“It's Not Just Hate”

"It's Not Just Hate": A Multi-Dimensional Perspective on Detecting Harmful Speech Online

Federico Bianchi
Stanford University
Stanford, California, USA

Stefanie Anja Hills
University of Stirling
Stirling, UK

Patricia Rossini
University of Glasgow
Glasgow, UK

Dirk Hovy
Bocconi University
Milan, Italy

Rebekah Tromble
George Washington University
Washington, DC, USA

Nava Tintarev
Maastricht University
Maastricht, The Netherlands

Abstract

Well-annotated data is a prerequisite for good Natural Language Processing models. Too often, though, annotation decisions are governed by optimizing time or annotator agreement. We make a case for nuanced efforts in an interdisciplinary setting for annotating offensive online speech. Detecting offensive content is rapidly becoming one of the most important real-world NLP tasks. However, most datasets use a single binary label, e.g., for *hate* or *incivility*, even though each concept is multifaceted. This modeling choice severely limits nuanced insights, but also performance. We show that a more fine-grained multi-label approach to predicting incivility and hateful or intolerant content addresses both conceptual and performance issues. We release a novel dataset of over 40,000 tweets about immigration from the US and UK, annotated with six labels for different aspects of incivility and intolerance. Our dataset not only allows for a more nuanced understanding of harmful speech online, models trained on it also outperform or match performance on benchmark datasets.

Warning: This paper contains examples of hateful language some readers might find offensive.

1 Introduction

Though once considered a problem driven primarily by reduced inhibitions in anonymous online spaces (Rösner and Krämer, 2016; Suler, 2004), offensive content has grown exponentially–to the point that many users no longer feel restricted by traditional conversational norms of tolerance and politeness, even when posting in their own names (Rossini, 2022). The pervasiveness of toxic discourse on social media in particular has helped sow the seeds of discord and hatred that harm the health and well-being of its targets and pose significant threats to the fundamental rights of individuals and social groups on the margins (Gelber and McNamara, 2016). Concerned about such outcomes, over the last two decades scholars and practitioners from a variety of fields have scrutinized online incivility and hateful speech. Those working in natural language processing, for example, have developed techniques to detect different types of offensive discourse, ranging from incivility to hate speech, while social scientists have focused extensively on the larger substantive effects of these phenomena.

Most computational approaches for detecting online toxicity are based on classifiers that predict the presence of a *single* main binary label (Basile et al., 2019; Stoll et al., 2020; Davidson et al., 2020, 2017), with some notable exceptions (Vidgen et al., 2021a,b; Mollas et al., 2022; Kennedy et al., 2022, inter alia).[1] However, while single-label binary approaches to harmful speech detection are conceptually tidy and tend to yield good predictive performance, they have major limitations. Most notably, such approaches *are unable to distinguish discourse that threatens democratic norms, values, and rights from expressions that are merely rude or impolite.* Prior work detecting incivility, for instance, has combined relatively harmless expressions that break traditional norms of polite speech–for instance, profanities and swearing–with discourse that is potentially more harmful, such as personal insults, stereotyping, or hateful speech (Stoll et al., 2020; Theocharis et al., 2016; Tang and Dalzell, 2019). Binary approaches to toxic and offensive content detection oversimplify these complex concepts, and ultimately undermine

[1] We refer to the works by Vidgen and Derczynski (2020) and Poletto et al. (2021) for in-depth surveys.

researchers' and practitioners' ability to understand potential harms and evaluate what content should receive most focus and intervention, including for the purposes of content moderation.

To address these open issues, we show that our *multi-label approach* rooted in insights drawn from social science is not only potentially more insightful, but also improves performance of detection models. In contrast to most previous work, we build upon a *conceptual model that disentangles uncivil from intolerant online discourse* (Rossini, 2022). The resulting labels can meaningfully distinguish discourse that is simply rude or offensive (*incivility*) from expressions that threaten democratic norms and values, such as equality, diversity, and freedom (*intolerance*).

We collect a dataset of more than 40,000 US- and UK-based tweets related to the topic of immigration, and annotate these tweets for four sub-types of *incivility* (profanities, insults, outrage, character assassination) and two sub-types of *intolerance* (discrimination, hostility). We refer to this dataset as *Not Just Hate* (NJH). We then fine-tune large pre-trained language models and show that these labels can be predicted with consistently good performance. We compare our results to other benchmark hate speech datasets to produce more insights about the dataset we introduce. Models trained on our data match or outperform state-of-the-art performance on those datasets.

Our approach, annotation methodology, and dataset can help the future development of automated harmful online speech detection, and foster a more nuanced understanding of the distinctive types of discourse that constitute online toxicity and abuse. Data and additional details on the annotation are available on OSF.[2] Details are also available on the GitHub repository.[3]

Contributions We describe a novel perspective on harmful online speech detection. We describe in detal our annotation pipeline and we release a dataset, NJH, of just over 40,000 tweet ids annotated with four sub-types of incivility (profanities, insults, outrage, character assassination) and two sub-types of intolerance (discrimination, hostility). We show that our data set generalizes to various types of offensive language and reaches state-of-the-art performance.

2 Data

Collection We collected our dataset via the Twitter Enterprise API, downloading over 150 million tweets over the course of 2020-2021. We selected the keywords used to collect these tweets in a multistage process. Beginning with a list of 30 keywords and phrases commonly associated with immigration in the US and/or UK (e.g., immigration, immigrant, refugee, illegals), we drew a random sample of tweets containing those words from the public streaming API, produced a list of words commonly co-occurring with the seeded terms, and qualitatively analyzed that list to identify the appropriateness of each additional keyword or phrase. We performed the same process with a series of subreddits related to immigration, carefully curating the list of subreddits to represent both pro- and anti-immigration sentiment, as well as a variety of immigration subtopics (e.g., subreddits dedicated to Asian or Muslim immigrants/immigration, refugees, etc.). Please see the Appendix for the final set of keywords used to collect tweets.

Annotation Our annotation approach was based on quantitative content analysis (Krippendorff, 2018), a social scientific method used by communication scholars to interpret meaning in textual data at scale. The annotation guidelines were broadly inspired by those of Rossini (2022), which we augmented and adapted with examples that were context- and country-specific to capture the nuances of immigration debates across the US and the UK.

The annotators were ten undergraduate researchers from the University of Liverpool (UK) and Syracuse University (US). We trained the students on the annotation guidelines until they achieved a satisfactory inter-annotator reliability score for two consecutive weeks (Krippendorff's α of 0.68 or above) and Gwet's AC1 (of .6 or above). We used these to correct for expected issues in the data quality–Krippendorff's α penalizes data scarcity, which is a problem in some of our labels, while Gwet's AC1 corrects for the probability that the annotators agree by chance (this is more likely for difficult annotation tasks such as this one). We continuously monitored the quality of the annotation by measuring inter-annotator reliability on a monthly basis.

How to aggregate annotation scores is a central

[2] https://osf.io/gxvsj/?view_only=12197981e47a47239a6f80c62db84b14

[3] https://github.com/vinid/not-just-hate

Label	Number
Outrage (O)	6,743
Insults (I)	5,040
Profanity (P)	4,074
Char. Assassination (C)	3,436
Discrimination (D)	10,437
Hostility (H)	2,699
No Label	22,007
Total Labels:	57,139

Table 1: 40,136 tweets. *No Label* = tweets with no labels. *Total Labels* includes tweets with multiple labels.

Example	Labels
@user and i don't trust dead liberal voters and illegals voting for more handouts. EC is where its at. Screw them lib frauds.	PROFANITY, INSULTS, CHARACTER ASSASSINATION, OUTRAGE, DISCRIMINATION
FUCK ICE!	PROFANITY, INSULTS

Table 2: Examples of tweets in our dataset with the respective labels. Text altered to preserve privacy.

and open problem in machine learning (Gordon et al., 2022; Davani et al., 2022). We divided the trained annotators into teams to individually annotate tweets from their respective country. We compared their individual annotations, and they met to discuss and adjudicate any disagreements–i.e., all annotators had to agree on a best label(s). We opted for this rigorous annotation process instead of a simple majority rule due to the complexity of the phenomena we investigate. While prior efforts on incivility and hate speech detection have relied on approaches such as 'majority rule' to determine labels based on crowdsourced annotations, leading to biases in interpretation that may undermine the quality of annotation, we leverage social scientific content analysis techniques to develop a rigorous annotation pipeline to ensure the quality of our dataset.

We randomly selected samples for annotation, stratifying by location. We restricted the data to the US and UK via geolocated tweets and user location, and use simple trigram tf-idf based classifier to distinguish between the UK, US, and Other (Macro F1=0.82, albeit we have updated our algorithm several times during the course of the research project) trained on 3 million tweets. Given the relative scarcity of intolerant discourse in particular, we used Jigsaw's Perspective API to over-sample potentially toxic tweets and increase the presence of some labels on approximately 60% of the annotated sample. Table 1 shows the number of tweets tagged with each label in the dataset.

3 Experiments

3.1 Datasets

HateEval (2,971 examples, 2 labels) This dataset was introduced during the SemEval2019 (Basile et al., 2019) challenge, and it is particularly well-suited for our task because it has been built around the topics of immigration, which we cover, and women. Note that we use the samples and splits provided by the recently introduced TweetEval benchmark (Barbieri et al., 2020).

HateCheck (421 examples, 2 labels) This dataset offers a checklist to evaluate and stress test different hate speech detection models. Examples in HateCheck address linguistic features like spelling variation or negation (e.g., "There is no hatred in my heart for you immigrants" is a non hateful example). HateCheck is an excellent dataset to verify how a model behaves when encountering these features. However, examples are manually generated and cannot be considered *real* examples of harmful speech. HateCheck covers different targets of harmful speech; however, we extract only the subset containing data related to immigration (Röttger et al., 2021).

Data Preparation We lighty pre-process the data: we replace user tags with an anonymous *USER* and links with *HTTPURL*.[4] This is done to prevent the model from learning spurious patterns regarding the occurrence of specific users. We split our dataset into 3 sets: train (85%, 34,115 examples), dev (7.5%, 3011 examples), and test (7.5%, 3011 examples).

3.2 Models and Training

We use RoBERTa (Liu et al., 2019) (base and large) and BERTweet (Nguyen et al., 2020) (base and large). BERTweet is a RoBERTa model additionally pretrained on Twitter data. Each model is fine-tuned three times, we report averaged results.

While we annotate for six labels, during training we let the classifier also predict the *supertypes*

[4]Note that this is a common approach for Twitter data in large language models (Nguyen et al., 2020).

Model	**NJH** (Macro-F1)	**HateEval** (Macro-F1)	**HateCheck** (Accuracy)	**Avg.**
Roberta-base	0.74 ± 0.00	0.63 ± 0.01	0.54 ± 0.03	0.64
Roberta-large	0.76 ± 0.00	**0.65** ± 0.02	0.69 ± 0.03	0.70
BERTweet-base	0.74 ± 0.01	0.64 ± 0.01	0.47 ± 0.06	0.62
BERTweet-large	**0.77** ± 0.00	0.63 ± 0.02	**0.71** ± 0.02	0.70
Best Other	–	0.52 ± 0.00	**0.71** ± NA	–

Table 3: Comparison of various models trained on our data and tested on several data sets. On HateCheck we evaluate on Accuracy as in the original paper by Röttger et al. (2021).

of the labels: *incivility* for PROFANITY, INSULTS, CHAR. ASSASSINATION, AND OUTRAGE and *intolerance* for DISCRIMINATION AND HOSTILITY. The total number of labels to predict is thus eight.

See the Appendix for the hyper-parameters used. We run a small parameter selection pipeline testing different learning rates {5e-4, 5e-5, 5e-6}. All models are trained for five epochs, but we select the model that performs best at validation time; validation is run every 200 steps. The learning rate of 5e-5 was the best performing, but we report all the results in the Appendix. We test all the trained models on the test portion of NJH and on the test set from HateEval and on HateCheck. We also report the best results on HateEval and HateCheck as described in the papers (Barbieri et al., 2020; Röttger et al., 2021) (marked as *Best Other* in Table 3).

Both HateEval and HateCheck focus on the binary hate/not hate annotations. To adapt to the binary setting, since models trained on NJH are multi-label, we consider a tweet *hateful* if the model predicts one or more of the following labels: *hostility*, *discrimination* and/or *intolerance*. Note that because these datasets have been annotated with different definitions of *hate*, results might not always be perfectly comparable.

3.3 Results

Table 3 shows the results of our fine-tuned models on the different datasets.

NJH Performance for all models is consistently above 0.70 Macro F1. Figure 1 shows the results per label for the best model, BERTweet-large.

HateEval Models trained on our dataset achieve better Macro-F1 on HateEval than previous work (Barbieri et al., 2020), reaching results comparable to those in the challenge (Barbieri et al., 2020). Best Other is described by (Barbieri et al., 2020), a RoBERTa model, fine-tuned on the HateEval training data (Basile et al., 2019).

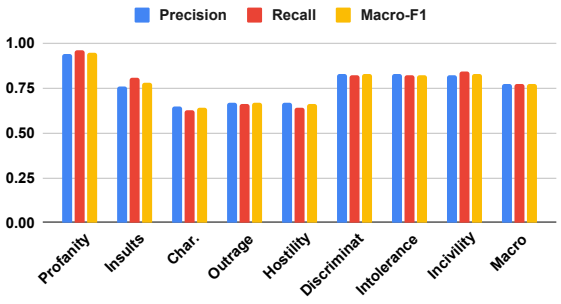

Figure 1: F1 results per-label and Macro F1.

HateCheck Performance on the immigration subset suggests that the base models do not learn as well as the large ones. However, BERTweet-large reaches a comparable performance to the best model. Best Other is the described by (Röttger et al., 2021), a BERT model fine-tuned on the Twitter dataset by Davidson et al. (2017).

3.3.1 Comparison with other Models

We compare the performance of other pre-trained models on NJH. This serves as a proof of concept that popular approaches do not capture the entire spectrum of incivility and intolerance proposed by our data. We use two models: one trained on data collected from different rounds of human-machine interaction to train better models (Vidgen et al., 2021b).[5] and one[6] that has been trained on HateEval data (Barbieri et al., 2020; Basile et al., 2019). We then compute the F1 score between the predictions of the models and each of our labels. Figure 2 shows the results. Both models seem to be able to effectively capture DISCRIMINATION; however they do not capture stronger harmful speech such

[5] https://huggingface.co/facebook/roberta-hate-speech-dynabench-r4-target

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

as HOSTILITY. In general, the models do not seem to capture **Incivility** (PROFANITY, INSULTS, OUTRAGE, or CHARACTER ASSASSINATION). This latter result can be expected since the models have been trained to predict just hateful content. Overall, our findings suggest that there is a need for models that can effectively distinguish different aspects of offensive and potentially harmful discourse.

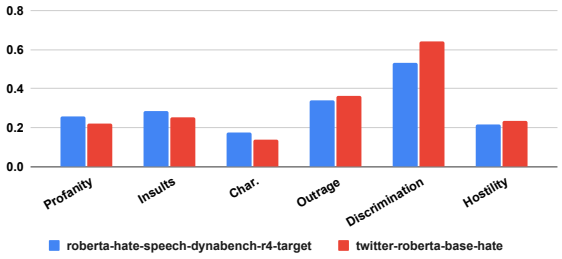

Figure 2: F1 of existing models on our dataset.

4 Conclusions

We suggest that a more fine-grained approach improves offensive and potentially harmful speech detection online, and, crucially, can allow for a better understanding of the spectrum of online toxicity. Our approach can successfully disentangle incivility from likely more harmful cases of intolerant content, allowing scholars and practitioners to better understand and detect discourse that undermines broader democratic norms, values, and rights (Rossini, 2022). We introduce a dataset of just over 40,000 tweets, annotated with six labels. Models trained on our dataset can predict labels with good confidence and perform well on other benchmark datasets.

Acknowledgments

DH is a member of the Bocconi Institute for Data Science and Analytics (BIDSA). This work was partially conducted while FB was a member of BIDSA.

We acknowledge financial support for this research in the form of a gift from Twitter, Inc.

Ethical Considerations

We anonymized Twitter handles as part of the data pre-processing, and any tweet text provided as an example here (i.e., in Table 2) has been edited to further preserve anonymity. The NJH dataset is shared in dehydrated format, i.e., as tweet IDs only, in full compliance with Twitter's Developer Policy.[7] We are aware that our dataset, if reconstructed, contains potentially harmful content. Though we use this content to help better examine, understand, and help mitigate the harms of online hate, we recognize that these tweets could be used for darker purposes. As any tweet successfully rehydrated from our list of tweet IDs remains in the public domain, we have assessed that the benefits of sharing this dataset outweigh the risks.

Limitations

HOSTILITY is an aggregated label that encompasses the originally annotated labels of HATEFUL SPEECH, DEHUMANIZATION, SERIOUS THREAT-PERSONAL ABUSE-HARASSMENT, and DEMOCRATIC THREAT. These labels did not yield enough annotated tweets to remain a part of our multi-label classifier in their own right. Although HOSTILITY is a suitable label that groups the original labels on the basis of hostile intent and/or effect, as well as the nature of their targets, we cannot claim that annotators would have interpreted tweets in the same way if they had annotated for HOSTILITY rather than following codebook guidance for each individual label. For full transparency, we are releasing all original, ungrouped, annotations for this dataset.

Although we are also releasing the unaggregated annotations alongside the aggregated annotations, it must be noted that the nature of our adjudication process means that our aggregated labels cannot be directly reproduced from the unaggregated ones. This is because we opted for a significantly more rigorous approach that involved annotators meeting to discuss and resolve every single annotation disagreement under expert supervision. At times, these discussions might have led to the decision to annotate for labels that previously no individual annotator had identified. Although more rigorous and fair - through ensuring every annotator's views are heard - this process has the downside of being less transparent retrospectively, as the discussion and decision-making that took place in these adjudication meetings cannot be easily documented and the final aggregated annotations ultimately only present the outcome of the process and not the process itself.

An additional limitation to replicability stems from the decay of tweets over time, wherein deleted tweets and/or tweets from suspended, deleted, or

[7] https://developer.twitter.com/en/developer-terms/policy

newly private accounts cannot be rehydrated based on their tweet IDs. This is a common event in all Twitter datasets (Tromble and Stockmann, 2017), and is particularly prevalent in hate speech datasets, where users are often suspended and individual tweets removed from the platform. We currently estimate approximately 25% of tweets in this dataset are no longer accessible for rehydration.

References

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.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a New Classifier for Automated Identification of Incivility in Social Media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. *arXiv:1703.04009 [cs]*. ArXiv: 1703.04009.

Katharine Gelber and Luke McNamara. 2016. Evidencing the harms of hate speech. *Social Identities*, 22(3):324–341.

Mitchell L. Gordon, Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. Jury learning: Integrating dissenting voices into machine learning models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, pages 1–16.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Patrícia Rossini. 2022. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, 49(3):399–425.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Leonie Rösner and Nicole C. Krämer. 2016. Verbal Venting in the Social Web: Effects of Anonymity and Group Norms on Aggressive Language Use in Online Comments. *Social Media + Society*, 2(3):205630511666422.

Anke Stoll, Marc Ziegele, and Oliver Quiring. 2020. Detecting Incivility and Impoliteness in Online Discussions. *Computational Communication Research*, 2(1):109–134. Number: 1.

John Suler. 2004. The Online Disinhibition Effect. *CyberPsychology & Behavior*, 7(3):321–326.

Yiwen Tang and Nicole Dalzell. 2019. Classifying hate speech using a two-layer model. *Statistics and Public Policy*, 6(1):80–86.

Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting

with party candidates. *Journal of Communications*, 66(6):1007–1031.

Rebekah Tromble and Daniela Stockmann. 2017. Lost umbrellas: Bias and the right to be forgotten in social media research. *Internet research ethics for the social age: New challenges, cases, and contexts*, pages 75–91.

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, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021a. 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.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021b. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682.

A Dataset Details

A.1 Data Statement

The data we share is is composed by tweet ids and does not directly contain personal information of the individual; however upon reconstruction, it shows tweet and author if it is still publicly available. The reconstructed data contains harmful messages. Annotators were all English native speakers.

A.2 Twitter Keyword Used

In the following we include the list of keywords used to extract the tweets.

"the wall" OR "fuck ice" OR undocumented OR illegals OR "an illegal" OR "muslim ban" OR "travel ban" OR refugee OR asylum OR #wherearethechildren OR "child cage"∼3 OR "children cage"∼3 OR #wall OR daca OR #dreamer OR "sanctuary city" OR "sanctuary cities" OR "baby cage"∼3 OR "babies cage"∼3 OR "abolish ice"∼3 OR "ice raid"∼3 OR #abolishice OR #muslimban OR #nobannohate OR #refugeeswelcome OR #refugeeswelcomehere OR ms-13 OR "build the wall" OR #buildthewall OR “ms- 13” OR "ms 13" OR deport OR citizenship OR birthright OR "illegal alien"∼3 OR ms13 OR #secureourborders OR #familiesbelongtogether OR #closethecamps OR #defenddaca OR #nocamps OR #noban OR #savedaca OR #immigrationreform OR #uslatino OR #openborders OR "open border"∼3 OR "kid cage"∼3 OR "kids cage"∼3 OR USCIS OR #proimmigration OR "farm worker" OR "farm workers" OR farmworker OR #farmworkerjustice OR #immigrationpolicy OR migrant OR amnesty OR #noamnesty OR #imalreadyhome OR #proopenborders OR #immigrantnation OR #nohumanisillegal OR #welcomeimmigrants OR "no human is illegal" OR #MSW52170 OR #immigrantsmatter OR #immigrantrights OR "learn to speak English" OR "steal jobs" OR "job stealing"∼3 OR "mexican border" OR "mexico pay"∼3 OR visa OR "chain migration" OR "dream act" OR "merit based" OR citizen OR foreigner OR "foreign national" OR "trump wall"∼3 OR "mexico policy"∼3 OR "foreign worker"∼3 OR "human trafficking"∼3 OR xenophobe OR xenophobia OR schengen OR "british national" OR #BNO OR "free movement"

B Model Training

Results on validation set are available in Figure 4

Model	5e-5	5e-6
Roberta-base	0.74 ± 0.01	0.67 ± 0.01
Roberta-large	0.76 ± 0.00	0.74 ± 0.00
BERTweet-base	0.73 ± 0.01	0.54 ± 0.02
BERTweet-large	0.77 ± 0.00	0.76 ± 0.00

Table 4: Results on the NHJ validation set. *models with a learning rate of 5e-4 obtained very low performance and were not working on the data.

Figure 5 shows the parameters used to train the models (excluding learning rate that is a parameter we found with grid search).

Param	Value
Batch Size	64
Learning Epochs*	5
Optimizer	AdamW
Betas	0.9 and 0.999
Max Length	80

Table 5: The main parameters we used to run the models. *While epochs are 5, we remark that we are running a step-wise evaluation. Batch size is achieved thanks to the use of gradient accumulation (8 steps)