Sociocultural Considerations in Monitoring Anti-LGBTQ+ Content on Social Media

Sidney G.-J. Wong[1,2,3]

[1]University of Canterbury, New Zealand
[2]Geospatial Research Institute, New Zealand
[3]New Zealand Institute of Language, Brain and Behaviour, New Zealand
{sidney.wong}@pg.canterbury.ac.nz

Abstract

The purpose of this paper is to ascertain the influence of sociocultural factors (i.e., social, cultural, and political) in the development of hate speech detection systems. We set out to investigate the suitability of using open-source training data to monitor levels of anti-LGBTQ+ content on social media across different national-varieties of English. Our findings suggests the social and cultural alignment of open-source hate speech data sets influences the predicted outputs. Furthermore, the keyword-search approach of anti-LGBTQ+ slurs in the development of open-source training data encourages detection models to overfit on slurs; therefore, anti-LGBTQ+ content may go undetected. We recommend combining empirical outputs with qualitative insights to ensure these systems are fit for purpose.

Content Warning: This paper contains unobfuscated examples of slurs, hate speech, and offensive language with reference to homophobia and transphobia which may cause distress.

1 Introduction

The proliferation of hate speech on social media platforms continues to negatively impact LGBTQ+ communities (Stefania and Buf, 2021). As a consequence of anti-LGBTQ+ hate speech, these already minoritised and marginalised communities may experience digital exclusion and barriers to access in the form of the digital divide (Norris, 2001). There have been considerable developments within the field of Natural Language Processing (NLP) in response to this social issue (Sánchez-Sánchez et al., 2024), with most of the methodological advancements in this area being made in the last three decades (Tontodimamma et al., 2021).

While much of hate speech research has focused on documentation and detection, there has been little attention on how these approaches can be applied across different social, political, or linguistic contexts (Locatelli et al., 2023). Just as the appropriateness of swear words is highly contextually variable depending on language and culture (Jay and Janschewitz, 2008), hate speech in the form of anti-LGBTQ+ hate speech is often predicated by social, cultural, and political attitudes towards diverse gender and sexualities. With minimal literature beyond just a system development context, we set out to investigate the suitability of implementing open-source anti-LGBTQ+ hate speech system on real-world sources of social media data.

This paper makes two contributions: firstly, we show the predicted outputs from classification models can be transformed into various time series data sets to monitor the rate and volume of anti-LGBTQ+ hate speech on social media. Secondly, we argue that social, cultural, and linguistic bias introduced during the data collection phase has an impact on the suitability of these approaches.

1.1 Related Work

Hate speech detection is often treated as a text classification task, whereby existing data can be used to train machine learning models to predict the attributes of unknown data (Jahan and Oussalah, 2023). The main focus of these systems are racism, sexism and gender discrimination, and violent radicalism (Sánchez-Sánchez et al., 2024). Both the production and deployment of hate speech detection systems are methodologically similar produced under the following pipeline (Kowsari et al., 2019):

a) *Data Set Collection and Preparation*: involves collecting either real-world or synthetic instances of hate speech in a language condition (i.e., keyword search). This phase may involve or manual annotation from experts of crowd-sourced annotators.

b) *Feature Engineering*: involves manipulating and transforming instances of hate speech.

Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, pages 84–97

This may involve anonymisation or confidentialisation depending on the privacy and data use rules for each social media platform.

c) *Model Training*: involves developing a hate speech detection system with machine learning algorithms. This may involve statistical language models or incorporating transformer-based large language models.

d) *Model Evaluation*: involves producing model performance metrics to determine the statistical validity of the system. This may involve making predictions on unseen or test data.

Despite their straightforward workflow, these systems pose a number of ethical challenges and risks to the vulnerable communities (Vidgen and Derczynski, 2020). Cultural biases and harms can be introduced at each stage of the data set production process (Sap et al., 2019). Some of this can be attributed to poorly designed systems which are not fit for purpose (Vidgen and Derczynski, 2020). For example, racial bias was identified in one open-source hate speech detection system developed by Davidson et al. (2017) which resulted in samples of written African American English being misclassified as instances of hate speech and offensive language (Davidson et al., 2019).

The presence of racial bias can be attributed to the decisions made during the *Data Set Collection and Preparation* phase during system development. Davidson et al. (2017) took a keyword search approach (i.e., slurs and profanities) to identify instances of hate speech and offensive language. These samples were then used in the development of the detection system. Although slurs and profanities are good evidence of anti-social behaviour, the same words can also be re-appropriated or reclaimed by target communities (Popa-Wyatt, 2020). Classifications algorithms are unable to account for implicit world knowledge.

Similarly, simple machine learning algorithms cannot account for linguistic variation which is another form of implicit world knowledge. Of interest to our current investigation, Wong (2023a) applied the same system developed by Davidson et al. (2017) on samples of tweets/posts originating in New Zealand. The system erroneously classified tweets/posts with words such as *bugger*, *digger*, and *stagger* as instances of hate speech. An unintended consequence of these misclassified tweets/posts is that rural areas exhibited higher rates of hate speech and offensive language when compared to the national mean.

However, not all forms of biases stem from decisions made during system development. Recent innovations in transformer-based language models, such as BERT (Devlin et al., 2019), have introduced new ethical challenges as the presence of gender, race, and other forms of bias have been observed in the word embeddings of large language models (Tan and Celis, 2019). This means there is potential for bias even in the later stages of system development during the *Model Training* phase.

While we grow increasingly aware of the impacts from these limitations (Alonso Alemany et al., 2023), the number of hate speech detection data sets and systems continue to increase (Tontodimamma et al., 2021). A systematic review of hate speech literature has identified over 69 training data sets to detect hate speech on online and social media for 21 different language conditions (Jahan and Oussalah, 2023). Seemingly, the solution to addressing social, cultural, and political discrepancies within hate speech detection is to develop more systems in different languages.

There remains little interest from NLP researchers to consider the issue of hate speech detection from a social impact lens (Hovy and Spruit, 2016). The primary concerns in this research area are largely methodological. For example, improving model performance of detection systems resulting from noisy training data (Arango et al., 2022). Laaksonen et al. (2020) critiqued the *datafication* of hate speech detection which in turn has become an unnecessary distraction for NLP researchers in combating this social issue.

In fact, the appetite in applying NLP approaches for social good has decreased over time (Fortuna et al., 2021). Some researchers are beginning to question whether the efforts put towards the development and production of hate speech detection systems is the ideal solution for this social issue (Parker and Ruths, 2023). In sidelining these pressing issues in hate speech detection research, we may unintentionally perpetuate existing prejudices against marginalised and minoritised groups these systems were meant to support (Buhmann and Fieseler, 2021).

In light of these ethical and methodological challenges in hate speech detection (Das et al., 2023), we are starting to see how sociolinguistic information can be used to fine tune and improve the social and cultural performance of hate speech de-

Hostility	Direct	Indirect	Total
Abusive	20	45	65
Disrespectful	5	56	61
Fearful	5	47	52
Hateful	36	106	142
Normal	13	71	84
Offensive	65	308	373
Total	144	633	777

Table 1: The distribution of English posts/tweets and the level of hostility by directness targeting sexual orientation in Ousidhoum et al. (2019). Note that all totals are total responses.

Class	ENG	TAM	TAM-ENG
HOMO	276	723	465
TRANS	13	233	184
NONE	4,657	3,205	5,385
Total	4,946	4,161	6,034

Table 2: The class distribution of YouTube comments based on the three-class classification system (homophobic (HOMO), transphobic (TRANS), and non-anti-LGBTQ+ (NONE) content) by language condition (English (ENG), Tamil (TAM), and Tamil-English (TAM-ENG)) in Chakravarthi et al. (2021).

tection (Wong et al., 2023; Wong and Durward, 2024) using well-attested methods such as domain adaptation (Liu et al., 2019). NLP researchers may still play an invaluable role in combating online hate speech by incorporating sociocultural considerations in the development and deployment of hate speech detection systems.

2 Methodology

As discussed in Section 1.1, hate speech detection research needs to undergo a paradigmatic shift in order to truly enable positive social impact, social good, and social benefit potential. The main purpose of this paper is to ascertain the influence of sociocultural factors (i.e., social, cultural, and political) in the development of hate speech detection systems. Our research questions are as follows:

RQ1 Can we use open-source hate speech training data to monitor anti-LGBTQ+ hate speech in real world instances of social media? and;

RQ2 How do the social, cultural, and linguistic contexts of open-source training data impact on the suitability of anti-LGBTQ+ hate speech detection?

In order to address RQ1, we compare and contrast two anti-LGBTQ+ hate speech detection systems. We provide an in depth description of the data sources in Section 2.1 and our system development pipeline in Section 2.2. Once we develop the detection systems, we apply the detection systems on real-world samples of social media data to monitor anti-LGBTQ+ hate speech across different geographic dialects.

We opted for a mixed-methods approach to address this emergent area of enquiry. This is because RQ2 can only be addressed qualitatively as we consider the suitability of the detection systems and the sociocultural relevance of the predicted outputs. We will address RQ2 in the discussion (Section 4); however, we have provided relevant sociolinguistic, cultural, and political information in Section 2.3 to contextualise our discussion.

2.1 Data Sources

As part of our investigation, we use two open-source training data sets to develop our anti-LGBTQ+ hate speech detection systems in our investigation: Ousidhoum et al. (2019) (*Multilingual and Multi-Aspect Hate Speech Data Set*; MLMA) and Chakravarthi et al. (2021) (LTEDI)[1]. The MLMA and LTEDI were chosen due to the availability of data and documentation to understand the data set collection and annotation process.

The MLMA is a multilingual hate speech data set for posts/tweets from X (Twitter) for English, French, and Arabic (Ousidhoum et al., 2019). The authors took a keyword search approach by retrieving posts/tweets which matched a list of common slurs, controversial topics, and discourse patterns typically found in a hate speech. This approach proved challenging due to the high-rates of code-switching in the English and French conditions and Arabic diglossia. The posts/tweets were then posted on the crowd-sourcing platform, Mechanical Turk, for public annotation.

One of the most well-documented anti-LGBTQ+ training data sets is the English, Tamil, and English-Tamil anti-LGBTQ+ hate speech data set developed by Chakravarthi et al. (2021). The data set contains public comments to LGBTQ+ videos on YouTube. The comments were manually annotated based on

[1]We refer to it as LTEDI with reference to its central role in the various shared tasks hosted as part of the *Language Technology for Equity, Diversity, and Inclusion*

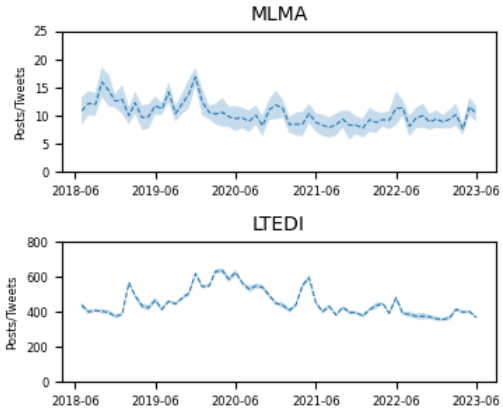

Figure 1: Model comparison of anti-LGBTQ+ hate speech on ten randomised samples of 10,000 posts/tweets per month from India between June 2018 to June 2023 including grouped mean and the upper and lower confidence intervals.

a three-class (i.e., homophobic, transphobic, and non-anti-LGBTQ+ hate speech). The training data was tested with three language models: MURIL (Khanuja et al., 2021), MBERT (Pires et al., 2019), and XLM-ROBERTA (Conneau et al., 2020).

The results show that transformer-based models, such as BERT, outperformed statistical language models with minimal fine-tuning. The best performing BERT-based system for English yielded an averaged F_1-score of 0.94 (Maimaitituoheti et al., 2022). This anti-LGBTQ+ training data set has since expanded to a suite of additional language conditions such as Spanish (García-Díaz et al., 2020), Hindi and Malayalam (Kumaresan et al., 2023), and Telugu, Kannada, Gujarati, Marathi, and Tulu (Chakravarthi et al., 2024).

We discuss the similarities and differences between the two data sets in relation sociocultural considerations regarding the data collection strategy in Section 2.1.1, the annotation strategy in Section 2.1.2, and the cultural alignment in Section 2.1.3 derived from available documentation.

2.1.1 Data Collection

The developers of the MLMA took a culturally-agnostic approach with limited information on the data collection points; however, evidence of code-switching between English with Hindi, Spanish, and French posed a challenge to annotators. The MLMA took a keyword search approach to filter X (Twitter) for instances hate speech. The keywords in relation to anti-LGBTQ+ hate in English included: *dyke*, *twat*, and *faggot*. This contrasts LTEDI which took a content search approach of users reacting to LGBTQ+ content from India.

The high-level of code-switching and script-switching between English and other Indo-Aryan and Dravidian languages provides some level of social, cultural, and linguistic information of the training data. Both training data sets are comparable in size; however, MLMA is 13.2% larger than LTEDI by number of observations. The proportion of anti-LGBTQ+ hate speech in the MLMA is 9.1% while the proportion of anti-LGBTQ+ hate speech in the LTEDI is 5.8%.

2.1.2 Annotation Process

Bender and Friedman (2018) proposed including data statement framework in the hope to mitigate different forms of social bias by dutifully documenting the NLP production process. Neither data sets provided annotator metadata (Bender and Friedman, 2018); therefore, we can only infer some of the annotator information from available documentation. Where the MLMA took a crowdsourcing approach, the LTEDI data set were annotated by members of the LGBTQ+ communities. Based on the limited details, LTEDI we know the annotators were English speakers based at the National University of Ireland Galway. Unsurprisingly, the MLMA at 0.15 is lower than LTEDI at 0.67 based on Krippendorf's alpha where 1 suggests perfect reliability while 0 suggests no reliability beyond chance.

2.1.3 Cultural Alignment

With limited documentation to the data set collection and annotation process beyond the system description papers, we tentatively determine the LTEDI is largely in alignment with anti-LGBTQ+ discourse from the South Asian cultural sphere and the MLMA as culturally-undetermined anti-LGBTQ+ rhetoric. This creates a useful contrast which not only compares the efficacy of two training data sets, but also anti-LGBTQ+ behaviour in different varieties of World Englishes which are influenced by their own unique social, cultural, and linguistic contexts (Kachru, 1982). We predict the data set collection and annotation approaches will have an impact on the outputs of the automatic detection systems.

2.2 System Development

The first phase of our investigation involves developing multiclass classification models to detect

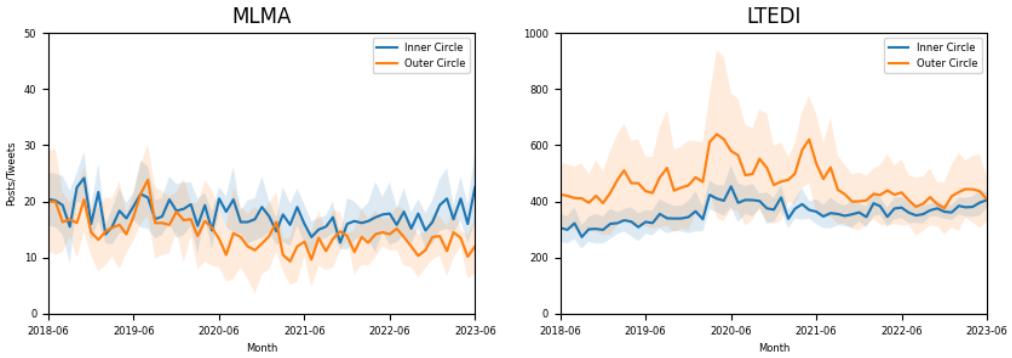

Figure 2: Comparison of anti-LGBTQ+ hate speech detected in 10,000 samples of posts/tweets from inner- and outer-circle varieties of English between June 2018 to June 2023 including grouped mean and the upper and lower confidence intervals.

	Macro		**Weighted**	
	Base	*Retrain*	*Base*	*Retrain*
LTEDI	0.78	0.81	0.95	0.96
MLMA	0.83	0.83	0.94	0.94

Table 3: Model evaluation metrics comparing the four candidate models by average macro F_1-score and average weighted F_1-score.

anti-LGBTQ+ hate speech in English. We opted for a transformer-based language modelling approach. Even though the focus of LTEDI is YouTube, we can adapt Pretrained Language Models (PLMs) to specific domains, or register of language, through pretraining with additional samples of text (Gururangan et al., 2020).

We initially trained two classification models with minimal feature engineering in order to determine the best approaches to develop our automatic detection systems. We split the training data into training, development, and test sets with a train:development:test split of 90:5:5. We used Multi-Class Classification model from the Simple Transformers[2] Python package to finetune and train the multi-class classification model. We trained each model for 8 iterations. We used AdamW as the optimiser (Loshchilov and Hutter, 2018). Our baseline PLM is XLM-ROBERTA, which is a cross-lingual transformer-based language model (Conneau et al., 2020).

2.2.1 Feature Engineering

Class imbalance had an effect on our detection system. Therefore, we collapsed the multiple classes from each training data set into a binary classification. We also removed the confidentialised user-names and URLs from Ousidhoum et al. (2019), as we could not mask these high-frequency tokens from the classification model. We used RandomOverSampler from the Imbalanced Learn[3] Python package to upsample the minority classes. We address the register discrepancy in Chakravarthi et al. (2021). We retrained XLM-ROBERTA with 120,000 samples of X (Twitter) language data from the CGLU (Dunn, 2020). The composition of the language data included 10,000 samples from each language condition.

2.2.2 Model Evaluation

We present the model evaluation metrics in Table 3. In Table 3, we compare the model evaluation results for the four candidate models (LTEDI_B, LTEDI_R, MLMA_B, and MLMA_R). The model performance improved in three of the four candidate models based on both macro average and weighted average F_1-score. Surprisingly, there were no differences between the two approaches for the MLMA models. With a focus on the anti-LGBTQ+ class, domain adaptation improved the F_1-score from 0.58 to 0.64 for the LTEDI_R model. The F_1-score for the MLMA_R remains unchanged at 0.69. Based on the model performance metrics for the four candidate models, we advanced with the LTEDI_R and MLMA_R classification models with domain adaptation and feature engineering during finetuning. We continued to apply domain adaptation in both systems despite not seeing significant improvements in the MLMA_R model to maintain consistency between the two classification models.

[2]https://simpletransformers.ai/

[3]https://imbalanced-learn.org/

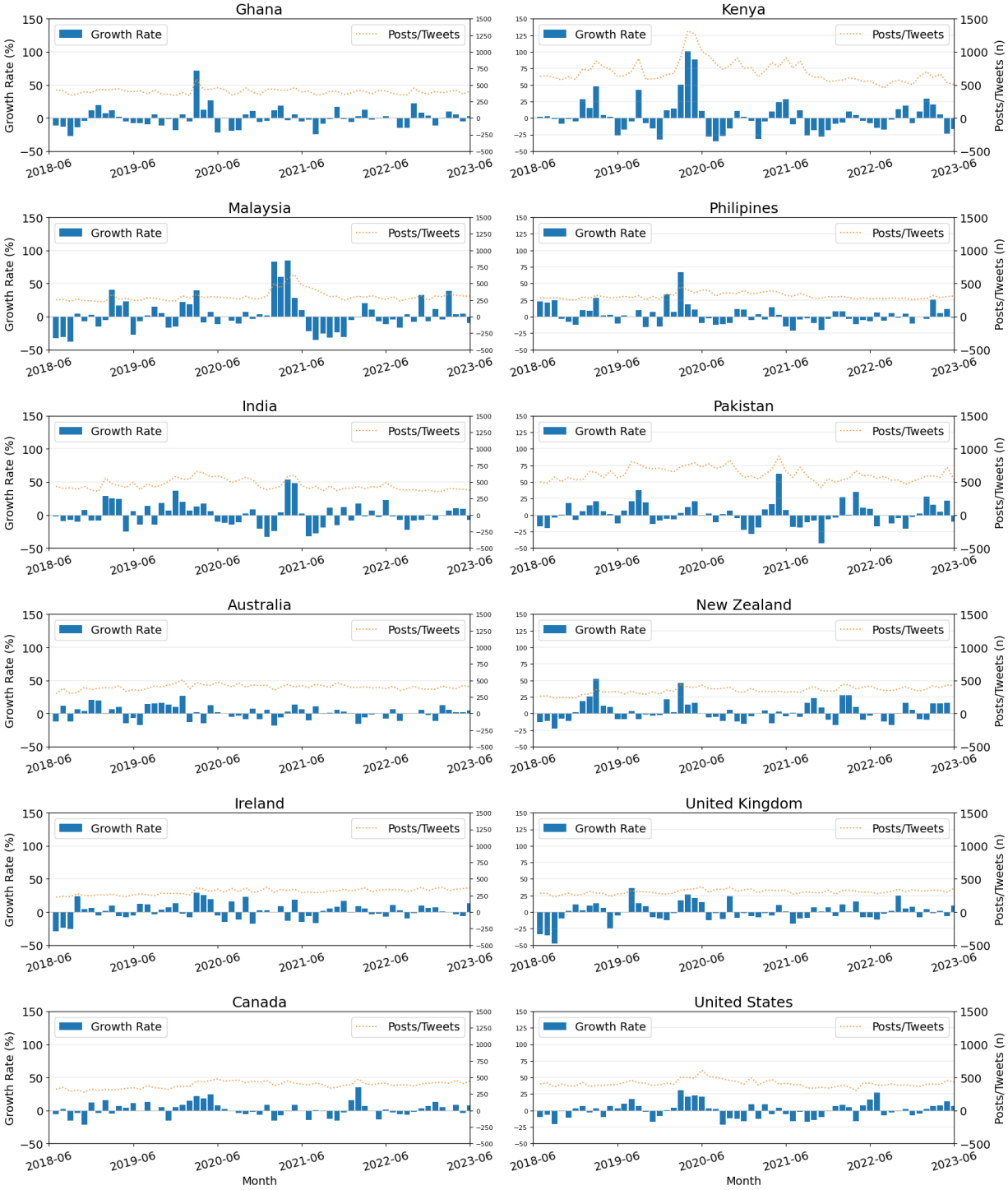

Figure 3: Quarterly growth rate of anti-LGBTQ+ hate speech detected with the LTEDI model with number of posts/tweets by country between June 2018 and June 2023.

2.3 Communities of Interest

Even though the MLMA is supposedly culturally-agnostic, we have broadly identified the cultural alignment within the LTEDI based on the data set collection and annotation process outlined in Chakravarthi et al. (2021). More specifically, high-levels of code-switching and script-switching between English, Hindi, and Tamil in the LTEDI suggests the presence of an Indian English substrate in the training data. Written English is often treated as homogeneous language; however, geographic-dialects represented by national-varieties of English maintain a constant-level of variation (Dunn and Wong, 2022).

Furthermore, the presence of Indian English on social media, or English spoken and written in India introduced as a result of British colonisation (Hickey, 2005), is uncontested (Rajee, 2024). In the three concentric circles model of World Englishes, Indian English is categorised as an outer-circle variety of English (Kachru et al., 1985). Outer-circle and inner-circle varieties of English are defined as national-varieties with British colonial ties. The distinguishing feature of outer-circle varieties is that English is not the primary language of social life and the government sector. These outer-circle varieties of English often co-exist alongside other indigenous languages.

In order to test for the influence of social, cultural, and linguistic factors, we retrieved samples of social media language from outer-circle and inner-circle varieties of English. Outer-circle varieties of English as written English originating from Ghana, India, Kenya, Malaysia, the Philippines, and Pakistan. Similarly, inner-circle varieties as written English originating from Australia, Canada, Ireland, New Zealand, the United Kingdom, and the United States. The data source of our social media language data comes from a subset of CGLU corpus which contains georeferenced posts/tweets from X (Twitter) (Dunn, 2020).

For each national-variety of English, we filtered the data for tweets in English. All posts/tweets were processed with hyperlinks, emojis, and user identifying information removed. In addition to the monthly samples for each country, we re-sampled monthly tweets from India over ten iterations to determine the impact of our sampling methodology. All posts/tweets were produced between July 2018 to June 2023. Of relevance to our analysis, the countries associated with these national-varieties all criminalised same-sex sexual activity as a legacy of the English common law legal system (with the exception of the Philippines) (Han and O'Mahoney, 2014). All but four of these countries (Kenya, Ghana, Pakistan, Malaysia) have since decriminalised same-sex sexual activity. However, LGBTQ+ rights vary significantly between countries and LGBTQ+ communities continue to face discrimination in response to increased anti-LGBTQ+ legislation in the United States disproportionately affecting transgender people (Canady, 2023).

3 Results

We dedicate the current section to describe the results of the second phase of our investigation. This phase involved applying the candidate models to automatically detect anti-LGBTQ+ hate speech on real-world instances of social media data in English. Firstly, we applied both anti-LGBTQ+ hate speech detection models on the ten randomised monthly samples of social media language data from India using the same sampling methodology for other national-varieties of English. The results are shown in Figure 1. As expected, the LTEDI_{R} model predicted higher rates of anti-LGBTQ+ hate speech; however, what was unexpected were the low number of predictions from the MLMA_{R} model. The narrow confidence intervals suggest little instability between the different samples and the predictions remained constant across samples.

After validating our sampling methodology by visually inspecting the ten randomised monthly samples from India, we applied both models on random samples of inner- and outer-circle varieties of English. We compared the results of the detection models as visualised in Figure 2. These were consistent with our initial results. The rate of anti-LGBTQ+ hate speech remained constant according to the MLMA_{R} model, while anti-LGBTQ+ hate speech has increased over time based on a visual inspection of the results. Of interest to our investigation, the MLMA_{R} model identified a higher proportion of anti-LGBTQ+ hate speech in inner-circle varieties of English. We saw an inverse relationship with the LTEDI_{R} where we see a higher proportion of anti-LGBTQ+ hate speech in outer-circle varieties of English. The wide confidence intervals of the LTEDI_{R} suggests greater between-variety instability in outer-circle varieties of English.

We calculated the quarterly growth rates for

(a) Training Data

(b) Predicted Outputs

Figure 4: MLMA Wordcloud.

(a) Training Data

(b) Predicted Outputs

Figure 5: LTEDI Wordcloud.

each variety of English for the predictions from the LTEDI$_R$. We included the total number of predicted posts/tweets in our visualisation as shown in Figure 3. The growth rates allowed us to determine the growth rate for each variety of English independently. The results suggest the growth rate of predicted anti-LGBTQ+ hate speech has remained stable over time.

4 Discussion

The results from our study raises some interesting questions on the efficacy of these systems on real-world instances of social media data. With regards to the first research question, our transformer-based multiclass classification model enabled us to detect instances of anti-LGBTQ+ hate speech from samples of georeferenced posts/tweets from X (Twitter). We were able to manipulate the predicted outputs into different forms of time series as shown in Figures 2 and 3. The level of anti-LGBTQ+ hate speech has maintained a constant rate of growth despite decreasing usership on the social media platform since the acquisition of X (Twitter) by Elon Musk in 2022. The results suggest that anti-LGBTQ+ hate speech on X (Twitter) is indeed increasing in both rate and volume over time (Hattotuwa et al., 2023).

When we compare the predicted results between the MLMA$_R$ and the LTEDI models, we can see significant differences between the two models. This is particularly obvious when we compare the predicted outputs in Figures 1 and 2, where the LTEDI$_R$ model on average predicted 50 times more instances of anti-LGBTQ+ hate speech than the MLMA$_R$. These was unexpected as the model evaluation metrics during model development suggested the MLMA$_R$ model performed marginally better than the LTEDI$_R$. Considering both the sampling methodology and the model development approaches were held constant between the models, we propose the differences we see in the predicted outputs is a result of the open-source training data.

One challenge of applying multiclass classification models on unknown data is that there is no simple method to validate the results. This is because we do not have access to labelled training, development, and test sets to evaluate the model performance. We are therefore reliant on qualitative methods to validate the performance of our detection models. Figure 4 is a visual representation of the word-token frequencies between the open-source training data (a) and the predicted anti-LGBTQ+ hate speech (b) from the samples of posts/tweets. The most prominent word-token in the training data is *faggot* followed by *dyke*. This is not unexpected as these word-tokens (including *twat*) were used to identify instances of anti-LGBTQ+ hate speech on X (Twitter). Counterintuitively, we did not see a similar distribution in the predicted outputs.

With reference to Figure 4, the word-tokens with the highest frequency in the predicted output were not *faggot* or *dyke*, but *sleep* and *gay*. When we

Variety	*dyke*	*faggot*	*twat*	*gay*
GH	8	2	6	353
IN	5	-	6	226
KE	1	4	7	295
MY	3	4	14	500
PH	8	4	8	701
PK	3	6	6	478

Table 4: Frequency of LGBTQ+ related slurs for outer circle varieties of English.

Variety	*dyke*	*faggot*	*twat*	*gay*
AU	5	12	48	635
CA	16	13	19	623
IE	15	16	62	659
NZ	6	14	53	627
UK	23	9	148	679
US	19	11	13	875

Table 5: Frequency of LGBTQ+ related slurs for inner circle varieties of English.

filtered for the keyword search terms in the samples, we found few instances across the varieties of English as shown in Tables 4 and 5. This is unexpected as the keyword search terms are highly prevalent in inner-circle varieties of spoken English (such as the United Kingdom and Ireland) (Love, 2021). This is supported by the higher word-token frequencies in inner-circle varieties of English as shown in Tables 4 and 5. We attribute the infrequent occurrence of LGBTQ+ slurs in direct response to X (Twitter) rules which discourages hateful conduct on the platform.

Our analysis of the MLMA_{R} model suggests a relationship between the training data and the resulting detection model. Incidentally, we also observe this bias towards inner-circle varieties of English in Figure 2 where the MLMA_{R} is more inclined to identify more anti-LGBTQ+ hate speech in inner-circle than outer-circle varieties of English. This leads our discussion to the second research question where we determine how the social, cultural, and linguistic context impacts the efficacy of anti-LGBTQ+ hate speech detection. Although anti-LGBTQ+ discourse is consistent across languages (Locatelli et al., 2023), slurs and swearwords are not (Jay and Janschewitz, 2008). This form of cultural bias toward inner-circle varieties of English (or oversight of outer-circle varieties) introduced during the data collection process, raises questions on the suitability of the MLMA_{R} model in monitoring anti-LGBTQ+ hate speech.

As we determined the LTEDI_{R} model to be more culturally aligned with the South Asian context, we initially predicted the LTEDI model would be more appropriate for South Asian contexts. However, the results suggest the LTEDI_{R} model as more fit for purpose in contrast to the MLMA_{R} model. Not only do we observe high-congruency between the LTEDI_{R} model output and the outer-circle varieties of English as shown in Figure 2, the word-token frequencies between the training data (a) and the predicted outputs (b) in the LTEDI appear to have a similar distribution as shown in Figure 5.

Curiously, both the training data and predicted output lack slurs. Instead, we see word-tokens associated with community (e.g., *people*) and religion (e.g., *bible*, *god*, and *Adam* possibly in reference to the Abrahamic creation myth of *Adam and Eve*). This is unsurprising as anti-LGBTQ+ legislation is often rooted in puritanical beliefs on morality (Han and O'Mahoney, 2014). With reference to Figure 3, we observed a possible link between the increased growth rate with nationwide response to the Covid-19 pandemic. Once again this raises a question on the validity of the predicted outputs and whether the posts/tweets are anti-LGBTQ+ or religious/spiritual in nature (or indeed, both).

5 Conclusion

The findings from this current paper raises a number challenges in applying hate speech detection in a real-world context. Even within national-varieties of English, we observed the impacts of social, cultural, and linguistic factors. For example, the LTEDI_{R} which was culturally aligned with Indian English was more sensitive to outer circle varieties of English, while the MLMA_{R} model was slightly more sensitive to inner circle varieties of English. We conclude that monitoring anti-LGBTQ+ hate speech with open-source training data is not problematic in itself; however, we must interpret these empirical outputs with qualitative insights to ensure these systems are fit for purpose.

Ethics Statement

The purpose of this paper is to investigate the suitability of using open-source training data to develop a multiclass classification model to monitor and forecast levels of anti-LGBTQ+ hate speech on social media across different geographic dialect

contexts in English. This study contributes to the efforts in mitigating harmful hate speech experienced by LGBTQ+ communities. In our investigation, we combine methods from NLP, sociolinguistics, and discourse analysis to evaluate the effectiveness of anti-LGBTQ+ hate speech detection.

We recognise the importance of advocate and activist-led research in particular by members of under-represented and minoritised communities (Hale, 2008). The lead author acknowledges their positionality as an active advocate and a member of the LGBTQ+ community (Wong, 2023b). The lead author is familiar with anti-LGBTQ+ discourse both in online and offline spaces and its harmful effects on members of the LGBTQ+ communities.

As discussed in Section 5, we support the critique of Parker and Ruths (2023) for NLP researchers to reflect on the efficacy and suitability of hate speech detection models. The development of hate speech data sets impose a 'diversity tax' on already marginalised LGBTQ+ communities. Originally coined by Padilla (1994), this refers to the unintentional burden placed on marginalised peoples to address inequities, exclusion, and inaccessibility particularly in a research context. NLP researchers need to work alongside key-stakeholders (e.g., affected communities, advocates, and activists) as well as social media platforms, non-profit organisations, and government entities to determine the solutions of this social issue.

The inclusion of unobfuscated examples of slurs, hate speech, and offensive language towards LGBTQ+ communities is a deliberate attempt to initiate the process of reclaiming and reappropriating some anti-LGBTQ+ slurs in NLP research. Currently, there are limited best practice guidelines on the obfuscation of profanities in NLP research (Nozza and Hovy, 2023). Worthen (2020) theorised that anti-LGBTQ+ slurs are used to stigmatise violations of social norms. Re-appropriating these stigmatising labels can enhance what were once devalued social identities (Galinsky et al., 2003). This process of 'cleaning' and 'detoxifying' slurs is also a process of resistance and to reclaim power and control (Popa-Wyatt, 2020).

We argue that within context of social media research giving unwarranted attention to slurs ignores the root of this social issue: hate speech expresses hate (Marques, 2023). Many social media platforms have already put in place procedures to censor sensitive word-tokens; however, social media users continue to adopt innovative linguistic strategies such as *voldermorting* (van der Nagel, 2018) and *Algospeak* (Steen et al., 2023) to contravene well-meaning moderation and censorship algorithms. Our results suggest hate speech training data sets do not identify the full breadth of hateful content on social media.

This paper does not include human or animal participants. Furthermore, we abide by the data sharing rules of X (Twitter) and posts/tweets with identifiable personal details will not be shared publicly. The authors have no conflicts of interests to declare.

Limitations

In this section, we address some of the known limitations of our approach in addition to limitations of the open-source training data and the social media data we have used in the current study.

Invisibility of Q+ identities This paper uses the LGBTQ+ acronym to signify diverse gender and sexualities who continue to experience forms of discrimination and stigmatisation (namely Lesbian, Gay, Bisexual, and Transgender people). While the Q+ refers to those who are not straight or not cisgender (Queer+), we acknowledge the invisibility of other minorities who are often excluded from NLP research including intersex and indigenous expressions of gender, sexualities, and sex characteristics at birth.

Sociocultural bias during data collection Despite including more training data, the MLMA identified significantly fewer instances of anti-LGBTQ+ hate speech than the LTEDI across the national-varieties of English. With reference to the word-clouds produced from the training data for MLMA and LTEDI as shown in Figures 4 and 5, there is a high likelihood the keyword search (on *dyke*, *twat*, and *faggot*) during the data collection process has caused the classification model to over-fit the training data. Similarly, the religious subtext in the LTEDI training data reinforces polarising beliefs that religion is anti-LGBTQ+. Furthermore, these detection systems do not account for semantic bleaching or the reclamation of slurs (Popa-Wyatt, 2020).

Pitfalls of large language models We acknowledge the cultural and linguistic biases introduced through the PLMs used in our transformer-based approach. However, we have mitigated some of these impacts through domain adaptation (Liu et al.,

2019). With reference to Figure 4, we have reason to believe the transformer-based detection systems erroneously classified *dylan*, *mike* and *like* with *dyke*. A breakdown of the character-trigrams (#DY, DYK, YKE, and #KE) confirms this belief.

Class imbalance and distribution We were able to improve the performance of the detection model during model development by up-sampling the minority classes. The LTEDI detected a constant proportion of anti-LGBTQ+ hate speech between 5-10% for all varieties of English which is a similar proportion of anti-LGBTQ+ hate speech in the training data (or 5.8% of the training data). This raises potential questions on the efficacy of transformer-based classification models.

Further work We welcome NLP researchers to address these limitations in their research especially on increasing the visibility of Q+ communities and the sociocultural biases shown in open-source training data sets and large language models.

Acknowledgements

The lead author wants to thank Dr. Benjamin Adams (University of Canterbury | Te Whare Wānanga o Waitaha) and Dr. Jonathan Dunn (University of Illinois Urbana-Champaign) for their feedback on the initial manuscript. The lead author wants to thank the three anonymous peer reviewers and the programme chairs for their constructive feedback. Lastly, the lead author wants to thank Fulbright New Zealand | Te Tūāpapa Mātauranga o Aotearoa me Amerika and their partnership with the Ministry of Business, Innovation, and Employment | Hīkina Whakatutuki for their support through the Fulbright New Zealand Science and Innovation Graduate Award.

References

Laura Alonso Alemany, Luciana Benotti, Hernán Maina, Lucía Gonzalez, Lautaro Martínez, Beatriz Busaniche, Alexia Halvorsen, Amanda Rojo, and Mariela Rajngewerc. 2023. Bias assessment for experts in discrimination, not in computer science. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 91–106, Dubrovnik, Croatia. Association for Computational Linguistics.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2022. Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, 105:101584.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Alexander Buhmann and Christian Fieseler. 2021. Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, 64:101475.

Valerie A. Canady. 2023. Mounting anti-LGBTQ+ bills impact mental health of youths. *Mental Health Weekly*, 33(15):1–6.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of Third Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Kayalvizhi Sampath, Durairaj Thenmozhi, Sathiyaraj Thangasamy, Rajendran Nallathambi, and John Phillip McCrae. 2021. Dataset for Identification of Homophobia and Transophobia in Multilingual YouTube Comments. *arXiv preprint*. ArXiv:2109.00227 [cs].

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward Cultural Bias Evaluation Datasets: The Case of Bengali Gender, Religious, and National Identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. 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 preprint*. ArXiv:1703.04009 [cs].

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Jonathan Dunn. 2020. Mapping languages: the Corpus of Global Language Use. *Language Resources and Evaluation*, 54(4):999–1018.

Jonathan Dunn and Sidney Wong. 2022. Stability of Syntactic Dialect Classification over Space and Time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Paula Fortuna, Laura Pérez-Mayos, Ahmed AbuRa'ed, Juan Soler-Company, and Leo Wanner. 2021. Cartography of Natural Language Processing for Social Good (NLP4SG): Searching for Definitions, Statistics and White Spots. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 19–26, Online. Association for Computational Linguistics.

Adam D Galinsky, Kurt Hugenberg, Carla Groom, and Galen V Bodenhausen. 2003. The reappropriation of stigmatizing labels: Implications for social identity. In Jeffrey Polzer, editor, *Identity Issues in Groups*, volume 5 of *Research on Managing Groups and Teams*, pages 221–256. Emerald Group Publishing Limited.

José Antonio García-Díaz, Ángela Almela, Gema Alcaraz-Mármol, and Rafael Valencia-García. 2020. UMUCorpusClassifier: Compilation and evaluation of linguistic corpus for Natural Language Processing tasks. *Procesamiento del Lenguaje Natural*, 65(0):139–142.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv preprint*. ArXiv:2004.10964 [cs].

Charles R. Hale. 2008. Engaging Contradictions: Theory, Politics, and Methods of Activist Scholarship. In *Engaging Contradictions*. University of California Press.

Enze Han and Joseph O'Mahoney. 2014. British colonialism and the criminalization of homosexuality. *Cambridge Review of International Affairs*, 27(2):268–288.

Sanjana Hattotuwa, Kate Hannah, and Kayli Taylor. 2023. Transgressive transitions: Transphobia, community building, bridging, and bonding within Aotearoa New Zealand's disinformation ecologies march-April 2023. Technical report, The Disinformation Project, New Zealand.

Raymond Hickey, editor. 2005. *Legacies of Colonial English: Studies in Transported Dialects*. Studies in English Language. Cambridge University Press, Cambridge.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Md Saroar Jahan and Mourad Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Timothy Jay and Kristin Janschewitz. 2008. The pragmatics of swearing. *Journal of Politeness Research Language Behaviour Culture*, 4(2):267–288.

Braj B. Kachru. 1982. *The Other tongue: English across cultures*. University of Illinois Press, Urbana-Champaign.

Braj B. Kachru, R. Quirk, and H. G. Widdowson. 1985. Standards, codification and sociolinguistic realism. *World Englishes. Critical Concepts in Linguistics*, pages 241–270.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint*. ArXiv:2103.10730 [cs].

Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. 2019. Text Classification Algorithms: A Survey. *Information*, 10(4):150.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Ruba Priyadharshini, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2023. Homophobia and transphobia detection for low-resourced languages in social media comments. *Natural Language Processing Journal*, 5:100041.

Salla-Maaria Laaksonen, Jesse Haapoja, Teemu Kinnunen, Matti Nelimarkka, and Reeta Pöyhtäri. 2020. The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. *Frontiers in Big Data*, 3.

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 [cs].

Davide Locatelli, Greta Damo, and Debora Nozza. 2023. A Cross-Lingual Study of Homotransphobia on Twitter. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 16–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Robbie Love. 2021. Swearing in informal spoken English: 1990s–2010s. *Text & Talk*, 41(5-6):739–762.

Abulimiti Maimaitituoheti, Yong Yang, and Xiaochao Fan. 2022. ABLIMET @LT-EDI-ACL2022: A Roberta based Approach for Homophobia/Transphobia Detection in Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 155–160, Dublin, Ireland. Association for Computational Linguistics.

Teresa Marques. 2023. The Expression of Hate in Hate Speech. *Journal of Applied Philosophy*, 40(5):769–787.

Pippa Norris. 2001. *Digital Divide: Civic Engagement, Information Poverty, and the Internet Worldwide*. Communication, Society and Politics. Cambridge University Press, Cambridge.

Debora Nozza and Dirk Hovy. 2023. The State of Profanity Obfuscation in Natural Language Processing Scientific Publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3897–3909.

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.

Amado M. Padilla. 1994. Ethnic Minority Scholars, Research, and Mentoring: Current and Future Issues. *Educational Researcher*, 23(4):24–27.

Sara Parker and Derek Ruths. 2023. Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10):e2209384120.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Mihaela Popa-Wyatt. 2020. Reclamation: Taking Back Control of Words. *Grazer Philosophische Studien*, 97(1):159–176.

Clarissa Jane Rajee. 2024. Analyzing Social Values of Indian English in YouTube Video Comments: A Citizen Sociolinguistic Perspective. *Strength for Today and Bright Hope for Tomorrow Volume 24: 3 March 2024 ISSN 1930-2940*, page 9.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Ella Steen, Kathryn Yurechko, and Daniel Klug. 2023. You Can (Not) Say What You Want: Using Algospeak to Contest and Evade Algorithmic Content Moderation on TikTok. *Social Media + Society*, 9(3).

Oana Stefania and Diana-Maria Buf. 2021. Hate Speech in Social Media and Its Effects on the LGBT Community: A Review of the Current Research. *Romanian Journal of Communication & Public Relations*, 23(1):47–55.

Ana M. Sánchez-Sánchez, David Ruiz-Muñoz, and Francisca J. Sánchez-Sánchez. 2024. Mapping Homophobia and Transphobia on Social Media. *Sexuality Research and Social Policy*, 21(1):210–226.

Yi Chern Tan and L. Elisa Celis. 2019. Assessing Social and Intersectional Biases in Contextualized Word Representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126(1):157–179.

Emily van der Nagel. 2018. 'Networks that work too well': intervening in algorithmic connections. *Media International Australia*, 168(1):81–92.

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

Sidney Wong and Matthew Durward. 2024. cantnlp@LT-EDI-2024: Automatic Detection of Anti-LGBTQ+ Hate Speech in Under-resourced Languages. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 177–183, St. Julian's, Malta. Association for Computational Linguistics.

Sidney Wong, Matthew Durward, Benjamin Adams, and Jonathan Dunn. 2023. cantnlp@LT-EDI-2023: Homophobia/Transphobia Detection in Social Media Comments using Spatio-Temporally Retrained Language Models. In *Proceedings of the Third Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 103–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sidney Gig-Jan Wong. 2023a. Monitoring Hate Speech and Offensive Language on Social Media. In *Fourth Spatial Data Science Symposium*, University of Canterbury.

Sidney Gig-Jan Wong. 2023b. *Queer Asian Identities in Contemporary Aotearoa New Zealand: One Foot Out of the Closet*. Lived Places Publishing.

Meredith Worthen. 2020. *Queers, bis, and straight lies: An intersectional examination of LGBTQ stigma*. Routledge.