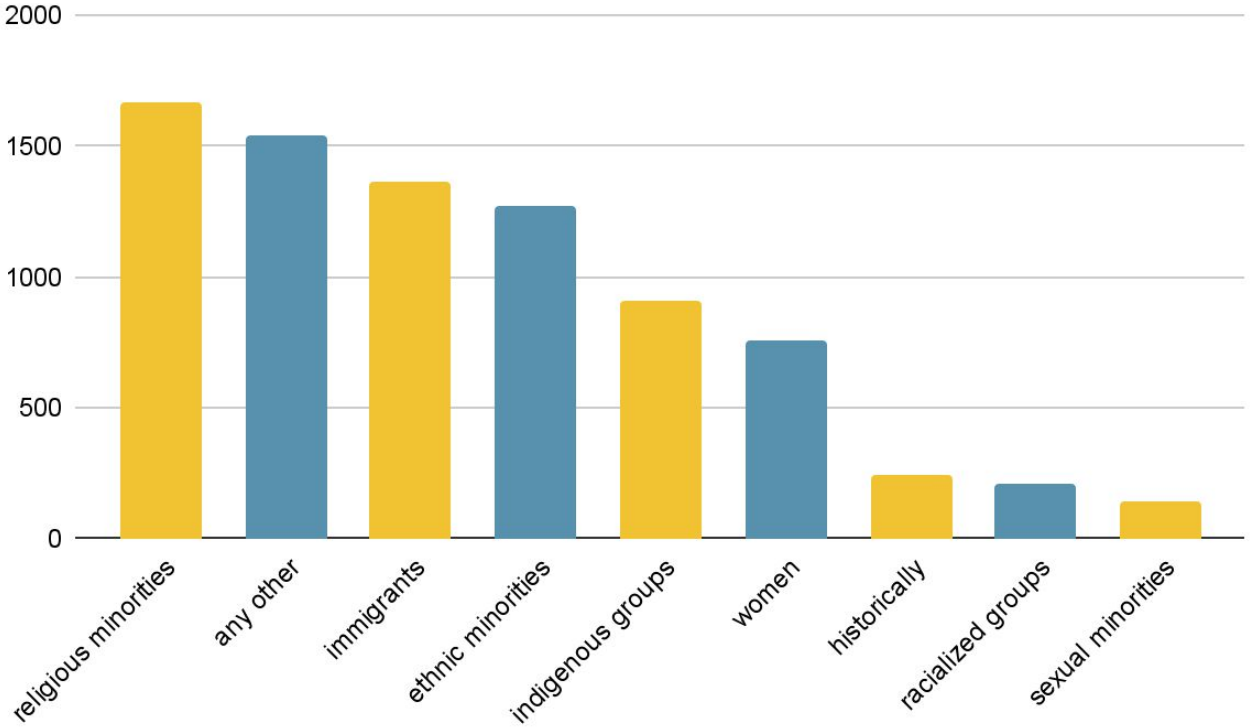
DATA (EXTREME SPEECH LABEL STATISTICS)



DATA (PROTECTED TARGET GROUPS)

- religious minorities
- immigrants
- ethnic minorities
- indigenous groups
- women
- historically oppressed caste groups
- racialized groups
- sexual minorities
- any other

DATA (TARGET GROUP STATISTICS)



DATA (EXTREME SPEECH PER COUNTRY)

	Brazil	Germany	India	Kenya	*Total*
Derogatory	4900	2659	2121	3406	*13086*
Exclusionary	121	2345	1376	874	*4816*
Dangerous	224	16	1319	839	*2398*

Table 2: Distribution of extreme speech labels across countries and in total.

DATA (PROTECTED GROUPS PER COUNTRY)

	Brazil	Germany	India	Kenya	*Total*
Religious Minorities	16	1269	3413	114	*4812*
Any Other	1058	37	340	1621	*3056*
Women	1677	372	402	397	*2848*
Immigrants	26	2383	108	290	*2807*
Ethnic Minorities	62	412	90	1316	*1880*
Indigenous Groups	61	6	4	1255	*1326*
Sexual Minorities	632	343	88	83	*1146*
Historically Oppressed Caste Groups	46	1	791	32	*870*
Racialized Groups	76	522	3	78	*679*

Table 3: Distribution of target groups across countries and overall.

DATA (INSTITUTIONS OF POWER PER COUNTRY)

	Brazil	Germany	India	Kenya	*Total*
Politicians	1120	794	255	2113	*4282*
Legacy Media	686	105	70	62	*923*
The State	56	168	19	77	*320*
Civil Society Advocates	30	57	35	10	*132*

Table 4: Distribution of institutions of power as targets of derogatory extreme speech.

04 SUMMARY

SUMMARY

- AI4Dignity is a multidisciplinary effort to tackle extreme speech online, combining the forces of anthropologists, fact-checkers and NLP researchers.
- A multilingual dataset of 20k online posts was collected, with labels for extreme speech and targets
- Unsurprisingly, our analysis shows large differences between countries and the extreme speech they have to deal with
- We hope this work will both help research in these marginalized communities and serve as a guide for more inclusivity

THE END

REFERENCES

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding
- Antigoni-Maria Founta et al. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. 2018.
- Maarten Sap et al. Social Bias Frames: Reasoning about Social and Power Implications of Language. 2019.
- Jae Yeon Kim et al. Intersectional Bias in Hate Speech and Abusive Language Datasets. 2020.
- Paula Fortuna and Sergio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text.
- Bjorn 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.
- Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny.