Aporophobia: An Overlooked Type of Toxic Language Targeting the Poor

Svetlana Kiritchenko,[1] Georgina Curto,[2] Isar Nejadgholi,[1] Kathleen C. Fraser[1]
[1]National Research Council Canada, Ottawa, Canada
[2]University of Notre Dame, Notre Dame, USA
{svetlana.kiritchenko, isar.nejadgholi, kathleen.fraser}@nrc-cnrc.gc.ca, gcurtore@nd.edu

Abstract

Content Warning: *This paper presents textual examples that may be offensive or upsetting.*

While many types of hate speech and online toxicity have been the focus of extensive research in NLP, toxic language stigmatizing poor people has been mostly disregarded. Yet, *aporophobia*, a social bias against the poor, is a common phenomenon online, which can be psychologically damaging as well as hindering poverty reduction policy measures. We demonstrate that aporophobic attitudes are indeed present in social media and argue that the existing NLP datasets and models are inadequate to effectively address this problem. Efforts toward designing specialized resources and novel socio-technical mechanisms for confronting aporophobia are needed.

1 Introduction

Online toxicity includes language that is offensive, derogatory, or perpetuates harmful social biases. Significant research effort has been devoted to addressing the problem of toxic language targeting several social groups, including women, immigrants, and ethnic minorities (Fortuna and Nunes, 2018; Kiritchenko et al., 2021). Yet, other groups (e.g., based on age, physical appearance, and socioeconomic status) also regularly experience stigmatization with severe consequences to the groups and to society at large. In this work, we focus on aporophobia—“rejection, aversion, fear and contempt for the poor” (Cortina, 2022). Cortina, the philosopher who coined the term in 1990s, argues that aporophobia is even more common than other forms of discrimination, such as xenophobia and racism. Moreover, aporophobia often aggravates intersectional bias (e.g., it is not the same to be a *rich* woman from an ethnic minority than a *poor* woman from the same ethnic group) (Hoffmann, 2019; Hellgren and Gabrielli, 2021).

In meritocratic societies, the rhetoric of equal opportunities—according to which everyone is provided with the same chances for success—assigns the responsibility for one’s welfare to each individual and results in blaming the poor for their fate (Mounk, 2017; Sandel, 2020). However, this principle does not reflect reality since every person has different abilities and disabilities, backgrounds, and experiences (Fishkin, 2014). In fact, economic indicators unveil a completely different picture: the overwhelming majority of poor people are those born into poverty (United Nations, 2018). Global levels of inequality are increasingly growing (Chancel and Piketty, 2021), social mobility is as low as 7% both in the United States and in Europe (Chetty et al., 2014; OECD, 2018), and the perception of social mobility in the US is higher than the actual opportunities to climb up the ladder, exacerbating even more the blamefulness and criminalization of the poor (Alesina et al., 2018).

Crucially, this bias has an impact on the actual poverty levels: if society considers the poor responsible for their situation and, therefore, “undeserving of help”, then measures for poverty mitigation would not be supported, thwarting the efforts towards achieving the first sustainable development goal of the United Nations to end poverty (Arneson, 1997; Applebaum, 2001).

Cortina (2022) states that evolutionary pressure has resulted in innate tendencies toward the search for reciprocity, which in market economies penalizes the poor when they are perceived as benefiting from social programs while offering nothing in return. These tendencies are further aggravated in the current Western capitalist context, where wealth is a symbol of success (Fraser and Honneth, 2003). What has been described as a “tyranny of merit” (Sandel, 2020) manifests unconsciously in our speech and writing as subtle and implicit stereotyping and rejection of the poor. Such implicit biased language can be challenging for NLP models that were not specifically trained to recognize this type of abuse (Wiegand et al., 2019; Nejadgholi et al., 2022).

To date, aporophobia has received little attention in NLP (Curto et al., 2022). In this position paper, we intend to raise awareness of this phenomenon in the community and advocate for the need to study such online behavior, its motivations and expressions, as well as its persistence and spread across online communications, and to design technologies to actively counter aporophobic attitudes. In particular, our goals are as follows:

- Characterize aporophobia as a distinct discriminatory phenomenon with significant societal impact, based on social science literature;
- Demonstrate that aporophobic attitudes are common in society and prominent in social media;
- Show that existing toxic language datasets are ill-suited for training automatic systems to address this type of prejudice due to (1) the lack of adequate sample of aporophobic instances, and (2) the failure of human annotators to recognize implicit aporophobic statements and attitudes as part of a general definition of harmful language.

The creation of resources and techniques to effectively confront aporophobia will contribute to both the safety and inclusiveness of online and offline spaces and to the effectiveness of poverty reduction efforts.

2 Societal Impact of Aporophobia

The current debate on bias and fairness mostly focuses on race and gender-based discrimination. Only recently, prejudice and bias against the poor, or aporophobia, has been described as a key distinctive discriminatory phenomenon in the social science literature (Cortina, 2022). However, international organizations have been denouncing the discrimination and criminalization of the poor for a long time (United Nations, 2018). Aporophobic attitudes have significant impact at different societal levels. At the micro (personal) level, stigmatization of the poor can inflict significant psychological harm, lead to the internalization of the continuous message of one's inferiority, and contribute to a self-fulfilling prophecy of failure (Habermas, 1990; Honneth, 1996). At the meso (institutional) level, policies for poverty reduction can be hindered by societal beliefs that the poor are responsible for their own fate and, therefore, undeserving of social assistance (Applebaum, 2001; Everatt, 2008; Nunn and Biressi, 2009). Finally, at the macro (international) level, aporophobic views are extended to blaming developing countries for their poverty, and prevent reaching fairer deals in international trade and financial markets (Reis et al., 2005; Yapa, 2002).

Aporophobia affects people across races, genders, and countries. In "Voices of the Poor," a series of publications that present poor people's own voices in 60 countries (Narayan and Petesch, 2002), a common concern has been raised that poor individuals face widespread social disapproval even from people of their own communities, races, genders, and religions. The testimonies describe situations in which "the mere fact of being poor is itself cause for being isolated, left out, looked down upon, alienated, pushed aside, and ignored by those who are better off. This ostracism and voicelessness tie together poor people's experiences across very different contexts" (Narayan and Petesch, 2002).

The impact of aporophobia is starting to be recognized by national and international organizations. Spain was the first country to include aporophobia as a distinct aggravation of hate crimes in the legal framework (Spanish Criminal Code, article 22.4), and aporophobia observatories are being created in several countries, in coordination with the United Nations. Examining and quantifying aporophobia provides NGOs and government officials with new approaches for poverty reduction policy making, acting on public awareness (in addition to redistribution of wealth) and treating poverty as a societal problem, as opposed to a problem of the poor. Mitigating aporophobia contributes to the fight against poverty for all ethnic groups and genders (Everatt, 2008) and NLP can play a key role in the identification, tracking and countering of online aporophobia.

3 Presence of Aporophobia in Twitter

In the first part of the study, we investigated the presence of aporophobia in Twitter. For this, we collected and analyzed tweets containing terms related to 'poor people' and contrasted them with tweets related to 'rich people'. Then, we performed topic modeling on tweets mentioning the group *giftofhome, encampments, encampment, sauda, dera, unsheltered, sacha, nudist,* ***defecating***, ***feces***, *blankets, shelters,* ***druggies***, *sidewalk,* ***crackheads***, ***addicts***, *doorways, tents, shelter, hostels, sidewalks, vagrants, gurmeet,* ***needles***, *evicting, rahim, sweeps, sleepers, tent, vets, skid,* ***junkies***, *toothless, outreach, camping, sacramento, sindh,* ***schizophrenic***, *portland, panhandling,* ***pooping***, *hobo, evict, motel, fatherless,* ***sodom***, *hud, isaiah, evicted, housed,* ***addict***, *motels, veterans, servicemen, fran, denver, camps, hemp, pdx, cashapp, eviction, downtown, accommodation,* ***meth***, *seattle, subways, depape, streets,* ***junkie***, *brettfavre, chinatown, unhoused, ebt, shalt, venice, hostel, freeway, newsom, sheltering, francisco, benches,* ***overdoses***, *surfing, huddled, rv,* ***overdose***, *reverend, homelessness, euthanasia,* ***addictions***, ***heroin***, *stray, houseless, belongings, cardboard, rendered,* ***urine***, ***alcoholics***, *favre, evictions*

Table 1: Top 100 words with the highest PMI-based association score (Eq. 1) for the group 'poor'. The words are presented in the decreasing order of the association score. The scores for the shown words range between 9.18 and 3.32. Words related to substance abuse, mental disorders, and health and environmental hazards associated with the homeless population are in bold.

'poor' and examined topics related to aporophobia. In the following, we discuss these steps in detail.

3.1 Tweet Collection

We polled the Twitter API to collect English tweets for a period of three months, from 25 August 2022 to 23 November 2022, using query terms related to poor and homeless people. The initial set of query terms was assembled from the social science literature on the "undeserving poor" (Everatt, 2008; Narayan and Petesch, 2002; Applebaum, 2001) and aporophobia (Cortina, 2022; Comim et al., 2020). The set was expanded with synonyms and related terms. Then, a one-week sample of tweets collected using this set of terms was manually examined. Terms that resulted in very small numbers of retrieved tweets or in many irrelevant tweets were discarded. We also excluded explicitly offensive and derogatory terms, such as *trailer trash*, *scrounger*, or *redneck*, which tend to be used in personal insults. The final list of query terms for the group 'poor' was: *the poor* (used as a noun as opposed to an adjective as in 'the poor performance'), *poor people*, *poor ppl*, *poor folks*, *poor families*, *homeless*, *on welfare*, *welfare recipients*, *low-income*, *underprivileged*, *disadvantaged*, *lower class*.

As a contrasting set, we also collected tweets related to the group 'rich' using the following query terms: *the rich* (used as a noun), *rich people*, *rich ppl*, *rich kids*, *rich men*, *rich folks*, *rich guys*, *rich elites*, *rich families*, *wealthy*, *well-off*, *upper-class*, *upper class*, *millionaires*, *billionaires*, *elite class*, *privileged*, *executives*. The single words *poor* and *rich* were not part of the search due to their polysemy (e.g., 'poor results', 'rich dessert'). Using the selected terms, we were able to collect a large amount of relevant tweets without costly manual filtering.

We excluded re-tweets, tweets with URLs to external websites, tweets with more than five hashtags, and tweets from user accounts that have the word *bot* in their user or screen names. This filtering step helped to remove advertisements, spam, news headlines, and so on. Further, tweets containing query terms from both 'poor' and 'rich' groups were also excluded. In the remaining tweets, user mentions were replaced with '@user' and query terms were masked with '<target>' to reduce the bias from the query terms in the analysis. In total, there were 1.3M tweets for the group 'poor' and 1.8M tweets for the group 'rich'.

3.2 Word Analysis

Words which are often used in tweets describing 'poor people', but rarely used in tweets describing 'rich people', are expected to be the most representative words associated with the group 'poor'. Thus, we calculated the score of association with the group 'poor' using the following formula:

$$s\,(w) = PMI\,(w, C_{poor}) - PMI\,(w, C_{rich}) \quad (1)$$

where PMI stands for Pointwise Mutual Information and was calculated as follows:

$$PMI\,(w, C) = log_2 \frac{freq\,(w, C) * N(T)}{freq\,(w, T) * N(C)} \quad (2)$$

where *freq* (w, C) is the number of times the word w occurs in corpus C, *freq* (w, T) is the number of times the word w occurs in corpus $T = C_{poor} \cup C_{rich}$, $N(C)$ is the total number of words in corpus C, and $N(T)$ is the total number of words in corpus T. Stopwords and low-frequency (< 300 occurrences in C_{poor}) words were disregarded.

Table 1 shows 100 words with the highest association to the group 'poor'. Note that these words include many terms related to alcohol and drug abuse (e.g., *addicts*, *meth*, *alcoholics*) and mental disorders (*schizophrenic*). Many tweeters also complained about unsanitary environments often

Topic words	# of tweets in topic	Example tweets
drug, addicts, mental, drugs, mentally, ill, addiction, health, addicted, addict	9,705	Homeless men are homeless because they are on drugs, got charges, are violent and up to no good. Most homeless are mentally ill, just put them into a home for the mentally ill.
crime, police, cops, criminals, jail, crimes, prison, arrest, commit, criminal	5,807	Crimes committed by the homeless against non-homeless or other homeless people occur weekly in our city. Put the homeless in jail and make work camps.
war, military, wars, army, soldiers, fight, join, recruitment, peace, die	1,687	The military preys on poor people to fight their wars. No rich kids go to war. They need more poor people to sign up to die for the state.
drunk, beer, drink, drinking, alcohol, cigarette, drunks, drinks, liquor, smoking	882	Poor folks can't run without alcohol. Drinks and deadbeat are the most beloved members of poor families.
fear, scared, scary, anxiety, afraid, terrified, scare, terrifying, fears, mongering	680	There are more homeless creeps hanging around a bicycle path than ever. People are scared they might need to walk past a homeless person when going to the mall.

Table 2: Examples of tweets expressing or discussing aporophobic views. The tweet texts were paraphrased to protect the privacy of the users.

surrounding homeless encampments and city sidewalks occupied by homeless people. Further, *euthanasia* appears in this list since many users were concerned (or some users supported) that it could become a solution to end the suffering of the poor.

3.3 Topic Modeling

Next, we analyzed the thematic content of tweets mentioning the group 'poor'. For this, we employed an unsupervised topic modeling toolkit, BERTopic (Grootendorst, 2022). The core component of BERTopic is a density-based clustering technique HDBSCAN (Campello et al., 2013), which can produce clusters of arbitrary shapes and leave documents that do not fit any clusters as outliers. This suited our case well as we wanted to discover the most commonly discussed topics in tweets mentioning poor people. The discovered topics are then represented with topic words, which are identified using class-based TF-IDF (c-TF-IDF). The 'topic words' are the words that tend to appear frequently in the topic of interest, and less frequently in the other topics.

To reduce computational costs, we applied topic modeling on a random subsample of 600K sentences from C_{poor}. For converting text to numerical representations, we used the sentence transformers method based on the *all-MiniLM-L6-v2* pre-trained embedding model.[1] For the vectorizer model, we used the CountVectorizer method,[2] and removed English stopwords and terms that appeared in less than 5% of the sentences ($min\_df = 0.05$). For the HDBSCAN clustering algorithm, we specified the minimum size of the clusters as $min\_cluster\_size = 500$. For all the other parameters, the default settings of the BERTopic package were used.

There were 142 topics extracted. We found a number of expected topics discussing the issues of homeless encampments in city parks and streets, the lack of affordable housing, the need to provide shelter and free meals to the homeless, (un)fair distribution of taxes among the socio-economic classes, Christian dogmas of helping the poor, criticism or support of government policies, and various related local issues. We also observed a number of topics with more derogatory and vilifying attitudes, portraying the poor, and especially the homeless, as drug addicts, drunkards, criminals, mentally disabled, and expendable, or expressing general feelings of fear and rejection of the group. Table 2 shows example tweets for some of these topics. Several topics tie the issues of poverty and homelessness with specific communities, such as Black people, immigrants and refugees, and veterans. Not all tweets on these topics express aporophobic views. Some report aporophobic situations, and many actually oppose such attitudes and criticize individuals and policies that hurt the poor. However, even when stereotypes are negated (e.g., 'not all homeless people are drug addicts'), the syntactic form preserves the stereotype-consistent

[1] https://www.sbert.net/docs/pretrained_models.html

[2] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

information (here: 'all homeless people are drug addicts'), confirming the stereotypic association (Beukeboom and Burgers, 2019). That is, the existence of such counter-speech is indirect evidence that such stereotypes and biases exist.

4 Unsuitability of Existing Datasets for Studying Aporophobia

Groups based on socio-economic status have been mostly overlooked in NLP research on toxic and biased language. Current lexicons designed to identify various types of social biases do not usually include status or socio-economic class categories, and in rare cases when they do, these lexicons are not tailored to identify aporophobia (Nicolas et al., 2021; Kozlowski et al., 2019; Smith et al., 2022).

Most existing datasets collected for specific targets of abuse (e.g., women, immigrants) do not include poor and low-income subpopulations as a target, with the exception of the dataset for patronizing and condensending language by Perez Almendros et al. (2020). To investigate whether these groups appear in datasets collected to study general toxicity (and its various forms), we examined nine frequently used, large, English-language toxicity datasets:

1. Civil Comments Dataset (Borkan et al., 2019): a dataset of public comments from English-language news sites annotated through crowd-sourcing for toxicity and six toxicity subtypes (severe toxicity, obscene, threat, insult, identity attack, and sexual explicit) with real-valued scores that represent the fraction of annotators who assigned the category to the comment. We considered a comment 'toxic' if it had score $>$ 0.5 for at least one of the seven toxic categories.

2. Wiki Toxicity (Wulczyn et al., 2017): a dataset of comments from Wikipedia talk page edits annotated through crowd-sourcing for six categories of toxicity: toxic, severe toxic, obscene, threat, insult, and identity attack. We considered a comment 'toxic' if it was labeled with any of the six categories of toxicity.

3. Abusive and Hateful tweet corpus by Founta et al. (2018): a large corpus of tweets annotated through crowd-sourcing for hateful, abusive, spam, and normal language. We considered a tweet 'toxic' if it was labeled as 'hateful' or 'abusive'.

4. Social Bias Inference Corpus (SBIC) (Sap et al., 2020): a collection of tweets, Reddit posts, posts from the hate communities Stormfront and Gab, and posts from a corpus of microaggressions annotated through crowd-sourcing for offensiveness, intent to offend, sexual content, target group, whether the speaker is part of the target group, and the implied statement. We considered a text 'toxic' if it was labeled as 'offensive'.

5. Unhealthy Comments Corpus (Price et al., 2020): a dataset of public comments from the Globe and Mail news website annotated through crowd-sourcing for 'healthy' vs. 'unhealthy' and for six potentially 'unhealthy' categories: (1) hostile; (2) antagonistic, insulting, provocative or trolling; (3) dismissive; (4) condescending or patronising; (5) sarcastic; and (6) an unfair generalisation. We considered a comment 'toxic' if it was labeled as 'unhealthy'.

6. Hate Speech and Offensive Language tweet corpus by Davidson et al. (2017): a dataset of tweets annotated through crowd-sourcing for three categories: (1) hate speech, (2) offensive language, and (3) neither hate speech nor offensive. We considered a tweet 'toxic' if it was labeled as either 'hate speech' or 'offensive language'.

7. Contextual Abuse Dataset (CAD) (Vidgen et al., 2021): a dataset of posts and comments from Reddit annotated for identity-directed abuse, affiliation-directed abuse, person-directed abuse, counter-speech, non-hateful slurs, and neutral. The annotations were done by two annotators, and the disagreements were resolved through an expert-driven group-adjudication process. We considered a text 'toxic' if it was labeled as 'identity-directed abuse', 'affiliation-directed abuse', or 'person-directed abuse'.

8. Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019a): a dataset of tweets annotated through crowd-sourcing for offensiveness (offensive or not), type of offense (targeted insult or untargeted), and target of offense (individual, group, or other). We considered a tweet 'toxic' if it was labeled as 'offensive'.

Dataset	**Data source**	**Categories considered ‘toxic’**	**Total # of instances**	**# of instances mentioning ‘poor’**	
				all classes	**‘toxic’**
Civil Comments	news site comments	score $>$ 0.5 for ‘toxicity’ or its subtype	1,999,515	19,140	867
Wiki Toxicity	Wikipedia comments	toxic, severe toxic, obscene, threat, insult, identity attack	312,735	168	7
Abusive and Hateful tweets	Twitter	hateful, abusive	99,996	51	9
SBIC	Twitter, Reddit, Stormfront, Gab, microaggressions	offensive	44,875	68	60
Unhealthy Comments	news site comments	unhealthy	44,355	81	3
Hate Speech and Offensive tweets	Twitter	hate speech, offensive	24,783	16	13
CAD	Reddit	identity-directed abuse, affiliation-directed abuse, person-directed abuse	23,417	84	17
OLID	Twitter	offensive	14,100	16	3
HASOC-2019	Twitter, Facebook	hate speech, offensive, profanity	7,005	19	6

Table 3: Number of instances containing the query terms for the group ‘poor’ in nine toxic language datasets. Train/dev/test splits for each dataset were merged.

9. HASOC-2019 (Mandl et al., 2019): a dataset of tweets and Facebook posts annotated by its creators for hate speech, offensive content, and profanity, and whether the offense is targeted or untargeted. We considered a tweet ‘toxic’ if it was labeled as ‘hate speech’, ‘offensive’, or ‘profanity’.

(More details on the datasets are provided in Appendix A.) We did not consider datasets annotated exclusively for hate speech since socio-economic status is not considered an attribute that defines a protected group in legal terms and, therefore, none of the existing hate speech datasets include the group ‘poor’ as a target in their definitions of hate speech.

We used the same query terms for the group ‘poor’ that we used in Sec. 3. Table 3 shows the number of instances containing these terms in the selected datasets.[3] While the sizes of the datasets vary from a few thousand to two million, most contain only a few dozen instances mentioning the group ‘poor’, and only a handful of these instances are labeled as toxic/offensive.[4] These datasets tend to be collected using query terms that frequently occur in toxic content targeting groups based on gender, ethnicity, or religion, and thus may not capture toxic content about poor people. The only noticeable exception is the Civil Comments dataset that includes over 19K instances mentioning the group ‘poor’. This is due to its size and to the fact that it comprises *all* online news comments collected through the Civil Comments platform, without any filtering. Notice, however, that the overwhelming majority of the messages mentioning the group ‘poor’ (over 95%) are labeled as non-toxic. This further demonstrates that topics related to poor people are frequently discussed online, but this group is rarely a target of NLP studies on toxicity.

In the Civil Comments dataset, toxicity was defined as ‘general incivility that would likely prompt users to leave the discussion.’ Not all instances mentioning the group ‘poor’ and originally labeled as ‘toxic’ are aporophobic as they can target other entities. We manually examined the Civil Comments test set for aporophobia, which we defined as ‘explicit or implicit expressions of rejection, aversion, or contempt towards poor or homeless people’. First, we looked at the instances originally labeled as ‘toxic’. Among 11,701 such instances in the test set, only 93 instances mention the group ‘poor’, and only 21 of them are instances of aporophobia. This clearly demonstrates that existing datasets do not contain a sufficient sample of aporophobia for classifiers to effectively learn the concept.

[3] Since SBIC has the targeted group explicitly labeled, we searched for words *poor*, *poverty*, and *homeless* in the targeted group description.

[4] Most instances mentioning the group ‘poor’ in SBIC are labeled ‘toxic’ as the majority of these instances are jokes collected from intentionally offensive subReddits.

Next, we examined instances of the Civil Comments test set originally labeled as 'non-toxic' by all of the annotators (i.e., with score of zero for all seven toxic categories). We manually annotated a random sample of 300 instances mentioning the group 'poor' and originally labeled 'non-toxic', and found 54 (18%) instances of aporophobia.[5] This indicates that aporophobic views can be expressed very subtly and are deeply rooted in our society so that none of the annotators considered these texts toxic.

To further demonstrate the unsuitability of the existing toxic language datasets for modeling aporophobia, we evaluated the performance of three publicly available, high-quality pre-trained RoBERTa-based toxicity prediction models on these aporophobic instances:

1. Detoxify[6] (Hanu and Unitary team, 2020): an open-source multi-class model fine-tuned on the Civil Comments dataset;

2. Wiki+Civil[7]: a binary toxicity model fine-tuned on the combination of the Wiki Toxicity and Civil Comments datasets;

3. TweetEval[8] (Barbieri et al., 2020): a RoBERTa-based model trained on 58M English tweets and fine-tuned on the OLID dataset.

Table 4 shows the recall these models achieve on the aporophobic instances originally labeled as either 'toxic' or 'non-toxic', i.e., the percentage of the aporophobic instances for which the models predicted the toxicity score > 0.5. For comparison, we also show precision and recall for the Toxic class these models achieve on the full Civil Comments test set. Observe that while the models demonstrate good overall performance on the test set and moderate to high recall on aporophobic instances originally labeled as 'toxic', they all fail to recognize aporophobia in more implicit instances that also proved challenging for human annotators. Overall, we conclude that the existing toxic language datasets are ill-suited for training effective models for aporophobia detection, and new datasets specifically targeting this phenomenon are urgently needed.

Model	**Full test set**		**Recall on aporophobia**	
	Prec.	**Recall**	**'toxic'**	**'non-toxic'**
Detoxify	0.57	0.83	0.57	0
Wiki+Civil	0.58	0.86	0.67	0.02
TweetEval	0.31	0.85	0.86	0.07

Table 4: Performance of three classification models on the Civil Comments test set (194,641 instances) and aporophobic instances originally labeled as 'toxic' (21 instances) or 'non-toxic' (54 instances) in the test set.

5 Discussion

Our exploratory analysis of tweets revealed a significant presence of aporophobic views expressed or confronted by the users. Since only a small percentage of people with low income use Twitter (at least in the U.S., the country with the highest number of Twitter users),[9] the views and opinions about this group come mostly from the out-group. Many users felt the need to dispute stereotypical beliefs and discriminatory actions against the poor and the homeless populations, indicating that such views are prevalent in social media and offline. Since aporophobia has not received the same attention as other types of discrimination (e.g., based on race or gender), and since it often manifests in subtle and implicit rejection or contempt, aporophobic language may not be perceived as hateful or threatening. Nevertheless, it can cause human suffering and jeopardize initiatives to fight poverty, since poverty reduction policies may not be supported when the persons in need are being blamed for their situation (Arneson, 1997).

In a context where the United Nations is calling for urgent action to end poverty, NLP techniques allow for a novel view to inform poverty reduction policies by measuring and tracking the various manifestations of aporophobia. Such instances can be organized according to the levels of negative action associated with prejudices, documented in cognitive science as avoidance, antilocution, discrimination and physical attack (Allport, 1954), at micro (individual), meso (institutional), and macro (national) levels (Comim et al., 2020). Furthermore, bias and discrimination have traditionally been studied for individual dimensions (e.g., gen-

[5] Similarly, we found instances originally labeled 'non-toxic' that contain aporophobic views in the CAD (6 out of 65) and Unhealthy Comments (11 out of 78 instances).

[6] https://huggingface.co/unitary/unbiased-toxic-roberta

[7] https://huggingface.co/SkolkovoInstitute/roberta_toxicity_classifier

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-offensive

[9] https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

der, race, etc.) (Hoffmann, 2019). Yet, different types of biases are often intertwined and aggravate one another (Lalor et al., 2022). Aporophobia can be incorporated in the intersectional view of bias and discrimination, with complex interrelations with racism, sexism, and xenophobia.

But while NLP techniques may be valuable in measuring aporophobic attitudes in written communications—such as news articles, social media, and educational material—current models, lexicons, and datasets are inadequate to effectively address this problem. In addition, expressions of aporophobia cannot simply be banned from public view. Alternative strategies for countering aporophobia and mitigating its harms need to be developed. Counter-speech and public awareness, as well as institutional and government policies, are some of the tools to reduce prejudice and discrimination against the poor. The NLP community can play a major role in developing such mechanisms in collaboration with social scientists and policy makers.

6 Conclusion and Future Work

Aporophobia is pervasive and entrenched in society, yet so far has been overlooked in NLP research on toxic language. This preliminary study indicates that existing toxic language datasets do not support the development of models for detecting and countering this type of societal bias, and new resources and methods need to be designed and built. However, since toxic and abusive language (including aporophobia) is a relatively rare phenomenon in online communications, random data sampling might be inefficient to collect appropriate amounts of aporophobic statements to characterize the phenomenon in language (Schmidt and Wiegand, 2017; Founta et al., 2018). Yet, other sampling techniques (e.g., keyword-based, content written by specific users) aiming at increasing the proportion of toxic messages can result in biased data distributions and learnt spurious correlations (Wiegand et al., 2019). Future work should address the problem of collecting data that adequately represents the phenomenon of aporophobia. Further, practical annotation guidelines and annotator training programs need to be developed to ensure that annotators have a proper understanding of aporophobia as a concept and can effectively recognize its explicit and implicit manifestations.

An aporophobia index (Comim et al., 2020), built and updated by tracking aporophobic views and actions reported or confronted on social media, can help government and non-governmental organizations analyze the trends of this phenomenon and correlate them with economic indicators on poverty and inequality. Such an aporophobia index, offering regular updates on aporophobia levels for different geographic locations, would provide valuable insights to tackle poverty as a societal problem, as opposed to a problem of the poor, and define alternative poverty reduction strategies that act on public awareness.

Research in this field is therefore critical for instrumental reasons: currently 685M people (10% of the total world population) still live in extreme poverty and the COVID-19 pandemic could make poverty levels increase by up to 8.3% (United Nations, 2022). Poverty is a worldwide problem that affects not only developing countries, but also a significant percentage of the population in thriving economies: for example, in the US, 37.9 million people live in poverty (Creamer et al., 2022). But fighting aporophobia is also essential because of intrinsic reasons: “Recognition of equal dignity and compassion is the key to an ethics of cordial reason and is indispensable to the overcoming of inhumane discrimination” (Cortina, 2022).

7 Limitations

In this exploratory study, we focused on English-language resources. Further, we examined only one social media platform, Twitter. As any other platform, Twitter has a biased demographic representation of users in terms of language, location, ethnicity, gender, age, socio-economic status, and other characteristics. In particular, Twitter is predominantly used in the United States.[10] As a result, user attitudes examined in this study primarily represent Western views and may differ significantly from views common in other regions of the world. Future studies on aporophobia need to include other languages and world regions and consider cultural differences while measuring and mitigating this type of social bias.

When searching for aporophobia-related texts, we excluded derogatory terms and slurs associated with the group ‘poor’ as such explicit forms of online abuse tend to be easier to detect by human

[10] https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

annotators and NLP models. Nevertheless, when designing tools for measuring and mitigating aporophobia such explicit expressions need to be taken into account. Furthermore, there is a wide variety of linguistic expressions referring to poor and homeless people, and sometimes this target group is not even mentioned at all, but could be inferred from the context (e.g., contexts referring to hunger, food stamps and/or other benefits, ghettos, etc.). To effectively confront aporophobia, NLP resources (lexicons, datasets, classification models) need to have a wide coverage of explicit and implicit linguistic expressions of the phenomenon.

Finally, we targeted only textual data. However, many social media posts combine text with other types of data, such as images and videos. Recent techniques for modeling multi-modal data can be employed to ensure a better coverage of various types of social media posts.

Ethics Statement

Confronting aporophobia, as an application similar to addressing other types of abusive and toxic language, poses a number of risks and ethical issues, including tension between freedom of speech and respect for equality and dignity, biased data sampling and data annotation, dual use, and many others, discussed in previous works by Hovy and Spruit (2016); Vidgen et al. (2019); Leins et al. (2020); Vidgen and Derczynski (2020); Cortiz and Zubiaga (2020); Kiritchenko et al. (2021); Salminen et al. (2021). Future research on this topic should comply with trustworthy AI principles of transparency, justice and fairness, non-maleficence, responsibility, and privacy (Jobin et al., 2019). Special attention should be paid to involving all legitimate stakeholders in the identification and definition of actions to counteract aporophobia, including the affected communities, non-governmental organizations (NGOs) and government officials working on poverty mitigation. In particular, the views and needs of the communities from both the Global North and the Global South should be included.

References

Alberto Alesina, Stefanie Stantcheva, and Edoardo Teso. 2018. Intergenerational Mobility and Preferences for Redistribution. *American Economic Review*, 108(2):521–54.

Gordon W Allport. 1954. *The Nature of Prejudice*. Basic Books.

Lauren D Applebaum. 2001. The influence of perceived deservingness on policy decisions regarding aid to the poor. *Political Psychology*, 22(3):419–442.

Richard J Arneson. 1997. Egalitarianism and the undeserving poor. *Journal of Political Philosophy*, 5(4):327–350.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Camiel J Beukeboom and Christian Burgers. 2019. How stereotypes are shared through language: A review and introduction of the social categories and stereotypes communication (SCSC) framework. *Review of Communication Research*, 7:1–37.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 491—500.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Proceedings of the Pacific–Asia Conference on Knowledge Discovery and Data Mining*, pages 160–172. Springer.

Lucas Chancel and Thomas Piketty. 2021. Global income inequality, 1820-2020: The persistence and mutation of extreme inequality. *Journal of the European Economic Association*.

Raj Chetty, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014. Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility. *American Economic Review*, 104(5):141–47.

Flavio Comim, Mihály Tamás Borsi, and Octasiano Valerio Mendoza. 2020. The multi-dimensions of aporophobia.

Adela Cortina. 2022. *Aporophobia: Why We Reject the Poor Instead of Helping Them.* Princeton University Press.

Diogo Cortiz and Arkaitz Zubiaga. 2020. Ethical and technical challenges of AI in tackling hate speech. *The International Review of Information Ethics*, 29.

John Creamer, Emily A Shrider, Kalee Burns, and Frances Chen. 2022. Poverty in the United States: 2021. *US Census Bureau*.

Georgina Curto, Mario Fernando Jojoa Acosta, Flavio Comim, and Begoña Garcia-Zapirain. 2022. Are AI systems biased against the poor? a machine learning analysis using Word2Vec and GloVe embeddings. *AI & Society*, pages 1–16.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 512–515.

David Everatt. 2008. The undeserving poor: poverty and the politics of service delivery in the poorest nodes of South Africa. *Politikon*, 35(3):293–319.

Joseph Fishkin. 2014. *Bottlenecks: A New Theory of Equal Opportunity*. Oxford University Press, USA.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of Twitter abusive behavior. In *Proceedings of the 12th International AAAI Conference on Web and Social Media*.

Nancy Fraser and Axel Honneth. 2003. *Redistribution or recognition? A political–philosophical exchange*. Verso Books.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Jürgen Habermas. 1990. *Moral consciousness and communicative action*. MIT press.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Zenia Hellgren and Lorenzo Gabrielli. 2021. Racialization and aporophobia: Intersecting discriminations in the experiences of non-western migrants and Spanish Roma. *Social Sciences*, 10(5):163.

Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, 22(7):900–915.

Axel Honneth. 1996. *The struggle for recognition: The moral grammar of social conflicts*. MIT press.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1:389–399.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

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(2):155–190.

Austin C Kozlowski, Matt Taddy, and James A Evans. 2019. The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

John P. Lalor, Yi Yang, Kendall Smith, Nicole Forsgren, and Ahmed Abbasi. 2022. Benchmarking Intersectional Biases in NLP. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3598–3609.

Kobi Leins, Jey Han Lau, and Timothy Baldwin. 2020. Give me convenience and give her death: Who should decide what uses of NLP are appropriate, and on what basis? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2908–2913, Online. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Yascha Mounk. 2017. *The Age of Responsibility: Luck, Choice, and the Welfare State*. Harvard University Press.

Deepa Narayan and Patti Petesch. 2002. *Voices of the poor: From many lands*. Washington, DC: World Bank and Oxford University Press.

Isar Nejadgholi, Kathleen Fraser, and Svetlana Kiritchenko. 2022. Improving generalizability in implicitly abusive language detection with concept activation vectors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5517–5529, Dublin, Ireland. Association for Computational Linguistics.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Heather Nunn and Anita Biressi. 2009. The undeserving poor. *Soundings*, (41):107.

OECD. 2018. A Broken Social Elevator? How to Promote Social Mobility. Technical report.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Ilan Price, Jordan Gifford-Moore, Jory Flemming, Saul Musker, Maayan Roichman, Guillaume Sylvain, Nithum Thain, Lucas Dixon, and Jeffrey Sorensen. 2020. Six attributes of unhealthy conversations. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 114–124, Online. Association for Computational Linguistics.

Elisa Reis, Mick Moore, Juliana Martínez Franzoni, and Thomas Pogge. 2005. *Elite perceptions of poverty and inequality*. Zed Books.

Joni Salminen, Maria Jose Linarez, Soon-gyo Jung, and Bernard J Jansen. 2021. Online hate detection systems: Challenges and action points for developers, data scientists, and researchers. In *Proceedings of the 8th International Conference on Behavioral and Social Computing (BESC)*, pages 1–7. IEEE.

Michael J Sandel. 2020. *The Tyranny of Merit: What's Become of the Common Good?* Penguin UK.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. 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.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180—-9211, Abu Dhabi, United Arab Emirates.

United Nations. 2018. Report of the special rapporteur on extreme poverty and human rights on his mission to the United States of America.

United Nations. 2022. The sustainable development goals report 2022.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos One*, 15(12):e0243300.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. 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.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. 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 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pages 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Lakshman Yapa. 2002. How the discipline of geography exacerbates poverty in the third world. *Futures*, 34(1):33–46.

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.

A Existing Toxicity Datasets

We used nine large, English-language, toxicity datasets:

- **Civil Comments Dataset**[11] (Borkan et al., 2019): The dataset includes public comments from about 50 English-language news sites across the world posted in 2015-2017 via the Civil Comments platform. The comments were annotated for toxicity and six toxicity subtypes (severe toxicity, obscene, threat, insult, identity attack, and sexual explicit) through crowd-sourcing. Each toxicity and toxicity subtype score is a real value which represents the fraction of annotators who believed the category applied to the given comment. The dataset is released under CC0, as is the underlying comment text, and was used in the Jigsaw Unintended Bias in Toxicity Classification Kaggle challenge.

- **Wiki Toxicity**[12] (Wulczyn et al., 2017): The dataset consists of comments from Wikipedia's talk page edits. The comments were annotated through crowd-sourcing for six categories of toxicity: toxic, severe toxic, obscene, threat, insult, and identity attack. The dataset is released under CC0, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. It was used in the Jigsaw Toxic Comment Classification Kaggle challenge.

- **Abusive and Hateful Tweet Corpus by Founta et al. (2018)**[13]: This is a large corpus of tweets annotated through crowd-sourcing for hateful, abusive, spam, and normal language. The tweets were selected using a boosted random sampling technique where a random sample was complemented with tweets that showed strong negative polarity and that contained at least one offensive word. This boosting procedure helped improve the coverage of the minority (non-normal) classes since hateful and abusive tweets tend to appear quite rarely in the Twitter stream. We used an updated version of the dataset with 100K annotated tweets. The dataset is available by requesting access from the authors.

- **Social Bias Inference Corpus (SBIC)**[14] (Sap et al., 2020): The dataset contains textual instances collected from various sources: posts from three intentionally offensive subReddits (r/darkJokes, r/meanJokes, r/offensiveJokes), posts from two English subreddits that were banned for inciting violence against women (r/Incels and r/MensRights), posts from known English hate communities Stormfront and Gab, posts from a corpus of microaggressions (Breitfeller et al., 2019), and a sample of tweets from three existing English Twitter datasets created by Founta et al. (2018); Waseem and Hovy (2016); Davidson et al. (2017). Each instance was annotated via crowd-sourcing for offensiveness, intent to offend, sexual content, target group, whether the speaker is part of the target group, and the implied statement. The dataset is publicly available. We used version 2.

- **Unhealthy Comments Corpus**[15] (Price et al., 2020): The dataset includes public comments from the Globe and Mail (a large Canadian newspaper) news website randomly sampled from the SFU Opinion and Comment Corpus dataset (Kolhatkar et al., 2020). Only comments with 250 characters or less were included in the sample. The comments were annotated through crowd-sourcing for the binary category 'healthy' vs. 'unhealthy' and for the presence of six potentially 'unhealthy' categories: (1) hostile; (2) antagonistic, insulting, provocative or trolling; (3) dismissive; (4) condescending or patronising; (5) sarcastic; and (6) an unfair generalisation. All labels are binary and include confidence scores. The labels and confidence scores were obtained as aggregated answers of multiple annotators taking into account the annotators' 'trustworthiness' scores. The dataset is released under CC BY-NC-SA 4.0.

- **Hate Speech and Offensive Language Tweet Corpus by Davidson et al. (2017)**[16]: The dataset consists of tweets collected using hateful words and phrases compiled by Hatebase.org. The tweets were annotated through crowd-sourcing for three categories: (1) hate speech, (2) offensive language but not hate speech, and (3)

[11] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

[12] https://www.kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge/data

[13] https://github.com/ENCASEH2020/hatespeech-twitter

[14] https://maartensap.com/social-bias-frames/

[15] https://github.com/conversationai/unhealthy-conversations

[16] https://github.com/t-davidson/hate-speech-and-offensive-language

neither hate speech, nor offensive. The dataset is publicly available.

- **Contextual Abuse Dataset (CAD)**[17] (Vidgen et al., 2021): The dataset contains a stratified sample of posts and comments from 16 subReddits, which were identified as likely to contain higher-than-average levels of abuse. The messages were collected over 6 months from 1st February 2019 to 31st July 2019. All posts and comments were manually annotated within the context of conversational threads for six primary categories: identity-directed abuse, affiliation-directed abuse, person-directed abuse, counter-speech, non-hateful slurs, and neutral. Each instance was assigned to one or more of the six categories. The annotations were done by two annotators, and the disagreements were resolved through an expert-driven group-adjudication process. The dataset is released under CC Attribution 4.0 International. We used version 1.1.

- **Offensive Language Identification Dataset (OLID)**[18] (Zampieri et al., 2019a): The dataset consists of tweets collected using query terms and constructions that are often included in offensive messages, such as 'you are', 'she is', 'gun control', 'MAGA', etc. The tweets were annotated via crowd-sourcing for offensiveness (offensive or not), type of offense (targeted insult or untargeted), and target of offense (individual, group, or other). The dataset is publicly available. It was used in the shared task SemEval 2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) (Zampieri et al., 2019b), and is part of the evaluation benchmark TweetEval (Barbieri et al., 2020).

- **Hate Speech and Offensive Content Identification in Indo-European Languages (HASOC-2019)**[19] (Mandl et al., 2019): The dataset consists of Twitter and Facebook posts collected using hashtags and keywords that contained offensive content. The posts were manually annotated by the creators of the dataset for hate speech, offensive content, and profanity, and whether the offense is targeted or untargeted. The dataset is publicly available and was used in the first edition of the HASOC track at FIRE 2019. We used only the English portion of the dataset.

[17] https://zenodo.org/record/4881008#.Y6dTinbMIuU

[18] https://github.com/cardiffnlp/tweeteval

[19] https://hasocfire.github.io/hasoc/2019/dataset.html