Multilingual Racial Hate Speech Detection Using Transfer Learning

Abinew Ali Ayele[1,2], **Skadi Dinter**[1], **Seid Muhie Yimam**[1], **Chris Biemann**[1]

[1]Language Technology Group, Universität Hamburg, Germany,
[2]Faculty of Computing, BiT, Bahir Dar University, Ethiopia

{abinew.ali.ayele, seid.muhie.yimam, chris.biemann}@uni-hamburg.de,
skadi.dinter@posteo.de

Abstract

The rise of social media eases the spread of hateful content, especially racist content with severe consequences. In this paper, we analyze the tweets targeting the death of George Floyd in May 2020 as the event accelerated debates on racism globally. We focus on the tweets published in French for a period of one month since the death of Floyd. Using the Yandex Toloka platform, we annotate the tweets into categories as hate, offensive or normal. Tweets that are offensive or hateful are further annotated as racial or non-racial. We build French hate speech detection models based on the multilingual BERT and CamemBERT and apply transfer learning by fine-tuning the HateXplain model. We compare different approaches to resolve annotation ties and find that the detection model based on CamemBERT yields the best results in our experiments.

1 Introduction

The rapid advancements of social media platforms like Facebook, Twitter, and YouTube during the last couple of years have enabled users to express and distribute their sentiments on events and ideas freely and conveniently. This eases the usage of hateful messages that can imply threats or harassment against minorities (Chiril et al., 2020). Since there are variations in defining hate speech globally, we took the following explanations as working definitions in this research. Therefore, hate speech is defined as a public communication consisting of messages that may express threats, harassment, intimidation, or disparagement of a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, culture or other characteristic (Nockleyby, 2000). Besides, offensive speech is also hurtful speech that is directed against another person. Compared with hate speech, offensive speech has fewer legal implications since it does not attack people based on their group identity, rather it hurts individuals based on personal characteristics and makes them offended. More specifically, racism is a type of discrimination that makes up a large portion of hate speech and is usually directed against the perceived ethnicity, appearance, religion, or culture (Rzepnikowska, 2019).

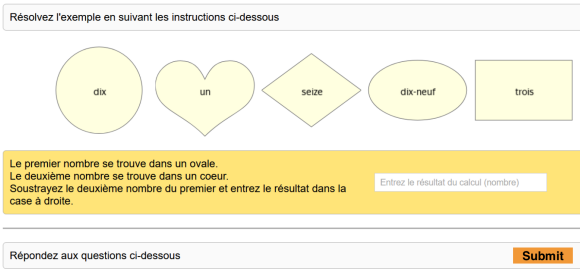

Figure 1: French language test example presented for performers

After the killing of George Floyd on May 25th, 2020, the number of racist comments on social media platforms, especially on Twitter, has increased substantially (Carvalho et al., 2022). Social media platforms use mainly content moderation systems, which are human-machine collaborative systems to detect and handle hate speech as an automatic detection system in spite of the limitations that such systems have to control the problem (Horta Ribeiro et al., 2021). These days, the task of automatic hate speech detection in general and racial hate speech, in particular, has attracted the attention of many natural language processing researchers.

To advance the development of hate speech detection algorithms in multiple languages, we extend the English hate speech detection model from HateXplain (Mathew et al., 2021) to the French language by employing our own annotated dataset. Despite there are various types of discrimination and intersections among them, we limit the scope

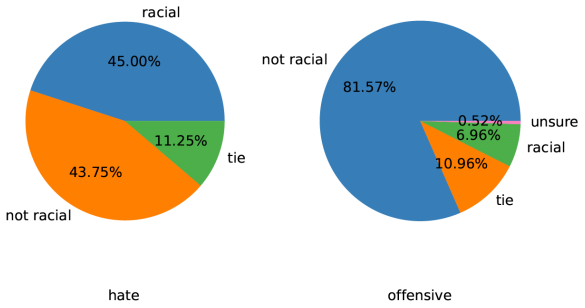

Figure 2: Class distributions of our French racial dataset

of our research to racial discrimination which is one of the most critical problems in society (Vanetik and Mimoun, 2022).

The study addresses the following research questions:

- Can BERT and HateXplain models be efficiently adapted to other languages or cultures, specifically to racial hate speech detection tasks in French?
- What are the main challenges of racial hate speech data annotation on the Toloka crowdsourcing platform?

In this paper, we employ a crowdsourcing-based racial hate speech data annotation using the Yandex Toloka platform[1]. Moreover, we fine-tuned HateXplain (Mathew et al., 2021), which is a BERT-based classification model for tweets in the French language.

The main contributions of this research include the following:

1. Collecting racial hate speech dataset in French,
2. Exploring the annotation challenges of racial hate speech annotation on the Yandex Toloka crowdsourcing platform, and
3. Adaptation of a racial hate speech detection model for the French Twitter dataset.

The remainder of the paper is organized as follows. The paper provides the related works in Section 2. While the data collection procedures and strategies are presented in Section 3, the data annotation strategies are briefly discussed in Section 4. We present our experiments including the baseline models, the results, and the error analysis in Section 5. Finally, the conclusion and future work are presented in Section 6, and the limitations of the research are indicated in Section 7.

[1] Yandex Toloka: https://toloka.yandex.com

2 Related Works

In academia, there is a strong interest in detecting hate speech and exploring the challenges facing the task. To address the issue, many researchers attempted hate speech studies by creating their own datasets and building classification models that can detect and classify hateful content from texts on social media platforms. In this regard Mozafari et al. (2020); Mathew et al. (2021); Ousidhoum et al. (2019); Davidson et al. (2017); Wang et al. (2021); Waseem and Hovy (2016); Vidgen and Derczynski (2020); Matamoros-Fernández and Farkas (2021); Vanetik and Mimoun (2022) and many other researchers investigated hate speech and developed classification models.

Most of the studies use Twitter data (Mathew et al., 2021; Vidgen and Derczynski, 2020). According to the work by Matamoros-Fernández and Farkas (2021), Twitter data is the most widely used source of data for computational social science such as hate speech and sentiment analysis tasks. Some researchers use lexical methods to retrieve social media texts based on the entries in a lexicon and build datasets for social computing (Njagi et al., 2015). The work by Davidson et al. (2017) analyzed the quality of lexical methods and proved that it is more effective to detect offensive language than hate speech. They also identified racism and homophobia more often as hate speech while sexism is more often offensive. Hate speech, racism, and racial profiling are less studied in French when compared with English (Vanetik and Mimoun, 2022). As indicated in Table 1 the study by Vanetik and Mimoun (2022) collected 2,856 French tweets and labeled them into racist and non-racist speech, and fine-tuned the BERT models for both multilingual with English dataset and monolingual models for French and English. Despite the dataset employed to build the models being a bit small in size, Vanetik and Mimoun (2022) achieved an F1-score of 67.4% for the monolingual French dataset and 64.7% for the multilingual dataset respectively as shown in Table 1. Table 1 also presented datasets and models focused on racial hate speech. The other tasks on racial hate speech presented by Waseem and Hovy (2016); Waseem (2016); Sanoussi et al. (2022) achieved F1-scores of 95.4%, 76%, and 65% class label per-

formance results respectively in different datasets.

There are fewer annotated datasets that deal with racist speech than for general hate speech, in particular for the French language (Vanetik and Mimoun, 2022). A few studies were conducted on racial hate speech in French. Chiril et al. (2020) created a French corpus of the sexist dataset by collecting tweets using keywords and becomes the first dataset to detect sexism and multi-target hate speech. Models developed for other languages such as English can not be properly adopted for racial hate speech classification in French due to contexts variations in culture and differences in linguistic features.

Mathew et al. (2021) presented a hate speech dataset annotated in three different perspectives such as:

1. the basic 3-class classification (hate, offensive or normal)
2. indicating the target community who are victims of hate/offensive speech and
3. the rationales behind the labeling decisions.

Mathew et al. (2021) adapted the CNN-GRU (Zhang et al., 2018), BiRNN (Schuster and Paliwal, 1997), BiRNN-Attention (Liu and Lane, 2016) and BERT (Devlin et al., 2019) models by modifying the original architectures.

For example, Mathew et al. (2021) fine-tuned the BERT model of Devlin et al. (2019) by adding a fully connected layer with the output corresponding to the classification tokens in the input where the token output usually holds the representation of the sentence to add attention supervision that matches the attention values corresponding to the token in the final layer.

3 Data Collection

Most of the existing hate speech datasets in French and other languages do not focus on racial hate speech. The dataset used in this research is collected from Twitter focusing on tweets that are published for one month following the death of George Floyd[2]. The death of George Floyd accelerated debates and demonstrations globally. Following the death, social media platforms such as Twitter, Facebook, and YouTube have become places for hate and offensive speeches in general and racial hate speech in particular.

[2]The New York Times: How George Floyd died, and ...: https://www.nytimes.com/article/george-floyd.html

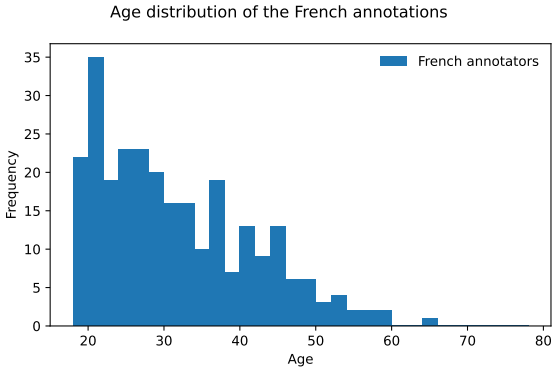

Figure 3: The age distribution of the annotators.

We employed 3,473 French hate speech lexicon entries adapted from the work of Stamou et al. (2022); Chiril et al. (2020) to filter the tweets that might contain racial hate speech content from the total 200m tweet corpus. We used the Python language detection[3] tool to filter tweets that are only written in French. We also removed truncated tweets since such tweets lack complete information and may confuse the annotators during annotation, and the model during experimentation. We removed retweets and kept only unique tweets that are not duplicated. Moreover, usernames and URLs are anonymized and replaced with <USER> and <URL> respectively. A total of 5k tweets are annotated using three independent annotators on Yandex Toloka crowdsourcing platform.

4 Annotation

Annotation by itself is a very complex task and becomes more challenging for hate speech annotations due to the lack of complete background contexts behind the texts scrapped from social media platforms (Davidson et al., 2017). We annotated 5k tweets on Toloka crowdsourcing platform and each tweet is annotated by three independent Toloka performers. We annotated 50 random tweets and evaluated the annotations by experts for the correctness of the corresponding labels. These control tweets were used to control malicious annotators engaging in the annotation task. Each task presented to performers contains 15 tweets and one of the tweets is a control question. Users are asked to classify tweets into **hate**, **offensive**, **normal** and **unsure**, and further classify hateful tweets into **racial**, **non-racial** and **unsure**. If **hate** is chosen

[3]Python Language detection library: https://pypi.org/project/langdetect/

Author	Language	Size	Labels	Best F1-Score
Vanetik and Mimoun (2022)	French	2,856	racist, not racist	67.4%
Sanoussi et al. (2022)	Chadian mixed French-Arabic	14,000	hate, insult, neutral, offensive	95.4%
Waseem and Hovy (2016)	English	16,914	racism, sexism, neither	76.0%
Waseem (2016)	English	6,909	racism, sexism, racism &sexism, neither	65.0%

Table 1: Status of racial hate speech studies (data size, labels, method, and best score and resource availability)

by an annotator, the targets **racial**, **non-racial**, and **unsure** will pop up immediately for the performer. The **unsure** label is provided to give performers the opportunity to indicate that a tweet is very hard to classify.

According to the work by Ross et al. (2017), providing the basic definitions and task descriptions of the annotation project beforehand improves the alignment of the opinions of the annotators on the class labels. We presented the annotation guideline to provide a complete description of the annotation task. Two training task pools structured in the same way as the actual task were presented to be completed by Toloka performers before joining the main annotation task. Such procedures can help Toloka performers to have sufficient knowledge and understanding of the annotation task.

One of the main challenges of crowdsourcing data annotation is the prevalence of malicious data annotators who merely participate in the annotation task to gain financial rewards (Öhman, 2020). In order to prevent potential malicious performers from engaging in the annotation task, we prepared a French language test and presented it to each performer as indicated in Figure 1. Toloka performers needed to pass the French language test in order to participate in the main French racial hate speech annotation task. We also limited the location of performers and allowed those performers who lived in France or Belgium. The performers who successfully completed the two training task pools, lived in France or Belgium, and passed the French language test were qualified and provided the privilege to access the main annotation task pools. A Fleiss kappa of 0.3 inter-annotator agreement, which indicated a fair agreement, is achieved. Each tweet was annotated by three annotators and the final gold label was aggregated from these three annotations with a majority voting scheme. As indicated in Figure 2, 45% of the tweets annotated as hate contained racial content and 11.25% had also ties. Hateful tweets had more probability to contain racial content and ties than offensive tweets. Figure 3 showed that the majority of Toloka performers who participated in the French racial hate speech annotation were young adults below 40 years. The summary of the overall annotation information is presented in Table 2. Moreover, the sample annotation task presented to Toloka performers for annotation is depicted in Figure 4, and the completed French racial Toloka project indicating the overview of the French racial hate speech annotation project is also provided in Figure 5. Each annotator earned $0.1 per task.

Fleiss Kappa score	0.3
Total number of Annotated tweets	5002
Number of annotators participated in the task	275
Mean age of annotators in years	31.11
Country distribution of annotators	265 Fr, 8 Be, 3 O
Accuracy for 50 random tweets	0.24
F1 score for 50 random tweets	0.24
Racial accuracy for 50 random tweets	0.12
Average time for 15 tweets	2 min 10 sec
Number of collected keywords	3473

Table 2: Basic annotation information (Fr= French, Be = Belgium, O = Others)

5 Experiments

5.1 Baseline Models

The BERT language model facilitates a lot of natural language processing tasks. It consists of transformer encoder layers with a self-attention mechanism (Devlin et al., 2019). The model has grown into a family of language models for a wide range of languages. The multilingual BERT and CamemBERT models are examples of such extensions. The works like HateXplain (Mathew et al., 2021),

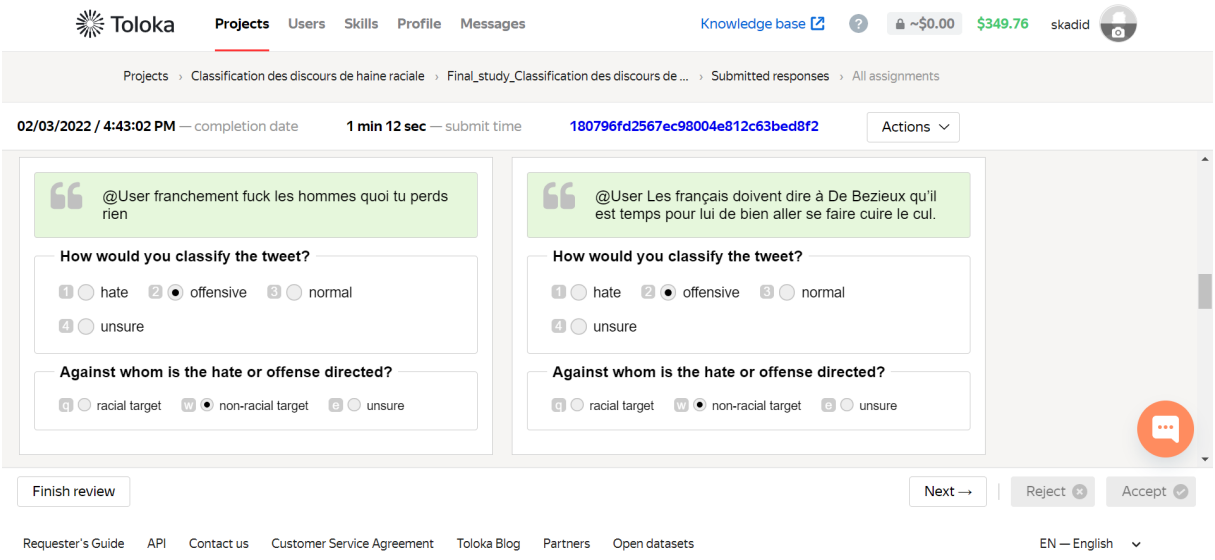

Figure 4: Example of the French annotation task.

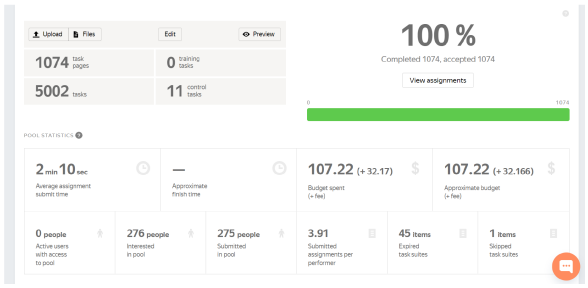

Figure 5: Completed French annotation project.

further fine-tuned the models with hate speech dataset collected from posts on Twitter[4] and Gab[5], which were filtered with keyword lists. The dataset was constructed for English and accommodated rationales to better explain the decisions of the crowd workers who annotated the posts. The HateXplain (Mathew et al., 2021) model achieved an accuracy of 70% and an F1-score of 69% on this dataset.

For this research, we employed the baseline BERT and other extended BERT models. The HateXplain dataset was used for fine-tuning the BERT models which are pre-trained for a wide range of language processing tasks. It was further preprocessed and applied for fine-tuning the multilingual BERT model. Additionally, the dataset was translated with Google Translate to French and trained on the French language model camemBERT[6]. CamemBERT is a pre-trained transformers language model developed for the French language on the original BERT (Martin et al., 2020).

We conducted different experiments by fine-tuning the HateXplain model with the multilingual BERT (ML BERT) and CamemBERT models on different datasets and class label generations. As indicated in Table 3, the first four experiments focused on the ML BERT and HateXplain model combinations (i.e., 1.0, 1.1, 1.2, and 1.3) while the next four experiments focused on the CamemBERT and HateXplain model combinations (i.e., 2.0, 2.1, 2.2, and 2.3). We analyzed the influence of different kinds of datasets and label aggregations on the performance of the models as shown in Table 3. One of them is the automatic aggregation of the three annotations for each tweet based on the Dawid-Skene aggregation method[7]. Opposed to automatic aggregation, some studies were conducted with a custom aggregation method that combines the votes in the following way: the classifications with at least two votes were considered the ground truth for each tweet. When there are three different classifications, the tweet is either removed (Experiment 1.1 and 2.1) or if there is at least one hateful label, it is considered hateful and otherwise offensive (Experiment 1.3 and 2.3) as shown in Table 3.

[4]Twitter: https://twitter.com
[5]Gab Social Network: https://gab.com
[6]CamemBERT: https://huggingface.co/camembert-base

[7]The Dawid-Skene Aggregation Model: https://toloka.ai/docs/guide/concepts/result-aggregation.html

Experiment	Pretrained Model	Label generation	Accuracy	F1-score	Ties	Training time
1.0	ML BERT	HateXplain	0.51	0.41	-	12m 47s
1.1	ML BERT+ HateXplain	self aggregated	0.84	0.77	no ties	3m6s
1.2	ML BERT+ HateXplain	Dawid Skene	0.78	0.69	automatically	4m3s
1.3	ML BERT+ HateXplain	self aggregated	0.65	0.51	if hate: hate, otherwise offensive	4m9s
2.0	camemBERT	HateXplain	0.592	0.57	-	10m45s
2.1	**HateXplain on camemBERT**	**self aggregated**	**0.888**	**0.86**	**no ties**	**3m19s**
2.2	HateXplain on camemBERT	Dawid Skene	0.806	0.75	automatically	3m54s
2.3	HateXplain on camemBERT	self aggregated	0.726	0.674	if 1 hate:hate, otherwise offensive	3m12s

Table 3: Studies for building a French hate speech detection model based on different BERT models and datasets

Experiment	Accuracy	F1	Epochs	Learn. rate
2.1 a)	0.886	0.859	3	5e-5
2.1 b)	0.899	0.882	2	5e-5
2.1 c)	0.888	0.876	1	5e-5
2.1 d)	0.882	0.869	4	5e-5
2.1 e)	0.852	0.784	3	5e-4
2.1 f)	**0.892**	**0.869**	**3**	**5e-6**
2.1 g)	**0.892**	**0.874**	**4**	**5e-6**

Table 4: Further experimental results based on Experiment 2.1 of Table 3

5.2 Results

For both of the BERT-based models, the datasets performed nearly similar results, as shown in Table 3. Hence, the model based on the Dawid Skene aggregation gained a better accuracy and F1-score than the aggregation based on the ones with a majority voting for both the multilingual BERT and camemBERT. The removal of the votes with ties has led to the best results for both base models. This implied that adding ties does not lead to better results. Experiments on the multilingual BERT such as Experiment 1.1 in Table 3 performed worse than the corresponding camemBERT (Experiment 2.1). This indicated that augmenting target datasets with translated English datasets like the HateXplain can improve the performance of the BERT modes.

The offensive tweets were predicted well but some normal tweets were also classified as offensive. There were remarkable differences between the performance of the models based on the multilingual BERT and the French camemBERT. Whilst the multilingual BERT always predicted *normal* as the class label with nearly the same score for every tweet, the camemBERT labeled the tweets appropriately. The multilingual experiments achieved a lower score than the camemBERT models. A random sample of 50 tweets that were incorrectly classified by the model was analyzed together with the reasons for the incorrect classification.

Despite all the three annotators agreed with 100% on the labels of some tweets, there were variations in the classification model where some were wrongly classified. For example, no tweet in the test set was classified as hate even though there were examples from annotators who all agreed that the corresponding tweet was hateful. This can be explained due to the class imbalance problem in the original dataset. Through further fine-tuning, the best performing model was chosen and hyperparameters like the number of epochs and the learning rate were varied as shown in Table 4. As the dataset has unbalanced classes, a stratified splitting of both the train and the test set was chosen as another experiment and showed improvements in the performance of the models.

6 Conclusion

This paper presented the collection of racial hate speech datasets from Twitter. The dataset was collected for a period of one month following the death of George Floyd in May 2020 as his murder was associated with racism. The debate regarding racism escalated during that time and racist speeches and expressions on almost all social media platforms were also aggravated. A total of 5k tweets are annotated as hate, offensive, normal, and unsure using Toloka. Furthermore, hate and offensive tweets were labeled as racial, non-racial, and unsure classes. This dataset can be used as a benchmark dataset for French racial hate speech research. The BERT model is successfully fine-tuned with the dataset together with the translated HateXplain dataset. Our experiment achieved an accuracy of 88% and an F1-score of 86% which are improving over the baseline HateXplain model.

In future work, we plan to work on further filtering the lexicon entries in order to reduce the class imbalance problem. Extending the dataset to include the racial targets and the rationales of the label decisions can also be future work. We published the resources in GitHUb[8].

7 Limitations

Due to the resources and time constraints, the annotators were not necessarily experts, which might have influenced the quality of the dataset. Since the task of racial hate speech is complex, distinguishing between hate and offensive content is even very difficult for the annotators. There are many cases where the annotators choose "unsure" as well as totally disagreed on the label's tweets during annotation. In addition to the low quality, the size of the dataset is also small and has a data imbalance problem that can be associated with the limitations of this research.

[8] https://github.com/uhh-lt/AmharicHateSpeech

References

Paula Carvalho, Bernardo Cunha, Raquel Santos, Fernando Batista, and Ricardo Ribeiro. 2022. Hate speech dynamics against African descent, Roma and LGBTQI communities in Portugal. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2362–2370, Marseille, France. European Language Resources Association.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media*, volume 11, pages 512–515, Montréal, QC, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Manoel Horta Ribeiro, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West. 2021. Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24.

Bing Liu and Ian R. Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pages 685–689, San Francisco, CA, USA. ISCA.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Ariadna Matamoros-Fernández and Johan Farkas. 2021. Racism, hate speech, and social media: A systematic review and critique. *Television & New Media*, 22(2):205–224.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on bert model. *PLOS ONE*, 15(8):1–26.

Dennis Njagi, Z. Zuping, Damien Hanyurwimfura, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10:215–230.

J Nockleyby. 2000. Hate speech in Encyclopedia of the American Constitution. *Electronic Journal of Academic and Special librarianship*.

Emily Öhman. 2020. Challenges in annotation: Annotator experiences from a crowdsourced emotion annotation task. In *Digital Humanities in the Nordic Countries Conference*, pages 293—301, Riga, Latvia.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the European refugee crisis. *arXiv preprint arXiv:1701.08118*.

Alina Rzepnikowska. 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.

Mahamat Saleh Adoum Sanoussi, Chen Xiaohua, George K Agordzo, Mahamed Lamine Guindo, Abdullah MMA Al Omari, and Boukhari Mahamat Issa. 2022. Detection of hate speech texts using machine learning algorithm. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0266–0273. IEEE.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Natalia Vanetik and Elisheva Mimoun. 2022. Detection of racist language in french tweets. *Information*, 13(7).

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *CoRR*, abs/2004.01670.

Simeng Wang, Xiabing Chen, Yong Li, Chloé Luu, Ran Yan, and Francesco Madrisotti. 2021. 'I'm more afraid of racism than of the virus!': racism awareness and resistance among Chinese migrants and their descendants in France during the Covid-19 pandemic. *European Societies*, 23(sup1):S721–S742.

Zeerak Waseem. 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*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on Twitter using a convolution-gru-based deep neural network. In *The Semantic Web: 15th International Conference, ESWC*, pages 745–760, Heraklion, Greece. Springer.