Monitoring Hate Speech and Offensive Language on Social Media

Sidney Gig-Jan Wong [1,2]

[1]Geospatial Research Institute [2]University of Canterbury

Content Warning

This poster contains references to and descriptions of hate speech and offensive language.

Background

Hate speech and offensive language content on social media platforms has increased in both volume and tone since early-2023 [4]. These insights were largely based on qualitative accounts. The main contribution of this study is to offer an innovative method to monitor levels of hate speech and offensive language content on social media across Aotearoa New Zealand.

Methodology

We used georeferenced Twitter data from the Corpus of Global Language Use [3]. The georeferenced tweets originated from within a 50-kilometre radius for each of the 100 data collection points across Aotearoa as shown in Figure 1. Data collection has been on-going since June 2018.

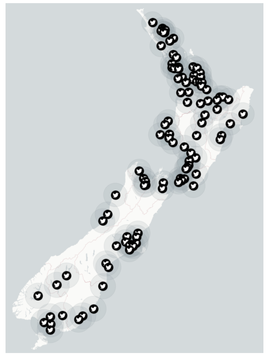

Figure 1. Map of data collection points across Aotearoa

We pretrained XLM-RoBERTa [1] with 50,000 samples of tweets from Aotearoa to fine-tune the language embeddings. We then trained a text multi-class text classification model using an open source hate speech data set [2]. Our final classification model had an average weighted F1 score of 0.90. An overview of our system is shown in Figure 2.

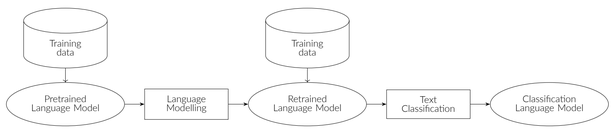

Figure 2. System Overview

We applied our text classification system on a random monthly sample of 1,000 tweets per location with replacement.

Key Findings

We took a transformer-based language approach using pretrained language models (PLMs) to develop a hate speech/offensive language classification system. We found that:

- Language models are a useful tool in monitoring social media behaviour
- Open source hate speech data sets may not be relevant within the context of Aotearoa

Therefore, we suggest further work is needed to develop language training data specific to the social, political, and linguistic context of Aotearoa.

Results

We ran the text classification model on the monthly samples of tweets for each location. The combined national totals for Aotearoa are presented in Table 1. We calculated the proportion (%) of tweets predicted as the sum of hate speech or offensive language for each location. The results show that the occurrence of hate speech and offensive language based on the classification model were increasing between the annual periods 2018-2019 to 2021-2022.

Year	Hate Speech (n)	Offensive Language (n)	Proportion (%)
2018-2019	540	29,373	4.35
2019-2020	581	33,279	4.52
2020-2021	577	34,398	4.53
2021-2022	588	30,563	4.23
2022-2023	486	27,569	4.28

Table 1. The occurrence of hate speech and offensive language for each period starting June 1 and ending May 31 the following year in Aotearoa.

Figure 3 plots the proportion of hate speech and offensive language by month. We compared the local measures for individual locations, the regional mean, and the national mean. Due to the paucity of results for some locations, we combined some of the regional council areas in Figure 3 as some locations were unable to yield monthly samples of 1,000 tweets. The West Coast/Otago/Southland area exhibited the highest proportion of hate speech and offensive language based on the classification model.

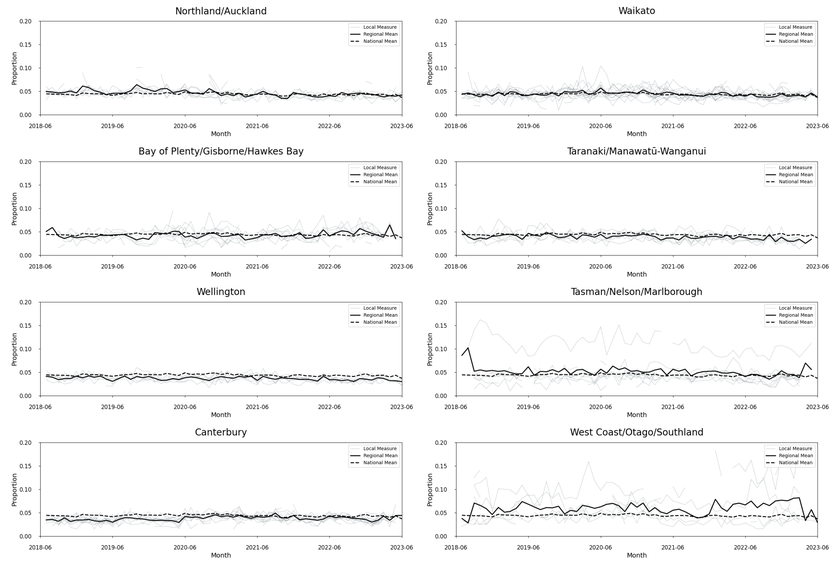

Figure 3. Proportion of hate speech and offensive tweets by broad regional areas

Discussion

The results suggest that the occurrence of hate speech and offensive language on social media has increased across Aotearoa. The results also suggest:

- Urban areas had similar rates of hate speech and offensive language when compared with the national mean
- Rural areas had higher rates of hate speech and offensive language when compared with the national mean

A closer inspection of the predicted tweets found that many of the samples would not be considered hate speech in the Aotearoa as shown in the word cloud in Figure 4 with stop words removed and slurs censored.

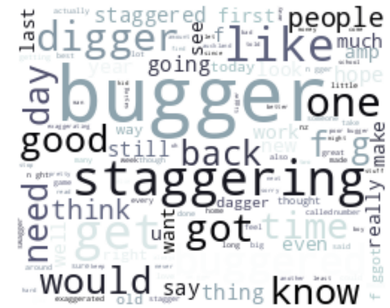

Figure 4. Word cloud of hate speech predicted by the classification model

The hate speech data set produced by used to train the classification model was not designed based on the social, political, or linguistic context of Aotearoa [2]. What might be a dialect feature words which (e.g., 'bugger') could be considered hate speech or offensive language outside the Aotearoa context. Furthermore, words which may contain structural similarities with slurs will be erroneously misclassified.

Conclusion

Despite the usefulness of language models to monitor social media behaviour such as hate speech and offensive language, we need to ensure that the language training data is relevant to a particular social, political, or linguistic context. We propose further work is needed to develop hate speech training data specific to Aotearoa.

References

[1] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale, April 2020. arXiv:1911.02116 [cs].

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

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

[4] Sanjana Hattotuwa, Kate Hannah, and Kayli Taylor. 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, April 2023.