Proceedings of the Eighteenth International AAAI Conference on Web and Social Media (ICWSM 2024)

Xenophobia Meter: Defining and Measuring Online Sentiment toward Foreigners on Twitter

Khonzoda Umarova, Oluchi Okorafor, Pinxian Lu, Sophia Shan, Alex Xu, Ray Zhou, Jennifer Otiono, Beth Lyon, Gilly Leshed

Cornell University, Ithaca, NY, USA
{ku47, oco6, pl444, js2896, ax39, rz95, jco66, mbl235, gl87}@cornell.edu

Abstract

Xenophobia, a form of hatred directed at foreigners, immigrants, and sometimes even people who are just perceived as foreigners, has been flooding social media in recent political climates. In order to capture language related to foreigners and those perceived-as-foreigners (F&PAF) we present the 7-scale Xenophobia Meter, ranging from anti– to pro-F&PAF sentiments with examples and application rationale. We also publish a dataset of over 7,000 tweets labeled according to this meter, from 11 U.S.-based accounts that are on the forefront of defining the rhetoric related to immigration and policy. We apply a number of models to automatically identify xenophobic and F&PAF-related language. We also present findings from qualitative interviews with human annotators about their labeling experiences. While we find xenophobia is a complex social phenomenon to identify by both humans and machine learning algorithms, we hope that our work inspires researchers, policymakers, and the public to learn about xenophobia and to make efforts to shift the rhetoric and policies toward allyship, equity and inclusion.

Introduction

The last decade has seen significant changes in global migration patterns due to involuntary displacement of millions of people in the Middle East, Africa, and Latin America, many of whom sought refuge in countries in Europe and North America (Achiume 2018). Recently, there has been a shift in political climates around the world with strong anti-immigration rhetoric resurging in many countries, including the U.S. Furthermore, the recent global COVID-19 pandemic has amplified xenophobic sentiment in politics, the media, and online, especially toward Chinese people and people of East Asian descent (Tahmasbi et al. 2021; Noel 2020; Esses and Hamilton 2021; Shen et al. 2022).

In this research, we are interested in quantifying and analyzing the landscape of xenophobic speech produced by high-profile public figures on social media, specifically Twitter. Researchers have been studying xenophobia both as a specific type of hate speech and on its own (Schmidt and Wiegand 2017; Khatua and Nejdl 2022). As a subset of hate speech, xenophobic speech is often context-dependent and nuanced (ElSherief et al. 2021). Yet, most work has focused on explicit anti-immigration expressions, often excluding content that is implicitly xenophobic, or content with pro-humanitarian or anti-xenophobic expression (Pitropakis et al. 2020). We aim to fill this gap, by providing a holistic and nuanced range of sentiments and attitudes towards immigration and "foreigners". We don't limit the definition of "foreign" to be based on citizenship or country of origin, but expand it to include actual or perceived immigration or foreigner status; hence we adapt the term F&PAF (foreigner and perceived-as-foreigner).

Working with a fine-grained measure of F&PAF-related sentiment on social media, we are interested in the perspectives of both humans and machine learning algorithms. Thus, we pose the following research questions:

- **RQ1:** How well can machine learning models identify and categorize nuances of F&PAF-related sentiment expressed on Twitter?
- **RQ2:** How do human annotators' experiences impact the identification and characterization of xenophobia?

Developed in the context of U.S.-based immigration and foreign policies and rhetoric, our first contribution is the Xenophobia Meter, a taxonomy of a range of sentiment toward F&PAFs, both positive and negative. The meter characterizes implicit and explicit variations of these sentiments, as well as more extreme cases of calls for action or incitement of violence. Our second contribution is a dataset[1] of over 7,000 tweets from 11 Twitter accounts that are on the forefront of U.S. immigration-related rhetoric, labeled by trained humans using the Xenophobia Meter. As a third contribution, we present comprehensive insights into the dataset through a range of quantitative and qualitative analyses.

We analyze primary dataset properties, such as distribution of labels and inter-rater agreement scores, as well as explore the language of tweets in the data. We find that existing sentiment analysis and general hate speech classification models are lacking when it comes to classifying xenophobia, especially if it is masked behind language that appears neutral in the absence of broader context. Further, we investigate the dataset's potential for identifying and classifying F&PAF-related sentiment by training a variety of models. Finally, to grasp the limitations of the labels in the dataset,

[1]The full dataset is published with Harvard Dataverse: https://doi.org/10.7910/DVN/KYR4IY.

we seek perspective into the human labeling efforts and experiences. Among other observations, we witness how individual experiences and past encounters with xenophobia impact labeling. Hence, discussions are important not only to clarify labeling criteria, but also to resurface these personal biases and gain a broader awareness of ways in which xenophobia presents or disguises itself.

Related Work

Xenophobia in Society and Online

Xenophobia is an irrational distrust, prejudice, fear or hatred towards individuals based on their perceived/actual foreigner status (Bordeau 2009). As population make up of countries change in the current age of globalization and immigration, tensions between migrant and indigenous groups, especially related to economic and cultural factors (differences in everyday lifestyle, values), give rise to xenophobia (Wimmer 1997). This tension "stems from the perception of equality and difference, of legitimate and illegitimate competition" (Wimmer 1997; Belanger and Pinard 1991). Some argue that internal societal crisis is the primary cause of xenophobia and racism, as "ways of reassuring the national self and its boundaries" (Wimmer 1997).

With the growth and increasing prevalence of social media platforms, xenophobia, along with other forms of hate speech and "othering" behaviors, found its way to online domains, where it is more visible, amplified, enabled, and even empowered (Tontodimamma et al. 2021; Harmer and Lumsden 2019). Hate speech targets a person or a whole group based on certain characteristics, including race, ethnicity, gender, sexual orientation, religion, or nationality, in some cases simply perceived by the perpetrator. Characterized by expressing hatred and intentionally causing harm to a group or an individual, hate speech incites bad actions and potentially violent responses (Sellars 2016).

The ambiguous, situational, and context-specific nature of hate speech makes it difficult to develop a universally-accepted definition of hate speech, and hence to quantify its presence online (Strossen 2016; Zhang and Luo 2019). Different world organizations and social media platforms have their own definitions of hate speech (Fortuna and Nunes 2018). While online platforms have regulations and mechanisms to ensure tagging and removal of hateful, offensive content, timely identification, the ambiguity of hate expressions, and nuances of language in different contexts make the task difficult.

Detecting Hate Speech on Social Media

Previous work on detecting and measuring hate speech online has looked at language on the levels of sentiment, lexicon, linguistic characteristics, topics, knowledge-based features, and other meta-information (Schmidt and Wiegand 2017; Tontodimamma et al. 2021; Zannettou et al. 2020). Previous research efforts used existing hate speech datasets to build various classifiers for automatic hate speech detection: Burnap and Williams (2016) looked at hate speech targeting race, disability, and sexual orientation; Watanabe, Bouazizi, and Ohtsuki (2018) studied differences between hateful, offensive and clean language; and Zhang and Luo (2019) explored deep learning approaches to hate speech detection. Expert-defined or crowd-sourced lexicons are a common way of identifying hateful content online (Davidson et al. 2017). Mathew et al. (2019) created a large dataset of hateful content and users on the social media platform Gab to further study patterns of hate diffusion across the platform via conversations.

As we searched for a detailed classification of content expressing sentiments toward foreigners, we found that many existing studies collapse expressions related to nationality (xenophobia) with race (racism) (Frías-Vázquez and Arcila 2019; Rzepnikowska 2019; Miller et al. 2016; Sylvia Chou and Gaysynsky 2021; Lloret-Pineda et al. 2022; Benitez-Andrades et al. 2022) or gender (misogyny) (Basile et al. 2019; Gallego, Gualda, and Rebollo 2017) Our work therefore focuses on xenophobia as a freestanding and unique phenomenon, detached from hate speech that refers to other protected categories.

Given the context-dependent nature of hate speech, substantial work has been applied to create labeled datasets for specific contexts. For instance, Frías-Vázquez and Arcila (2019) collected tweets associated with a specific context (refugees on MS Aquarious) to label sentiments and hate speech directed towards refugees vs. politicians. Similarly, Pitropakis et al. (2020) collected and annotated a large immigration-related Twitter dataset to study language features associated with anti-immigrant speech. Burnap and Williams (2015) used crowd-sourcing to label a Twitter dataset with language offensive or antagonistic toward race, ethnicity, or religion and further created models to identify hateful tweets and explore their language. Khatua and Nejdl (2022) created a Twitter dataset annotated with social perceptions (sympathy / antipathy) and behaviors (solidarity / animosity) towards migrants.

Unlike previous work, our collection of tweets is not based on specific events (Frías-Vázquez and Arcila 2019) or "relevant" keywords or hashtags (Pitropakis et al. 2020; Burnap and Williams 2015; Khatua and Nejdl 2022). Instead, we focus on a set of entities and organizations with active Twitter accounts, which are at the forefront of mainstream media rhetoric on topics related to immigration. Furthermore, our 7-category Xenophobia Meter allows to capture more nuanced and fine-grained sentiments towards F&PAF, in comparison to commonly used binary rating scales (Pérez-Landa, Loyola-González, and Medina-Pérez 2021).

When annotating a dataset for the tasks of detecting complex social phenomena such as xenophobia, racism, or misogyny, the labeling itself becomes a complex social process of constructing the ground truth of the phenomenon (Muller et al. 2021). In addition, the labeler's perception and understanding of the phenomenon, developed metric, and application criteria change with time in so called the "concept evolution" effect (Kulesza et al. 2019). Therefore, the nature of the phenomenon studied and the quality of the final labeled dataset are bounded by the labeler's knowledge (Waseem 2016). As such, when building models and classifiers based on human-labeled data, taking annotation qual-

ity metrics into account is essential for a more accurate understanding of performance and limitations of such models (Gordon et al. 2021). Thus, in addition to the dataset we also provide insight into the labeling process from the perspective of annotators.

Methodology

The Xenophobia Meter

In order to understand and quantify xenophobia online, in consultation with a team of experts in migration law & policy, human rights, and AI, we developed a "meter" for the expression of attitudes toward F&PAFs, and over the course of the first 2.5 years the meter was iterated and refined. To the best of our knowledge, the Xenophobia Meter is the first effort to classify social media data on a detailed scale of expressions of attitudes toward F&PAF.

Addressing the gap in literature for studying xenophobia as a freestanding phenomenon with its own complexities, the project founders created the detailed typology to incorporate the insights gathered from a group of experts, including a legal scholar with decades experience in migrant rights and advocacy, researcher at the intersection of law and technology, international law scholar, and an information scientist. This 7-scale Xenophobia Meter captures an ordinal range of attitudes, from "Very Xenophobic" to "Very Pro-Foreigner".

We intentionally developed a more detailed scale than previous efforts that detect hate speech as a binary classification task (Waseem 2016; Watanabe, Bouazizi, and Ohtsuki 2018; Schmidt and Wiegand 2017), in order to capture a nuanced insight into language use as it relates to F&PAFs, immigrants, and immigration policy. We also decided to include both positive and negative sentiments with the purpose of highlighting both problematic xenophobic speech and the possibility of allyship, equity, and inclusion.

Building on the distinction between attitudes and behaviors toward F&PAF (Khatua and Nejdl 2022), we developed the scale (presented in full detail in Table 1, with reasoning criteria and examples) with a positive and negative category in each level:

- Level 3 is applied to most extreme content with active, explicit language that translates to offline behaviors:

  (-3) calls for violent or hurtful actions toward F&PAFs.

  (+3) calls for action *against* xenophobic acts or to benefit F&PAFs.
- Level 2 is used for content with explicit language that is less active and behavior-oriented:

  (-2) openly hateful content toward F&PAFs or in support of xenophobic policies or movements.

  (+2) criticizing xenophobic acts or openly supporting pro-humanitarian policies or movements.
- Level 1 is for language that is more passive, implicit, and subtle that indirectly indicates suggestive opinions:

  (-1) indirectly hurtful content towards F&PAFs.

  (+1) generally humanitarian content.
- (0) is reserved for neutral content that shares facts and statistics or for content that is unrelated to F&PAFs, thus outside of the scope of this meter. This includes tweets that exhibit racism or anti-racism in the absence of any F&PAF reference. This was a difficult decision, as xenophobia and racism are often intertwined (Rzepnikowska 2019). We come back to this challenge later in our qualitative findings.

The original topology contained the seven-category scale, and over the years of labeling, we expanded the reasoning criteria accompanying each category. The multiple reasoning criteria are helpful for training labelers to understand the nuances of xenophobia and what it may look like and to increase the efficiency of labeling. The result is highly nuanced encoding that may give more flexibility to future researchers who use the dataset and facilitate correlations.

Although initially developed in the context of Twitter data related to U.S.-based immigration and foreign policies and sentiments, in the long term, we hope to extend the Xenophobia Meter to other kinds of online content, such as other social media platforms, news articles, memes, imagery, videos, and ads, as well as to other global and local contexts and languages.

Constructing the Dataset

Dataset collection Since one of our goals is to highlight the accountability of public actors bearing legal or moral responsibility for their speech, we focus on discourse by figures and entities such as journalists, government agencies, and non-profit organizations with verified Twitter accounts. This is in contrast to other dataset collected using searched keywords or hashtags (Pitropakis et al. 2020; Burnap and Williams 2015; Khatua and Nejdl 2022). While these actors have presence on a variety of media platforms, at the time of gathering the data, due to its ease of API usage in the past, we settled on the Twitter platform.

To choose which accounts to include in the dataset, we assembled a group of 5 politically-savvy and legal scholars who are highly active on Twitter. In a brainstorming session, they proposed a range of Twitter accounts that produce a high volume of tweets related to immigration, immigrant communities, and F&PAFs. From this list, the group selected 12 accounts, three in each of four categories with the likelihood of being: highly pro-F&PAF, generally pro-F&PAF, generally anti-F&PAF, and highly anti-F&PAF. After we started labeling, one account proved to contain little foreigner-related content, so we dropped it and focused on the remaining 11 accounts. These accounts (Table 2) include government agencies that enforce immigration-related policies, entities supporting certain groups for which the topic of immigration is of interest, public figures known to express opinions about immigration, and other organizations promoting different relevant agendas.

For each source, we collected 3200 of the latest tweets as of September 2020. We obtained information such as partial tweet textual content, tweet id, and the tweet link. Using Twitter API we obtained information such as reply, retweet, and like counts. Tweets and relevant information were stored in Google spreadsheets for labeling.

Category (Value)	Reasoning Criteria	Example Tweet
Very Xenophobic (-3)	• Calls for violent or hurtful action/policies • Uses xenophobic slur • Calls for action that clearly disadvantages immigrants • Xenophobic hashtags that do call for action or use a verb (Ex. #deportallimmigrants)	"The fight for a border wall along our southern border is far from over. We need your help to keep the pressure on the amnesty opposition. Can we count on your help?" (@FAIRImmigration)
Xenophobic (-2)	• Openly hateful towards a foreign group • Openly supports xenophobic orgs/policies/-movements • Places foreign groups or pro-immigrant entities in a negative light by giving concrete examples. • Commending action that clearly disadvantages immigrants • Criticizes policy that clearly advantages immigrants • Xenophobic hashtags that do not call for action or use a verb (Ex. #immigrantssuck)	"Secretary of State Michael Pompeo on Wednesday traveled to Wisconsin to speak to state officials about how the Chinese Communist Party (CCP) is increasingly attempting to influence and divide Americans." (@BreitbartNews)
Slightly Xenophobic (-1)	• Indirectly hateful towards a foreign group using coded or suggestive language • Anti-humanitarian comments	"In fairness, among Americans there is persistent disbelief that the Chinese allow themselves to be genocided by their own government every few decades." (@TuckerCarlson)
Neutral (0)	• Shares facts, statistics or news concerning foreign groups and foreigner-related organizations /policies but does not comment • Not foreigner related	"ICE announces extension to I-9 compliance flexibility" (@ICEgov)
Slightly Pro-Foreigner (1)	• Expresses abstract opposition against xenophobia • Pro-humanitarian comments • Anti-racist comments that do not specifically refer to foreigners	"Syrians who have returned home to rebuild their lives are now faced with many difficult challenges." (@UNHRCUSA)
Pro-Foreigner (2)	• Criticizes/shames specific xenophobic acts or policies that clearly disadvantage foreigners (ex. quoting xenophobic sentiments, replying to a xenophobic tweet in opposition) • Openly supports movements/policies/orgs that are pro foreigner • Places foreign groups or pro-immigrant entities in a positive light by giving concrete example • Commending action/policy that clearly benefits foreigners • Pro-foreigner hashtags that do not call for action or use a verb (Ex. #ICEsucks)	"It can be tough to find an accessible way to explain complex #immigrationlaw processes to clients. Thanks to Peter Choi for this idea of describing the green card process through the lens of a DMV visit: *shared_link*" (@AILANational)
Very Pro-Foreigner (3)	• Calls for action against xenophobic acts or to benefit foreigners • Shares the story of a victim of xenophobia • Pro-foreigner hashtags that do call for action or use a verb (Ex. #abolishICE)	"Please sign this petition asking Congress to give permanent protections to TPS holders who continue working during #Covid19 as healthcare professionals, meat packaging & food processing workers, delivery drivers, warehouse workers, service industry staff. *shared_link*" (@BAJItweet)

Table 1: Xenophobia Meter with reasoning criteria and examples for each of the 7 categories.

Account	Organization description	Type	# of tweets labeled	IRR all labelers	IRR no deviators
AAAJ_AAJC	A non-profit organization advocating for Asian American communities	Non-profit	785	0.568	0.616
AILANational	The American Immigration Lawyers Association	Government	701	0.433	0.545
BAJItweet	Black Alliance for Just Immigration	Non-profit	471	0.560	0.622
BreitbartNews	A far-right news network	News	789	0.590	0.685
FAIRImmigration	Federation for American Immigration Reform	Non-profit	728	0.570	0.641
ICEgov	U.S. Immigration and Customs Enforcement	Government	661	0.621	0.629
IngrahamAngle	Political talk show on Fox News Channel	News	514	0.670	0.778
splcenter	Southern Poverty Law Center	Non-profit	623	0.565	0.624
StatePRM	U.S. Bureau of Population, Refugees, and Migration	Government	524	0.324	0.345
TuckerCarlson	Tucker Carlson Tonight show host on Fox News Channel	News	828	0.633	0.682
UNHCRUSA	U.S. chapter of United Nations High Commissioner for Refugees	Government	408	0.384	0.466

Table 2: Information about the 11 selected Twitter accounts and the number of tweets labeled in each account. The last two columns correspond to Inter-Rater Reliability scores (IRR) achieved by labelers when (i) considering all labels and (ii) removing labels contributed by "deviators."

Dataset labeling The labeling efforts spanned over two years, and involved a changing team of human labelers trained to apply one of the seven Xenophobia Meter ratings to a tweet. Labelers were recruited among students on the university campus for a 10-weekly hourly-paid job. Highly committed labelers who worked for longer periods were promoted to a supervisor position.

All recruited labelers start with a training session conducted by supervisors. In the training, labelers are provided with information about U.S. immigration and foreign policy and learn about the meter, and the labeling workflow instructions. Together, they label a set of tweets curated by supervisors specifically for training and conclude with a discussion about the process and the assigned labels.

Henceforth, labelers are provided with weekly sets of about 200 tweets to label independently, at their own pace. Supervisors distribute the sets across the accounts and the labelers, ensuring that each tweet has at least two labelers. To label a tweet, the labeler starts by reading it in the spreadsheet, and uses the tweet URL to review any additional media and thumbnails in the Twitter platform. Based on these, the labeler first identifies if the tweet is "relevant to F&PAF" (Yes/No); "Not F&PAF-relevant" tweets are automatically assigned a rating of (0). Then, the labeler decides which rating from the Xenophobia Meter to apply, and indicates the rationale for their chosen rating in a drop-down menu in the spreadsheet, with options corresponding to *Reasoning Criteria* in Table 1. Finally, they indicate their level of confidence in choosing the rating (Yes/No); labelers may also skip a tweet they are unsure how to rate. Retweets are skipped[2]; specifically, BAJItweet, StatePRM, and UNHCRUSA had the highest proportions of retweets in the dataset.

At the end of each week, labelers reconvene in a *calibration meeting* facilitated by the supervisor. Prior to the meeting, the supervisor lists tweets for discussion: those marked as unconfident or with omitted ratings, and where multiple labelers disagreed on their ratings by an absolute difference of 2 or more. In the meeting, labelers discuss their ratings and rationale for each of the listed tweets. Discussions cover U.S. policies, nuanced language use, and possible political motives related to a tweet and its language. The supervisor documents the discussion in the spreadsheet, and after the discussion, a labeler may change their rating.

Inter-rater reliability To measure agreement between labelers, we computed Krippendorf's alpha for inter-rater reliability (IRR) (Krippendorff 2011) across all the labelers for each source (Table 2). We observe higher IRR scores for openly partisan sources, such as IngrahamAngle and TuckerCarlson, and lower scores for accounts belonging to some government entities (in particular StatePRM, UNHCRUSA, and AILANational, but not ICEgov). We also observe, for ICEgov especially, that tweets often appear as neutral facts, although repeatedly presenting facts about immigrants committing crimes paints an unfavorable picture of F&PAF.

We further looked into identifying individual labelers who may be responsible for more disagreements, which we define when the rating they gave for a tweet is more than one standard deviation away from the mean rating. By quantifying such disagreements, we identified labelers – "deviators" – for whom the number of disagreements is greater than one standard deviation from the average number of disagreements for each source. Several labelers appear as deviators across multiple sources. Computing IRR scores without labels provided by the "deviators" improves the agreement scores for sources such as AILANational, IngrahamAngle, BreitbartNews, UNHCRUSA, and FAIRImmigration. However, for some accounts the improvement is minimal (StatePRM and ICEgov). We come back to these "deviators" later, when we present the experience of being a

[2]Labelers were instructed to not label retweets, and as a rule of thumb used that a tweet that starts with "RT" is a retweet. Retweets that didn't have this indicator were labeled.

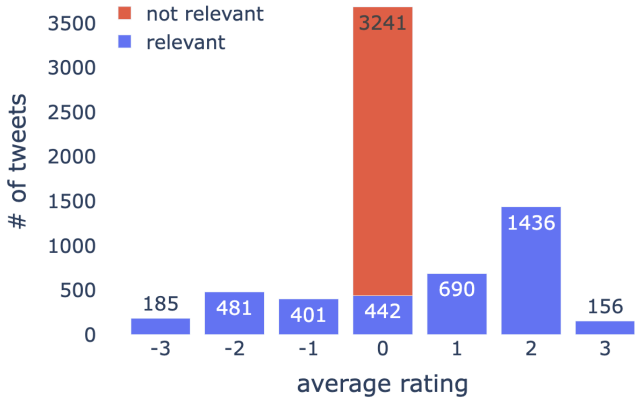

Figure 1: Distribution of average tweet ratings across all accounts for tweets relevant and not relevant to F&PAF.

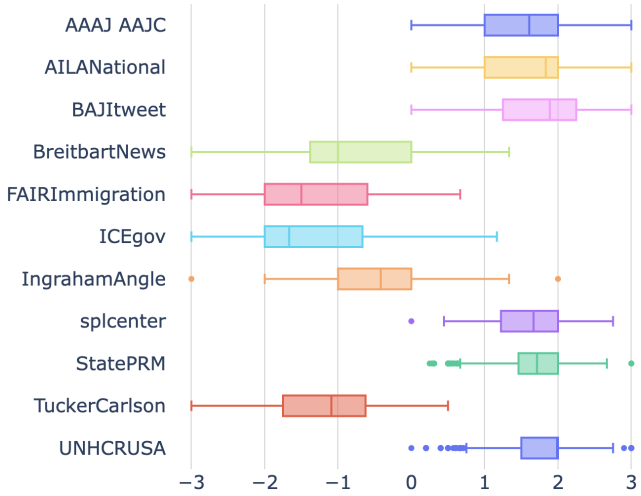

Figure 2: The distribution of ratings for F&PAF-relevant tweets for each of the 11 accounts.

labeler and conflicts around ratings in the labeling team.

Dataset Analysis

Over the course of 3 years, 34 labelers provided over 39,000 unique labels to over 7,032 tweets across all selected accounts, with over 6,423 tweets labeled by at least three labelers. Our analyses below use the data obtained from all labelers, including the "deviators" we identified in the previous section. For each tweet, we calculated the average rating given by different labelers for the tweet. The distribution of these ratings is presented in Figure 1. We see that the majority of tweets are neutral, with more tweets falling on the pro-F&PAF side than anti-F&PAF. Figure 2 presents the distribution of average tweet ratings for each account. This distribution corresponds with our expectations given their U.S. partisan perspective, interests, and activities.

Next, we present statistical analyses we conducted of the dataset, to demonstrate its efficacy to capture attitudes toward F&PAF across the sources. We begin by presenting sentiment analysis, and continue with models for classifying F&PAF-relevance and for predicting levels of xenophobia.

Sentiment Analysis

In order to understand and quantify the sentiment expressed in our dataset, we applied a pre-trained model on the task of Sentiment Analysis (Rosenthal, Farra, and Nakov 2017) with *positive*, *neutral*, and *negative* label categories. Specifically, we used the Twitter RoBERTa model[3] by (Loureiro et al. 2022). For all analyses below we first removed mentions and urls from tweet text, then applied pre-trained models, and finally normalized the scores across metrics and standardized by tweet length. We observe the highest *negative* sentiment in tweets by TuckerCarlson and most *positive* sentiment in tweets by StatePRM. Other accounts are mostly *neutral*, with ICEgov and FAIRImmigration having the highest neutral scores.

Next, we analyzed the distribution of language in each account across six emotions – *joy*, *sadness*, *anger*, *surprise*, *disgust*, and *fear* – for the task of Emotion Recognition (Plaza del Arco et al. 2020). Due to majority of tweets being classified as "other", we adjusted the counts to exclude such tweets. TuckerCarlson and BreitbartNews have predominantly language associated with *disgust*. Interestingly, we observe the language of *fear* in AILANational and UNHCRUSA, which we believe to be used in the pro-F&PAF context when it comes to the topic of immigration.

Finally, we investigate patterns captured by existing Hate Speech models pre-trained on the task of detecting *hateful* tweets that *target* women or immigrants, while simultaneously classifying language as *aggressive* or *non-aggressive* (Basile et al. 2019). We find that highly partisan accounts, such as TuckerCarlson[4] had very high scores of hateful and aggressive speech, followed by FAIRImmigration and IngrahamAngle. This finding is interesting, because earlier we found FAIRImmigration having predominantly *neutral* tweets. This discrepancy may be explained by the example of using the term "illegal aliens", which may sound neutral, but in the context of immigration is a hateful slur. Curiously, BreitbartNews and TuckerCarlson had very few tweets tagged by the model, which is likely due to a large number of tweets from these accounts being on topics unrelated to F&PAFs. To our surprise, the model tagged BAJItweet, which is known to be an organization fighting for immigration justice, for hateful language. It is possible that this account uses African-American English in some tweets, which is wrongly detected as abusive by many hate speech classifiers (Davidson, Bhattacharya, and Weber 2019; Sap et al. 2019). Alternatively, it is possible that its language is not in itself xenophobic, but rather is directed toward anti-F&PAF policies and groups.

[3]For analyses in this section we use models by CardiffNLP: https://huggingface.co/cardiffnlp/.

[4]https://www.vox.com/2018/3/21/17146866/tucker-carlson-demographics-immigration-fox-news

Classifying Relevance

Our data collection approach of identifying sources rather than keywords or hashtags, means that not every tweet in the dataset is relevant to immigration. Figure 1 presents the number of tweet labels that are relevant vs. not relevant to F&PAF in the labeled dataset. Within the accounts, there is a high number of "not F&PAF-related" tweets from news media accounts like TuckerCarlson, BreitbartNews, and splcenter, as they publish content on a variety of topics. Given that a large portion of the collected tweets are not relevant to F&PAF and considering the cost of labeling, moving forward, it may be possible to streamline the labeling process by only assigning human labelers with tweets that are relevant to F&PAF. Therefore, we train a machine learning classifier that identifies tweets as relevant or not relevant.

First, we marked the tweet as "relevant" if at least half of the labelers rated it as such. With this we had 3,791 and 3,241 tweets relevant and not relevant to F&PAF, respectively. We removed urls, mentions, and emojis from tweet text, tokenized them, and further obtained tweet embedding vectors to use as input for the classifier by averaging embedding vectors of its words. For classification, we tried simple regression models as well as DistilBERT (Sanh et al. 2019). The results are in Table 3 along with standard deviation values computed against a range of random seeds, with DistilBERT achieving the best performance, F1 score of 0.870. Further, we experimented by adding features from sentiment analyses. To do so, we augmented each tweet with an additional sentence: *"this tweet is mostly* {*polarity*} *and* {*hate speech*} *and contains mostly* {*emotion*}*"*. We found a slight improvement in F1 scores (to 0.873) when utilizing sentiment, emotion, and hate speech features (See Table 3).

A: Classifying relevance (models)	F1 score (SD)
Logistic Regression	0.766 (0.011)
SVM	0.754 (0.011)
Multi-Layer Perceptron	0.756 (0.013)
DistilBERT	0.870 (0.010)
DistilBERT+sentiment	0.873 (0.008)
B: Classifying xenophobia (models)	Accuracy (SD)
(7-category) Linear Regression	0.416 (0.017)
(7-class) K-nearest Neighbors	0.445 (0.010)
(3-class) Logistic Regression	0.811 (0.007)
(3-class) DistilBERT	0.893 (0.010)
(3-class) DistilBERT+sentiment	0.893 (0.009)

Table 3: A: Performance of different classifiers for identifying tweets as relevant to F&PAF. B: Performance of different classifiers for applying Xenophobia Meter for 7- and also simplified 3- categories.

Classifying Xenophobia

With the goal to discover and quantify the language of xenophobia, next we trained and evaluated machine learning models to apply Xenophobia Meter on tweets in the dataset. Similarly to the procedure above, after filtering out tweets that were not relevant to F&PAFs, we cleaned and obtained vector representations for the remaining ones. We started by applying linear regression models for the continuous 7-category scale, as well as models for multi-class classification, achieving the highest accuracy of 0.445 with K-nearest Neighbors classifier.

We also experimented with a simplified version of the problem, where 7 categories are collapsed into 3-class system of positive, neutral, and negative (+1, 0, -1). In this 3-class classification setting we try Logistic Regression, DistilBERT and DistilBERT augmented with sentiment features (like in the previous section). We achieved much better results in this 3-class classification problem (Table 3), although, the classifier struggled with the "neutral" category. We achieved the best accuracy of 0.893 with DistilBERT.

The Labeler Experience

Aware that the capabilities of machine learning models are bounded by the quality of the data they are trained on (Geiger et al. 2020) , we seek to provide a holistic understanding of our dataset through exposition of the labeling process and its complexities (Denton et al. 2020; Gebru et al. 2021) . Un-blackboxing the labeler experience puts into context the usage and application of the Xenophobia Meter and the dataset. It also clarifies some of our quantitative findings: low inter-rater reliability scores, shortcomings of existing hate speech recognition models, and worse performance of our xenophobia categorization models in 7-category setting in comparison to a 3-class task. Uncovering the human side of the labeling process (Muller et al. 2021) can also be useful for outlining challenges and strategies in labeling similar politically-laden datasets.

Qualitative Labeler Interviews

Between June and August 2022, we conducted interviews with labelers who participated in the project in the past two years. We recruited 11 labelers and supervisors for semi-structured interviews over Zoom, lasting 40-60 minutes, and compensating each with a $15 Amazon gift card. See Table 4 for an anonymized list of the interviewees.

The interview protocol was intended to explore perspectives, challenges, and experiences of the human labelers in the project. We started with obtaining informed consent, followed by introductions and demographics questions, and asking about their previous knowledge and experience with xenophobia.

The majority of the interview was about the labeling process: we asked labelers to share their screen, demonstrate a typical labeling session workflow, and walk us through the labeling procedure and elaborate on their thought processes while assigning labels. We also asked about the training and labeling guidelines they received, if and how they were useful, and suggestions for improvements. We also asked about the calibration meetings and their dynamics: how they were run, how conflicting labels were discussed, and how labelers handled and reconciled disagreements over labels.

Supervisors were asked additional questions, about the development and evolution of the labeling guidelines, and the logistics of supervisor workflow, including creating la-

Pseudonym	Pronouns	Age group	Race/Ethnicity	Country of Origin	Role
Valerie	she/her	18-24	Hispanic/Latino	USA	Supervisor
Riya	she/her	25-34	South Asian	India	Supervisor
Lizzy	she/her	18-24	Hispanic Asian	USA	Supervisor
Marley	she/her	18-24	Asian American	USA/China	Labeler
Tamir	he/him	35-44	South Asian	Pakistan	Labeler
Susan	she/her	18-24	Black/African American	USA	Labeler
Mary	she/her	18-24	White	USA	Labeler
Grace	she/her	18-24	Asian	China	Labeler
Bridget	she/her	18-24	Black/African	Rwanda	Labeler
Gabriel	he/him	18-24	Asian American	USA	Labeler
Martha	she/her	18-24	White/Hispanic	USA	Labeler

Table 4: Interview participants and their demographics information.

beler spreadsheets, distributing tweets between labelers, supervising the labeling progress, and running calibration meetings. We inquired about challenges experienced by supervisors during various steps of the process and suggestions for improvements.

All interviews were audio-recorded and fully transcribed. To analyze the data, we followed open-coding and axial-coding (Saldana 2012). Four researchers read the transcripts and open-coded each paragraph. We then used a shared whiteboard[5] to cluster the codes, going back to the transcripts to read the underlying data to ensure cluster consistency. Finally, we extracted insights that emerged from the clusters and narrated them based on the underlying data, using excerpts from the interviews to preserve the labelers' perspectives. Next, we detail the primary insights that emerged.

Findings

Growth throughout the labeling journey While all labelers are university students, they come from different backgrounds and experiences (Table 4). Their incoming knowledge about xenophobia varied, from no knowledge at all to having taken several relevant courses. The training, interviewees reported, was useful for getting everyone familiar with xenophobia in a structured way, through the lens of the Xenophobia Meter and the training examples. Riya, who is from India, said: *"I have no context about a lot of these slur words, etc. I know they exist, but I don't know the history behind it. I don't know why it is so offensive and why it could be so triggering for people. I learned so much about that."*

Learning and growth then continued during calibration meetings, where discussions ensued about language use, policies, current events, each others' perspectives. Riya, who was initially a labeler and then a supervisor, said about calibration meetings:

> *"The biggest benefit was that you could reorient people into how to label. [...] The second was it fostered a place where everyone could learn from each other. [...] And third, it was a place to actually bring your doubts, like you might be struggling, so just talk about what is happening, what is going wrong, or what you are struggling with in terms of linguistics [...], so it was just a place for everyone to learn, to ask questions."*

All the interviewees talked about how they had benefited from labeling. Over time, they became confident in their labeling decisions and developed their own labeling styles. They learned about xenophobia as a concept, xenophobic and anti-xenophobic language, and became more aware of current events and politics: *"The more you see a certain type of tweet, the more you start to see keywords, for example, 'our country,' 'our jobs,' [...] so you start to understand the subtle ways that people are exhibiting xenophobia"* (Susan). And, while the labeling task was detail-oriented, time consuming and demanding, they felt that they contributed to social good by being part of the project.

Strategies to streamline labeling As university students, labelers needed to find ways to manage the labeling workload: *"balancing school and other extracurricular events, just being able to find a good time to sit down and read tweets and label things"* (Gabriel). They reported that labeling was initially difficult and time consuming, as their knowledge about xenophobia, U.S. immigration policies, and nuanced language use was limited, and they had to frequently consult the labeling guidelines. The more they labeled and learned about the topic, the easier labeling became. Martha said: *"Once I became more knowledgeable in that I had to look at the rating rationale, I just kind of knew off the top of my head."*

With experience, labelers also developed strategies to speed up the labeling, save time and reduce the work difficulty. Marley, for instance, learned the tone of different accounts, whether they were more likely to be on the positive or negative side of the meter, and what to expect in their tweets: *"There were some accounts that we would follow that a lot of them just didn't have xenophobia so that would be another way that I would filter them. Like Laura Ingraham, Tucker Carlson, Breitbart News, they would say a xenophobic thing once in a while, but a lot of it was Trump fans storming the capital. 'We love Trump'. That is in my opinion problematic in other ways, but that is not xenophobic. And there would be masses of non-xenophobic ones in a*

[5]We used mural.co for the affinity diagramming.

row." Marley also used the keyboard shortcut Command+F to search for xenophobic slur words, such as "illegal" and "alien", which are a reason to label a tweet as *Very Xenophobic* (-3). She reported that these strategies made the labeling task easier.

Exposure to negative content Being a more experienced labeler also came with more exposure to large amounts of xenophobic or negative content. This sometimes impacted labelers personally, as international student Grace said: *"when I was actually getting into the tweets for a long time, [...] that is when I started to think about, oh, this is really disrespectful to me because I'm a foreigner myself here in this country."* Labelers developed strategies to deal with the negativity. Bridget said she expected some accounts to be more negative and prepared by mentally distancing herself from the negativity. Lizzy reported taking breaks during labeling to cope with the negativity. And, Grace reminded herself of the project goal to fight xenophobia: *"I tried to force myself into that mindset by saying, I am doing this for the social good."*

To address this problem, supervisors implemented various strategies, such as onboarding new labelers with more positive accounts, providing experienced labelers with an equal mix of tweets from positive and negative accounts, and using the calibration meetings: *"The weekly meetings were a great way to be like, wow, this week my God, the tweets I saw. So it was a safe space where people could talk about their emotions."* (Riya).

Individual biases and inconsistencies The labeling procedure was designed with the ultimate goal to develop a valid and reliable xenophobia-labeled dataset. This included the training session, labeling guidelines, multiple labelers for each tweet, and calibration meetings. Yet, inconsistencies in labeling were evident during calibration meetings and later, as we analyzed inter-rater reliability.

Initially, some labelers were unfamiliar with policies and xenophobic rhetoric. For example, Bridget, an international student from Rwanda, said: *"China and the U.S. over the Covid-19, that is something that I did not expect to be xenophobic."* For others, prior personal experiences, ranging from offensive remarks to hateful comments to aggressive and threatening behavior, made them more sensitive to anti-foreigner expressions. For example, Marley, who is Chinese-American, told us about an encounter with her White athletic coach: *"my coach said, well I'm not going to get Chinese food because I don't want Coronavirus. Actually, you are Chinese, should I get away from you too?"*

Even after training, learning about policies and rhetoric, and gaining experience, some conceptual challenges remained, especially at the intersection of racism and xenophobia. Valerie described a tweet she found difficult to decide if it was "relevant to F&PAF": *"it could be referencing anti-Asian, but also Asian-Americans, so again, is this actually foreigner-related? Because if it is Asian-American, are they foreigners? But then again [...] what is perceived foreigner and what is not perceived foreigner, or how are we identifying that with xenophobia?"*

Strong opinions also impacted labeler choices. Some labelers, for example, disagreed with the idea of labeling tweets that included racist speech that was not related to F&PAF as Neutral (0). Susan eloquently explained her views on the intersection between xenophobia and racism:

> *"When people are from France or from Spain who come over and are immigrants and study in schools, people are like, oh my gosh, that is so cool, verses when you get people who are brown, Black, Asian, now people are taking our jobs, now people are disruptive to society, and so that is where you see building a wall, that is where you see coronavirus named as something else other than coronavirus."*

Earlier, we reported on inter-rater reliability scores, identifying "deviators", labelers who had typically different ratings than others. Susan acknowledged being one:

> *"I had a very different outlook on racism and xenophobia as being one of the only people who was Black indigenous person of color, and so a lot of times my labeling would be very different and more extreme. [...] I think it is good to have someone like me who is a radical labeler, who will look for racism and xenophobia in everything that she reads. I also think it is good to have people who will look at it from a standpoint of like, no, I just read it as a tweet."*

Calibration meetings were designed to resolve inconsistencies and conflicts in labeling. However, both labelers and supervisors recognized that conflicts weren't always easy to resolve, as labelers are human with opinions and beliefs. Supervisor Riya explained: *"their stance was politically very different, not politically, but because of their background, because of where they came from, etc. [...] and in the end, it was not as if they were right or someone else was wrong, it was just difference of opinion."* Gabriel explained his viewpoint: *"Sometimes I just felt strongly in the way that I rated something. Not necessarily that I disagreed with how people rated theirs, but [...] I wanted to stick with it."*

Despite our best efforts to develop a reproducible labeling process and a valid xenophobia meter, we eventually have to rely on human labelers to create the labeled dataset. As humans with past experiences, background knowledge, and political opinions, identifying one correct ground truth for xenophobic expression may be tricky.

Discussion and Future Work

Making Data Collection Decisions Keeping in mind the goal of accountability for public actors and contrasting undesired xenophobic speech with alternative pro-humanitarian sentiment, we intentionally selected a small set of prominent Twitter accounts and avoided putting individual Twitter users under the spotlight and exposing them to unnecessary public attention. Given our team's expertise in U.S. immigration law and foreign policy, we focused on U.S.-based accounts. Moreover, the choice of accounts was based on the knowledge and views of the team members who participated in the initial brainstorming session, which is an important limitation. This means that our dataset is not representative of the entire Twitter landscape and has

a distinct U.S.-centric focus. Even though models trained on this data might not be applicable to randomly-selected tweets, the Xenophobia Meter criteria can still be applied to reason about pro- or anti-F&PAF sentiments. Further work is needed to enrich Xenophobia Meter with examples from more accounts, and to expand to other countries and contexts (Frías-Vázquez and Arcila 2019). Another potential extension of Xenophobia Meter is to identify, similar to ElSherief et al. (2021), which groups sentiments are directed at and why.

Further, recent changes associated with Twitter, functionalities of the platform and the API not only led to disruptions in Twitter ecosystem (Schulman et al. 2023; Conger 2023) but also question the future accessibility of Twitter data for researchers. Fortunately, in the context of this work, the 11 selected actors have presence on other media platforms, where they actively pursue their respective agendas. Future work involves adapting the Xenophobia Meter to other social media platforms and continuing to monitor pro- and anti-F&PAF sentiment of public figures and entities. The already collected and labeled data would still be useful for identifying and reasoning about F&PAF-related sentiment from prominent public actors in the sphere of immigration and foreign policy, for instance if looking from a historical perspective (the tweets were collected in the fall of 2020 prior to the U.S. presidential elections).

Considerations for Politically-Laden Labeling Consistent with previous research, from interviews we, too, observe the multitude of ways in which involving a diverse group of labelers, a nuanced 7-scale meter, and the sensitive and potentially triggering topic of xenophobia became a complex social process (Muller et al. 2021). Given calls to unblackbox dataset construction and human labeling processes (Denton et al. 2020; Gebru et al. 2021), our qualitative findings contextualize some of the quantitative findings, and can be used by researchers working on politically-laden datasets and topics.

One important outcome of this research is training university student annotators to learn about and become aware of xenophobia and the ways it manifests in public speech. Our interview findings highlight their learning and growth journeys. In future labeling iterations we implement a mechanism to consistently track history and order of tweet batches released for labeling. With this we hope to better model and quantitatively track trajectories of growth. Such mechanisms could also be used to test, cautiously, whether crowdworkers, who represent a more diverse population, could apply labeling of politically-laden content. At the same time, we built structures of training, supervising, and calibration that don't necessarily transfer well to crowdsourcing platforms, and that go beyond simply achieving accurate gold standard datasets (Sen et al. 2015).

Calibration meetings, where labelers discuss tweets, how each should be rated, and why, became a space for learning about xenophobia and each other's world views. At the same time, some labelers noticed that their ratings and views tended to be different from other labelers, which overtime has resulted in some identifying themselves as "radical". Individual biases stemming from labelers' past experiences and backgrounds led to differing interpretations of the same tweet, which in turn affected reliability scores. As part of the inter-rater reliability analysis, we used a heuristic to quantify such "deviators". This technique could be a resource for supervisors during training to ensure labeler's understanding of Xenophobia Meter and labeling guidelines. We caution against applying this technique broadly to get rid of extremes, as this would limit the variety of opinions and the range of valid interpretations of tweets and Meter categories (Kairam and Heer 2016), even at the cost of lower reliability scores. Hence, in our quantitative modeling, we intentionally included all labels, including those created by "deviators".

One of the immediate next steps for our work is to apply the models we trained as AI tools to streamline continued labeling efforts of Twitter data (Desmond et al. 2021; Karlos et al. 2019). With the given accuracy results, the F&PAF-relevance classifier would the a good initial step in the labeling pipeline and would concentrate the work of labelers on applying Xenophobia Meter only to relevant tweets. Finally, it would be interesting to adapt Xenophobia Meter classifier for balanced sampling of tweets from each of the both pro-F&PAF and anti-immigration sides of the scale, as well as regulate labeler exposure to potentially negative content.

Conclusion

We describe the creation of the dataset of over 7,000 tweets from public entities labeled according to the nuances of sentiment and attitudes toward people perceived as foreigners in the U.S. This dataset would be useful not only for researchers to study nuances of language of xenophobia, but also to educate the general public on examples of both xenophobia and pro-F&PAF sentiment on U.S. Twitter. Beyond dataset creation, the labeling process and especially calibration meetings had a positive effect of being a platform for labelers, university students, to learn more about xenophobia and the range of real-life experiences from one another.

Ethics Statement The team has considered ethical implications of our work at multiple stages of the project. We collect publicly available Twitter data, only selecting accounts that represent organizations, government agencies, or public figures. During training and labeling, supervisors implemented various strategies to minimize the harm from exposure to negative content (see details in The Labeler Experience section). We envision this work to help identifying and understanding xenophobia in its various forms and to push the public rhetoric towards inclusivity and allyship. However, we are also aware and understand that the meter and data can be used to target that same rhetoric we are trying to advocate for.

Acknowledgements The Xenophobia Meter Project was launched with support from Global Cornell's Mario Einaudi Center for International Studies and Migrations initiative, with support from a Just Futures partnership with the Andrew W. Mellon Foundation. We are very grateful for the collaboration from Bao Kham Chau, Pranoto Iskandar, Lily Pagan, Marten van Schijndel and his students, Joshua Jacob, Cornell Hack4Impact, and generations of student labelers.

References

Achiume, E. T. 2018. Governing xenophobia. *Vand. J. Transnat'l L.*, 51: 333.

Basile, V.; Bosco, C.; Fersini, E.; Nozza, D.; Patti, V.; Rangel, F.; Rosso, P.; and Sanguinetti, M. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics.

Belanger, S.; and Pinard, M. 1991. Ethnic movements and the competition model: some missing links. *American Sociological Review*, 446–457.

Benitez-Andrades, J. A.; González-Jiménez, Á.; López-Brea, Á.; Benavides, C.; Aveleira-Mata, J.; Alija-Pérez, J.-M.; and García-Ordás, M. T. 2022. BERT Model-Based Approach for Detecting Racism and Xenophobia on Twitter Data. In *Metadata and Semantic Research: 15th International Conference, MTSR 2021, Virtual Event, November 29–December 3, 2021, Revised Selected Papers*, 148–158. Springer.

Bordeau, J. 2009. *Xenophobia*. The Rosen Publishing Group, Inc.

Burnap, P.; and Williams, M. L. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2): 223–242.

Burnap, P.; and Williams, M. L. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data science*, 5: 1–15.

Conger, K. 2023. Twitter Begins Removing Check Marks From Accounts. *The New York Times*.

Davidson, T.; Bhattacharya, D.; and Weber, I. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. Florence, Italy: Association for Computational Linguistics.

Davidson, T.; Warmsley, D.; Macy, M.; and Weber, I. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, 512–515.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; Nicole, H.; and Scheuerman, M. K. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets.

Desmond, M.; Muller, M.; Ashktorab, Z.; Dugan, C.; Duesterwald, E.; Brimijoin, K.; Finegan-Dollak, C.; Brachman, M.; Sharma, A.; Joshi, N. N.; et al. 2021. Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. In *26th International Conference on Intelligent User Interfaces*, 392–401.

ElSherief, M.; Ziems, C.; Muchlinski, D.; Anupindi, V.; Seybolt, J.; De Choudhury, M.; and Yang, D. 2021. Latent Hatred: A Benchmark for Understanding Implicit Hate Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 345–363. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Esses, V. M.; and Hamilton, L. K. 2021. Xenophobia and anti-immigrant attitudes in the time of COVID-19. *Group Processes & Intergroup Relations*, 24(2): 253–259.

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

Frías-Vázquez, M.; and Arcila, C. 2019. Hate speech against Central American immigrants in Mexico: Analysis of xenophobia and racism in politicians, media and citizens. In *Proceedings of the Seventh International Conference on Technological Ecosystems for Enhancing Multiculturality*, 956–960.

Gallego, M.; Gualda, E.; and Rebollo, C. 2017. Women and refugees in Twitter: Rhetorics on abuse, vulnerability and violence from a gender perspective. *Journal of Mediterranean Knowledge (ISSN 2499-930X)*, 2(1).

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Geiger, R. S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R.; and Huang, J. 2020. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 325–336. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Gordon, M. L.; Zhou, K.; Patel, K.; Hashimoto, T.; and Bernstein, M. S. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14.

Harmer, E.; and Lumsden, K. 2019. Conclusion: Researching 'online othering'—Future agendas and lines of inquiry. In *Online Othering*, 379–395. Springer.

Kairam, S.; and Heer, J. 2016. Parting crowds: Characterizing divergent interpretations in crowdsourced annotation tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1637–1648.

Karlos, S.; Kanas, V. G.; Aridas, C.; Fazakis, N.; and Kotsiantis, S. 2019. Combining active learning with self-train algorithm for classification of multimodal problems. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–8. IEEE.

Khatua, A.; and Nejdl, W. 2022. Unraveling Social Perceptions & Behaviors towards Migrants on Twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 512–523.

Krippendorff, K. 2011. Computing Krippendorff's alpha-reliability.

Kulesza, T.; Charles, D.; Caruana, R.; Amershi, S. A.; and Fisher, D. A. 2019. Structured labeling to facilitate concept evolution in machine learning. US Patent 10,318,572.

Lloret-Pineda, A.; He, Y.; Haro, J. M.; Cristóbal-Narváez, P.; et al. 2022. Types of Racism and Twitter Users' Responses Amid the COVID-19 Outbreak: Content Analysis. *JMIR Formative Research*, 6(5): e29183.

Loureiro, D.; Barbieri, F.; Neves, L.; Espinosa Anke, L.; and Camacho-collados, J. 2022. TimeLMs: Diachronic Language Models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 251–260. Dublin, Ireland: Association for Computational Linguistics.

Mathew, B.; Dutt, R.; Goyal, P.; and Mukherjee, A. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, 173–182.

Miller, C.; Arcostanzo, F.; Smith, J.; Krasodomski-Jones, A.; Wiedlitzka, S.; Jamali, R.; and Dale, J. 2016. From brussels to Brexit: islamophobia, xenophobia, racism and reports of hateful incidents on twitter. *Demos*.

Muller, M.; Wolf, C. T.; Andres, J.; Desmond, M.; Joshi, N. N.; Ashktorab, Z.; Sharma, A.; Brimijoin, K.; Pan, Q.; Duesterwald, E.; et al. 2021. Designing ground truth and the social life of labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16.

Noel, T. K. 2020. Conflating culture with COVID-19: Xenophobic repercussions of a global pandemic. *Social Sciences & Humanities Open*, 2(1): 100044.

Pérez-Landa, G. I.; Loyola-González, O.; and Medina-Pérez, M. A. 2021. An explainable artificial intelligence model for detecting xenophobic tweets. *Applied Sciences*, 11(22): 10801.

Pitropakis, N.; Kokot, K.; Gkatzia, D.; Ludwiniak, R.; Mylonas, A.; and Kandias, M. 2020. Monitoring users' behavior: anti-immigration speech detection on Twitter. *Machine Learning and Knowledge Extraction*, 2(3): 192–215.

Plaza del Arco, F. M.; Strapparava, C.; Urena Lopez, L. A.; and Martin, M. 2020. EmoEvent: A Multilingual Emotion Corpus based on different Events. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1492–1498. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Rosenthal, S.; Farra, N.; and Nakov, P. 2017. SemEval-2017 Task 4: Sentiment Analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 502–518. Vancouver, Canada: Association for Computational Linguistics.

Rzepnikowska, A. 2019. Racism and xenophobia experienced by Polish migrants in the UK before and after Brexit vote. *Journal of Ethnic and Migration Studies*, 45(1): 61–77.

Saldana, J. 2012. *The Coding Manual for Qualitative Researchers*. SAGE Publications. ISBN 978-1-4462-7142-1.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, 1668–1678.

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*, 1–10.

Schulman, J.; Qu, H.; Lazer, D.; Perlis, R.; Ognyanova, K.; Baum, M.; Cadenasso, S.; Druckman, J.; Green, J.; Quintana, A.; et al. 2023. The COVID States Project# 97: Twitter, Social Media, and Elon Musk.

Sellars, A. 2016. Defining hate speech. *Berkman Klein Center Research Publication*, (2016-20): 16–48.

Sen, S.; Giesel, M. E.; Gold, R.; Hillmann, B.; Lesicko, M.; Naden, S.; Russell, J.; Wang, Z.; and Hecht, B. 2015. Turkers, scholars," arafat" and" peace" cultural communities and algorithmic gold standards. In *Proceedings of the 18th acm conference on computer supported cooperative work & social computing*, 826–838.

Shen, X.; He, X.; Backes, M.; Blackburn, J.; Zannettou, S.; and Zhang, Y. 2022. On Xing Tian and the Perseverance of Anti-China Sentiment Online. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 944–955.

Strossen, N. 2016. Freedom of speech and equality: Do we have to choose. *JL & Pol'y*, 25: 185.

Sylvia Chou, W.-Y.; and Gaysynsky, A. 2021. Racism and xenophobia in a pandemic: interactions of online and offline worlds. *American Journal of Public Health*, 111(5): 773–775.

Tahmasbi, F.; Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2021. "Go eat a bat, Chang!": On the Emergence of Sinophobic Behavior on Web Communities in the Face of COVID-19. In *Proceedings of the web conference 2021*, 1122–1133.

Tontodimamma, A.; Nissi, E.; Sarra, A.; and Fontanella, L. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1): 157–179.

Waseem, Z. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, 138–142.

Watanabe, H.; Bouazizi, M.; and Ohtsuki, T. 2018. Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection. *IEEE access*, 6: 13825–13835.

Wimmer, A. 1997. Explaining xenophobia and racism: A critical review of current research approaches. *Ethnic and racial studies*, 20(1): 17–41.

Zannettou, S.; Finkelstein, J.; Bradlyn, B.; and Blackburn, J. 2020. A quantitative approach to understanding online antisemitism. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 786–797.

Zhang, Z.; and Luo, L. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5): 925–945.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, and we discuss some of these aspects and potential concerns in the Discussion and Future Work section and the Ethics Statement.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract accurately reflects contributions of this paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes, we discuss and justify our methods in the Methodology section and address limitations in the Discussion and Future Work section.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, we address these as part of the Discussion and Future Work section.
   (e) Did you describe the limitations of your work? Yes, we discuss the limitations of this work primarily as part of the Discussion and Future Work section, but also in The Labeler Experience Section and the Ethics Statement.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, we address potential negative societal impacts as part of the Discussion and Future Work section and the Ethics Statement.
   (g) Did you discuss any potential misuse of your work? Yes, we address potential misuse of this work as part of the Discussion and Future Work section and the Ethics Statement.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, we discuss steps to mitigate negative outcomes of this research in The Labeler Experience, Discussion and Future Work sections as well as in the Ethics Statement.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes, we have read the ethics guidlines and made sure that the paper conforms to them.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A
   (b) Have you provided justifications for all theoretical results? N/A
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A
   (e) Did you address potential biases or limitations in your theoretical framework? N/A
   (f) Have you related your theoretical results to the existing literature in social science? N/A
   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A
3. Additionally, if you are including theoretical proofs...
   (a) Did you state the full set of assumptions of all theoretical results? N/A
   (b) Did you include complete proofs of all theoretical results? N/A
4. Additionally, if you ran machine learning experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes, the code to reproduce experimental results is shared on GitHub: https://github.com/khonzoda/Xenophobia-Meter-Project.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes, experiment details are presented as part of the published code.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes, the experiments were conducted on multiple seeds, and results are presented along with the standard deviation.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes, the details are included in the code release.
   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes, machine learning experiments serve as a demonstration of a use-case of the dataset.
   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes, although not in the context of machine learning experiments, we discuss the implications of labeling politically-laden datasets in detail in The Labeler Experience section.
5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...
   (a) If your work uses existing assets, did you cite the creators? Yes, please see the Dataset Analysis section.
   (b) Did you mention the license of the assets? Yes, we provide information about the specific HuggingFace model used.
   (c) Did you include any new assets in the supplemental material or as a URL? NA

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes, please refer to the Methodology and The Labeler Experience sections.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes, please refer to our Ethics Statement.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see **?**)? Yes, we make our dataset publically accessible through the Harvard Dataverse and provide detailed information about the dataset and its creation process.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes, the details about this dataset are presented both the Methodology and The Labeler Experience sections, as well as in metadata as part of the published dataset.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes, the full text of the interview protocol can be found here: https://docs.google.com/document/d/1Orxl-GVotdBSIopmAd3wTEldolhBsA-82HMNRVhLUf0/edit?usp=sharing. And also here is the link to the IRB-approved informed consent for the interview: https://cornell.ca1.qualtrics.com/jfe/form/SV_eKSKoF4VTWxb1CS.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes, see the full text of the informed consent form her: https://cornell.ca1.qualtrics.com/jfe/form/SV_eKSKoF4VTWxb1CS.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes, participants were offered a $15 Amazon gift card as compensation for participating in a 40-60 minute interview.

(d) Did you discuss how data is stored, shared, and deidentified? Yes, this is described in the informed consent form herer: https://cornell.ca1.qualtrics.com/jfe/form/SV_eKSKoF4VTWxb1CS.