THOS: A Benchmark Dataset for Targeted Hate and Offensive Speech

Saad Almohaimeed [1] **Saleh Almohaimeed** [1] **Ashfaq Ali Shafin** [2] **Bogdan Carbunar** [2] **Ladislau Bölöni** [1]

Abstract

Detecting harmful content on social media, such as Twitter, is made difficult by the fact that the seemingly simple yes/no classification conceals a significant amount of complexity. Unfortunately, while several datasets have been collected for training classifiers in hate and offensive speech, there is a scarcity of datasets labeled with a finer granularity of target classes and specific targets. In this paper, we introduce THOS, a dataset of 8.3k tweets manually labeled with fine-grained annotations about the target of the message. We demonstrate that this dataset makes it feasible to train classifiers, based on Large Language Models, to perform classification at this level of granularity.

Warning: *Due to the nature of the research, the body of this paper contains offensive language.*

1. Introduction

Despite the obvious benefits of social media in connecting and entertaining people and providing a forum for public conversation about topics of interest to society, many social media platforms are plagued by harmful content (hate speech, cyberbullying, offensive statements). In recent years, it has been both a societal expectation, good business practice and sometimes a legal requirement for social media companies to filter harmful content. Due to the large number of messages on social media platforms, such a filtering can only be achieved through automatic classification.

Early approaches to the task of identifying abusive content treated this as a classification into two or, perhaps, several classes, such as hate, offensive and neither (Davidson et al., 2017; Waseem & Hovy, 2016). Out of these, identifying offensive content using a dictionary-based approach is trivially simple. However, it has been recognized early on that the labels of offensive and hate speech need to be considered separately, as not every message containing swear words is hate speech, while a message without offensive words can be hateful and even incite violence.

However, the seemingly straightforward classification of messages into two classes (harmful vs. not) obscures the fact that these are complex ideas that need to be identified and treated in the societal context. To mock somebody's favorite sports team is not hate speech; however, to mock somebody's religion, is. Making fun of somebody's sunburn after a beach party is not hate speech, but in most contexts, joking about other people's skin color can be hateful. Finally, in many societies, the severity of an inappropriate statement depends on the target: for instance, negative statements directed at vulnerable minorities might be judged differently compared to similar messages targeting majority populations.

Of course we could hope that a dataset of messages simply labeled as hate speech or not would allow machine learning models to infer features that implicitly take into account the target. However, it's unlikely that such an emergence will occur with small- or medium-sized datasets. Furthermore, we should also consider the fact that for many applications it is preferable to have a classifier that explicitly identifies the target of the message. For instance, the percentage of hate messages targeting a certain ethnic or religious group might be used to assess dangerous societal trends.

To create the classifier algorithms for target detection, it is necessary to have well-annotated datasets at the appropriate level of resolution. While some datasets that include the target in the annotation exist (Mathew et al., 2021; Ousidhoum et al., 2019), they cover only some of the possible ways in which the target can be classified at a finer resolution. A logical next step would be to consider separately the *topic* of the message (such as religion, politics or country), and within a topic, a more specific *subtopic* such as China or India. Apart from the target of the message, the dataset would also be labeled as to whether it contains offensive words (OFF) and whether, in the human annotator's opinion, it contains hate speech (HS).

[1]Department of Computer Science, University of Central Florida, Orlando, FL 32816, USA [2]Florida International University, Miami, FL 33199, USA. Correspondence to: Saad Almohaimeed <almohaimeed.saad@gmail.com>, Ladislau Bölöni <ladislau.boloni@ucf.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, DMLR workshop, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

The THOS dataset described in this paper aims to provide such a labeled collection of messages for training and validating fine-grained classifiers. In the remainder of this paper, we describe the structure of the dataset and the procedure used to create it. We also describe some initial results that use the dataset to train fine-grained classifiers based on large language models.

The main contributions of this paper are as follows:

- We created and manually labeled a Twitter dataset (THOS)[1] for hate and offensive speech, which exhibits a finer level of granularity in its labeling compared to previous such datasets. For instance, THOS includes labels identifying the targets as general topics such as country, religion, ethnicity, politics or individuals, as well as specific subtopics such as the USA, China, Christian, Jewish, or Muslim.
- We analyze the dataset with regards to the co-occurences of its labels. One of our findings confirms the conjecture that relying solely on offensive language detection, such as employing dictionary-based approaches, is insufficient in effectively tracking hate speech. Surprisingly, approximately 55% of instances classified as hate speech do not contain offensive language, while a significant proportion of offensive language does not meet the criteria for hate speech.
- We show that the THOS dataset, though of a moderate size, is sufficient to train fine-grain classifiers. We demonstrate that fine-tuning large language models (LLMs) on THOS can yield classifiers of sufficient accuracy for tracking trends of hate speech against specific subgroups on social media.

2. Related Work

As hate speech against specific groups emerged as a major problem on social media, a growing community of researchers started to focus on its detection and classification. These projects were motivated through various objectives, ranging filtering offensive language and the study of trends in hate against specific communities to complex problems such as the identification of hate campaigns organized by operatives of political organizations.

The first generation of such systems used traditional machine learning techniques and focused on classifying the messages in a handful of classes. From the beginning, the creation and labeling of datasets were the primary prerequisites for the classification systems. (Waseem & Hovy, 2016) developed a dataset named NAACL_SRW_2016, consisting of 16k tweets, for multi-class classification focusing on sexism, racism, and normal speech. They explored the impact of various features, including the geographic location of the tweet and the gender of the user. The authors conducted experiments using logistic regression as the machine learning model and evaluated the model's effectiveness using F1-score, precision, and recall metrics. Their findings revealed that the best performance was achieved when utilizing character four-gram features combined with the gender feature.

(Davidson et al., 2017) have created a Twitter dataset (24.8k tweets) that considers the majority of the annotators' voting as multi-class labeling (i.e. hate, offensive, or neither). The authors use 1 to 3-gram features weighted by TF-IDF for the Part-of-Speech tags using Natural Language Toolkit (NLTK) along with the sentiment's score of each tweet using VADER (Hutto & Gilbert, 2014). They experimented with different machine learning models and with the best performance optained by logistic regression with L2 regularization. An interesting finding of the paper was that sexist content was mostly classified as offensive while homophobic and racist content was mostly detected as hate speech.

(Waseem, 2016) examined the impact of annotator knowledge by extending the NAACL_SRW_2016 dataset with 6.9k tweets labeled by four different annotators. The author found that the performance of the expert's annotation model was comparable to NAACL_SRW_2016, but including the labels of amateur annotators lowered the performance compared to the NAACL_SRW_2016 dataset on both binary and multi-class classification.

One of the focus of the second generation of projects was the recognition that the *target* of the speech also holds significant relevance in hate speech detection.

(Zampieri et al., 2019) created an offensive language identification dataset consisting of 14.1k tweets that consider three levels of aspects (i.e. speech type, offense type, and target type). They created a baseline by experimenting with two types of embedding, FastText[2] and custom on a linear SVM using word unigram, pre-trained BiLSTM and CNN models by (Kim, 2014; Rasooli et al., 2018).

(Ousidhoum et al., 2019) created a multi-lingual (Arabic, English, and French) hate speech dataset of 13k tweets that considers the aspects of directness, hostility, target, group and annotator. The dataset was used to train classifiers using both legacy approaches (Bag-of-Words with logistic regression) and deep learning based ones (BiLSTM using the (Ruder et al., 2017) learning network). The authors tested the model with different settings according to the number of tasks (number of labels to predict) and languages.

[1] https://github.com/mohaimeed/THOS

[2] fastText - https://github.com/facebookresearch/fastText

They found that single-task single-language (STSL) model outperforms the multi-task multi-lingual (MTML) model in *directness* aspect while MTML performs better than STSL in the other aspects.

(Fortuna et al., 2020) analyzed six public abusive language datasets by using FastText to investigate the compatibility between different datasets and their definitions of classes. The analysis found that there is a significant incompatibility between the definitions of classes by different datasets.

(Mathew et al., 2021) had extended previous datasets from (Davidson et al., 2017; Ousidhoum et al., 2019) and (Mathew et al., 2019) to create the HateXplain dataset containing 20k Twitter and Gab annotated samples. The dataset considers hate and offensive as the two classes of abusive language, and also includes the target group. The dataset also contains an unusual label listing the tokens that represent the position of the abusive content in the text (i.e. *rationales*). The dataset was used to train CNN-GRU, BiRNN and BERT models and word embedding by GloVe for non-BERT models. The performance was evaluated through a variety of metrics such as accuracy, macro F1 and AUROC, the prediction bias (i.e. AUC) by (Borkan et al., 2019) and finally the explainability metrics (i.e. Plausibility and Faithfulness). It was found that the rationale labels can improve both performance and bias metrics, but not the explainability metric.

Table 1 summarizes the size, granularity and annotator expertise of the datasets discussed above, positioning THOS in the context of this work.

3. Dataset creation

3.1. Motivation and general principles

The goal of the THOS dataset is to support research for the creation and evaluation of accurate and detailed classifiers of harmful speech on the internet. As Table 1 shows, THOS seeks to address an existing gap in the collection of hate speech datasets by focusing on higher resolution labels with respects to the target and a moderate size, but higher quality labeling performed by experts. Like the majority of comparable datasets, THOS includes explicit labels of hate speech (HS) and offensive language (OFF). We will call the combination of HS and OFF as *explicit hate speech* (HSOFF), while the presence of the HS label without the OFF label as *implicit hate speech*. The dataset also includes labels identifying the target of the message at two levels of the resolution: the broader, more general topics (TPC) and the more specific sub-topics (STPC). To ensure the accuracy and reliability of the annotations, domain experts were engaged in the annotation process, rather than relying on crowd-sourcing. This approach ensures the uniformity of the annotations, resulting in a more robust and accurate dataset (Waseem, 2016).

3.2. Content of the dataset

The THOS dataset contains 8,282 tweets, with 46 data fields, which include the text of the tweet, 5 meta-data fields relating to the identity of the tweet and annotator and 40 boolean labels. The labels can be grouped into four classes.

The **Hate** label answers the questing whether the speech or text promotes or incites hatred toward an explicitly or implicitly targeted individual or group. The **Offensive** indicates whether the speech or text contains offensive (swear) words regardless whether it incites or promotes hate.

The third class comprises six labels that encode TPCs. Based on previous research, which identified ethnicity, origin(country), race and religion (Fortuna & Nunes, 2018; Relia et al., 2019; Scheitle & Hansmann, 2016) as the most common target groups in hate speech, these were added as available TPC selections. We decided to group the ethnicity/race in a single TPC. In the pilot phase of the dataset creation, we also identified as large classes of hate targets *political groups* and *individuals*. Finally, a TPC representing *other* was added to capture targets outside these groups. These labels are not intended to be mutually exclusive: for instance, 1166 tweets in THOS fall under two TPCs, while 66 under three TPCs.

The fourth class of labels in the THOS dataset are the 31 labels describing STPCs. As shown in Table 4, STPCs can be viewed as a second level of classification hierarchy after TPCs. However, there is also a qualitative difference: while TPCs delineate targets in terms of high level *concepts*, most STPCs correspond to *concrete* instantiations of those concepts, such as specific countries, ethnicities or political orientations.

Table 6 shows several examples from the dataset and the associated labels.

3.3. Data acquisition and labeling

The tweets forming the THOS dataset were collected through the Twitter API using the Tweepy[3] library. To avoid collecting tweets that don't fit under any of the TPCs, we used keyword search to look up tweets that are relevant to **countries** *(i.e. USA, UK, China, and India)*, **religions** *(i.e. Muslim, Christian, and Jewish)* as well as **ethnicities** and races *(i.e. Black, White, Asian, Arab, European, African, Hispanic, and Latino)*. The keywords were combined with prefixes such as *all, all of the, most, most of the some, some of the*. We collected a total of 31k tweets using the process.

To align the THOS dataset with the existing datasets, we

[3]Tweepy - https://www.tweepy.org/

Table 1. Dataset Comparison (existing work)

Dataset	Speech Type		Target			Annotator		Size
	Hate	Offensive	TPCs	STPCs	Indvs	Expert	Crowdsource worker	
(Waseem & Hovy, 2016)		✓				16,907		16,907
(Waseem, 2016)		✓				unknown	6,909	6,909
(Davidson et al., 2017)	✓	✓					24,783	24,783
(Zampieri et al., 2019)		✓	not explicit		✓	300	14,100	14,100
(Ousidhoum et al., 2019)	✓	✓	religion, sex, gender, origin, disability	15		300	13,000	13,000
(Mathew et al., 2021)	✓	✓		23			20,148	20,148
THOS (this paper)	✓	✓	country, religion, ethnicity, politics	31	✓	8,282		8,282

Table 2. Speech Type Distribution

Speech Type	# of samples	Percentage %
Hate only	1878	22.7
Offensive only	933	11.3
Hate and Offensive	1523	18.4
Normal	3948	47.6

Table 3. TPCs Distribution

TPC	# of shared samples	# of samples	total
Country	914	544	1458
Religion	384	394	778
Ethnicity/Race	1024	810	1834
Politics	125	42	167
Other	1303	1771	3074
Individuals	1388	2281	3669
Total	5138	5842	-

started by creating the definitions for hate speech and offensive speech drawing on the ones used by other datasets in the literature. Starting with these definitions, we proceeded with a pilot phase, during which 280 sample tweets have been carefully annotated. These annotated samples were then provided to the expert annotators for study, revision, and feedback. Using this feedback we created a rules and process document (RPD)[1] to serve as a guideline for the experts to resolve potential ambiguities in the content.

The annotation of the dataset was carried out by five experts. Each expert was assigned a 3k tweets to annotate, out of which they were required to submit 1.5k tweets. For THOS, the quality of the annotation was prioritized over coverage, thus experts were allowed to skip tweets that are unclear or meaningless, to enhance the overall accuracy of the annotations. The annotators reported an average annotation rate of 80 tweets/hour (45s/tweet). After every 200 submitted tweets, the annotations were reviewed and checked for conformance to the RPD.

As a note, the classification of hate speech did not check whether the tweet refers to a fake news or not, as the fact-checking of the tweet would be highly expensive.

3.4. Comments on the class distribution in the dataset

As we discussed before, the THOS dataset was designed as a tool for training and validating hate speech classifiers. The tweets we annotated were extracted using keywords that are known to refer to targets frequently subjected to hate speech. The objective was to create a rough balance between tweets with harmful content and those without. As Table 2 shows, about half of the messages are marked as hate, offensive or both. We need to emphasize that this is not intended as an accurate representation of the messages in Twitter.

Likewise, we sought to achieve a roughly equal distribution of the TPCs (see Table 3) and STPCs (see Table 4). However, the distribution over these classes proved to be more uneven. At the TPC level, for instance, we have a

Table 4. STPCs Distribution

	STPC	# of shared samples	# of samples	total
Country	USA	362	186	548
	UK	66	34	100
	China	232	80	312
	India	113	48	161
	Other_Country	401	115	516
	Undefined_Country	4	2	6
Religion	Muslim	126	105	231
	Christian	126	169	295
	Jewish	154	110	264
	Other_Religion	14	-	14
	Undefined_Religion	1	-	1
Ethnicity/Race	White	314	124	438
	Black	240	78	318
	Arab	127	94	221
	African	88	54	142
	European	101	73	174
	Asian	203	104	307
	Hispanic	155	86	241
	Other_Ethnicity	163	81	244
	Undefined_Ethnicity	-	1	1
Politics	Republican	37	18	55
	Democrat	35	9	44
	Other_Politics	66	12	78
	Undefined_Politics	-	-	-
Other	People	24	66	90
	Male	168	252	420
	Female	288	353	641
	Other_Other	1008	828	1836
	Undefined_Other	97	164	261
Indv	Someone	1148	1889	3037
	Undefined_Individual	385	329	714
	Total	6246	5464	-

significantly smaller number of tweets in the Politics TPC, as the method for searching for tweets did not explicitly search for this topic. For the STPCs, a substantial number of tweets were classified under the labels prefixed with "Other", denoting values specified but outside the list of the enumerated ones, and the "Undefined", where the topic implied the existence of a particular STPC that, however, was not defined in the tweet.

Lastly, we did not label STPCs with specific individuals. Thu, here we have only two classes - when an individual was named (the STPC Somebody) and when an individual is implied but not identifiable from the tweet (the STPC Undefined Individual).

Table 5. HSOFF and TPCs correlation

TPCs	Normal	HS	OFF	HSOFF
Country	615	426	138	279
Religion	314	211	71	182
Ethnicity	750	394	252	438
Politics	55	82	0	30
Individual	1927	901	334	507

4. Experimental Study

In this section, we verify that the THOS dataset is suitable for its intended purpose, namely, serving as training and validation data. We anticipate that the use of THOS will improve the detection of various types of harmful content on social media in at least two ways. First, it should enable fine-grain classification of the various types of hateful and offensive speech, identifying the various TPCs and STPCs. This is important for the social media companies and society in general, because different targets might represent different levels of risk, and require different societal responses. Second, the fine-grained annotations of the training data might help classifiers develop the appropriate inductive bias that would allow them to achieve a more accurate classification, even if the desired output is only a two-way or four-way classification along the hate and offensiveness axes.

The development of such new generations of classifiers are beyond the scope of this paper. However, it is incumbent upon us that, to validate the value of the dataset, to verify whether it can serve as appropriate training and validation data for such classifiers.

4.1. Experimental setup

The current state-of-the-art in text classification relies on the use of large language models (LLMs). One frequently used approach is based on starting from a foundation model and fine-tuning it to predict a specific label through a linear layer. LLMs such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and the Twitter-specialized BERTweet (Nguyen et al., 2020) all offer the possibility to train classifiers following this general idea with relatively modest computational expenditure. As a note, for the newest generation of very large LLMs, such a training model would not be feasible. Therefore, different training approaches would be necessary, the discussion of which is beyond the scope of this paper.

In order to perform experiments with the dataset we implemented multi-label classifiers based on BERT, RoBERTa and BERTweet. These models were fine-tuned over the course of 20 epochs, with a batch size of 16, the learning rate of 2e-5 and maximum sequence length of 200.

Regarding the metrics used to evaluate the classifiers, we

Table 6. Dataset samples

Tweet	TPCs	STPCs	HS	OFF
@USER It is Allah's anger, we need to repent and make tawba all muslims of the world.	Religion	Muslim	NO	NO
@USER she Hispanic? cuz my mom would say some sh*t like that lol	Ethnicity, Individual	Hispanic, Undefined	NO	YES
African immigrants all have the same annoying accent lol	Ethnicity, Other	African, Other_Other	YES	NO
@USER They always lie	Other	Undefined	YES	NO
@USER It is a question that all Christians must always ask themselves.	Religion	Christian	NO	NO
@USER Latinos can't have sh*t.	Ethnicity	Hispanic	NO	YES
Why the f*ck do I have an Asian teacher for a English teacher tf is this Tom foolery.	Ethnicity, Other	Asian, Teacher	YES	YES
@USER He is one of the China a*s kissers in DC.	Country, Individual	China, Undefined	YES	YES

note that while the dataset is well-balanced with regards to the hate (HS) and offensive (OFF) labels, it is unbalanced in terms of the TPCs and even less so with regards to the individual STPCs. Thus, in our experiments, in addition to the cross-entropy loss, we also reported the macro-averaged F1 and AUROC scores, which are more fair in an unbalanced dataset. These values had been independently calculated for each TPC and STPC and then averaged over all the TPCs and STPCs, respectively.

4.2. Experimental results

The experiments we conducted aimed to predict the hate and offensive (HSOFF), TPC, and STPC values using the three models we trained. We considered three separate scenarios.

In the first scenario, the input of the neural network was the text of the tweet, while the outputs were the HSOFF, TPC, and STPC values. This scenario is typically encountered when the social media network company or external observers are evaluating the messages based on only the text. The results are shown on the top row of Table 7. We can make several observations about the results. First, the quality of the classification is significantly above random, validating the fact that the THOS dataset can be used to train classifiers, both for HSOFF as well as the finer grain TPC and STPC values. In fact, we found that the TPCs classification has a higher accuracy than the HSOFF values. This is likely due to the fact that the definitions of hate and offensive are more complex and open to interpretation. Overall, the accuracy values might not be sufficient for applications such as filtering (except maybe as a system that tags possible harmful messages to be verified by a human), but they could be instrumental for the study of hate speech trends.

In the second scenario, we considered the case where the classifier receives both the text and the TPCs as input. From a practical point of view, this corresponds to instances where we have access to an oracle that can provide the TPC. In practice, such a prediction might come from the context - for instance, in social networks such as Reddit, the subreddit in which a message was posted can often determine the TPC with high accuracy. The results of this experiment are shown in the middle row of Table 7. As expected, supplying the TPC as input boosts the accuracy of both the HSOFF classification (marginally), as well as the STPC classification (significantly).

Lastly, in the third scenario, we investigated the case when the classifier receives the text and the STPCs as input (Table 7, bottom row). We observe that this scenario also marginally increases the accuracy of the HSOFF classification. As expected, in this case, the prediction of the TPC is almost perfect, as the settings of the STPC essentially predicts that the encompassing TPC will also be set.

For the second experiment, we know that for every STPC there is only one parent TPC. However, we believe that the confusion between the countries and ethnicities (e.g. European, African, Arab, Asian), is the reason behind not getting a 100% score in TPCs classification since these kinds of STPCs depend on another token in the tweet text. Similar to the first experiment in the correlation experiment, the STPCs improved the HSOFF F1_Macro score by 2 points while improving the AUROC_Macro by 1 point to be 0.82.

5. Conclusion

In this paper, we have proposed THOS, a new benchmark hate speech dataset consisting of 8.3k tweets. The dataset considers the hate and offensive speech along with the speech's target at two levels of resolution. In an experimental study, we demonstrated that the dataset can be

Table 7. Experimental results: classification of HSOFF, TPCs, STPCs based on various inputs

		HSOFF			TPCs			STPCs		
Model	Input	F1	AUROC	CE	F1	AUROC	CE	F1	AUROC	CE
BERT	txt	0.78	0.81	1.23	0.87	0.92	0.27	0.76	0.88	0.07
RoBERTa	txt	0.77	0.81	1.20	**0.89**	**0.94**	0.26	0.70	0.85	0.07
BERTweet	txt	0.78	0.82	**1.06**	0.88	0.93	**0.23**	0.64	0.82	0.07
BERT	txt+TPCs	0.78	0.81	1.23	-	-	-	**0.86**	**0.93**	**0.04**
RoBERTa	txt+TPCs	**0.79**	**0.83**	1.11	-	-	-	0.82	0.90	**0.04**
BERTweet	txt+TPCs	0.78	0.82	1.08	-	-	-	0.70	0.84	0.05
BERT	txt+STPCs	0.78	0.82	1.19	0.998	0.999	0.004	-	-	-
RoBERTa	txt+STPCs	**0.79**	0.82	1.13	0.998	0.999	0.002	-	-	-
BERTweet	txt+STPCs	0.78	0.81	1.10	0.999	0.999	0.001	-	-	-

successfully used to train LLM-based classifiers for the hate/offensive classification as well as the target classification. Furthermore, we found that knowing the target labels can improve the classification accuracy along the hate/offensive dimensions.

Acknowledgements: This work was partially supported by the National Science Foundation under awards 2114948 and 2114911.

References

Borkan, D., Dixon, L., Sorensen, J., Thain, N., and Vasserman, L. Nuanced metrics for measuring unintended bias with real data for text classification. In *Proc. of the 2019 World Wide Web Conference (WWW)*, pp. 491–500, 2019.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. Automated hate speech detection and the problem of offensive language. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 11, 2017.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

Fortuna, P., Soler, J., and Wanner, L. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proc. of the 12th Language Resources and Evaluation Conference (LREC)*, pp. 6786–6794, 2020.

Hutto, C. and Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 8, pp. 216–225, 2014.

Kim, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746–1751. Association for Computational Linguistics, October 2014.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Mathew, B., Dutt, R., Goyal, P., and Mukherjee, A. Spread of hate speech in online social media. In *Proc. of the 10th ACM Conference on Web Science (WebSci)*, pp. 173–182, 2019.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proc. of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pp. 14867–14875, 2021.

Nguyen, D. Q., Vu, T., and Nguyen, A. T. BERTweet: A pre-trained language model for English tweets. *arXiv preprint arXiv:2005.10200*, 2020.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. Multilingual and multi-aspect hate speech analysis. *arXiv preprint arXiv:1908.11049*, 2019.

Rasooli, M. S., Farra, N., Radeva, A., Yu, T., and McKeown, K. Cross-lingual sentiment transfer with limited resources. *Machine Translation*, 32:143–165, 2018.

Relia, K., Li, Z., Cook, S. H., and Chunara, R. Race, ethnicity and national origin-based discrimination in social media and hate crimes across 100 US cities. In *Proc. of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 13, pp. 417–427, 2019.

Ruder, S., Bingel, J., Augenstein, I., and Søgaard, A. Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*, 2017.

Scheitle, C. P. and Hansmann, M. Religion-related hate crimes: Data, trends, and limitations. *Journal for the Scientific Study of Religion*, 55(4):859–873, 2016.

Waseem, Z. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proc. of the First Workshop on NLP and Computational Social Science (NLP+CSS)*, pp. 138–142, 2016.

Waseem, Z. and Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proc. of the NAACL Student Research Workshop (SRW)*, pp. 88–93, 2016.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666*, 2019.