Twitter Hate Speech Analysis Against Japanese Minorities

Ethan Nhu Tran
North Carolina State University
Raleigh, NC
entran@ncsu.edu

***Abstract*—Minorities in Japan face various forms of prejudice and discrimination. With the advances in technology and popularity of social media, the Internet provides a platform for both positive and hateful speech. This paper applies BERT, a state-of-the-art machine learning model for languages, to classify tweets about these minority populations as either toxic or nontoxic. The results of this paper show the diverse impact of social media on Japanese minority groups. All relevant code pertaining to this paper can be found at https://github.com/iskytran/fljnlp.**

I. Introduction

There is a myth of homogeneity that the Japanese possess a "highly uniform culture" widely believed both outside and inside of Japan [1]. This belief has led to the perception that Japan lacks minorities. However, this could not be further from the truth. The Ainu, Burakumin, Okinawan, and Zainichi Korean populations have historically been discriminated against, and more recently, newer immigrant populations such as the Vietnamese, Pakistanti, and others have also faced challenges lining and working in Japan. In this paper, we explore how the Ainu, Burakumin, and Zainichi Korean populations are perceived on social media by scraping Twitter data and applying a machine learning model to that data to classify the tweets.

A. Historical Context

Siddle [2] writes that the Ainu are an indigenous people that resided in what is now known today as Hokkaido. He mentions that due the abundant natural resources in Hokkaido, Japan colonized the area and the Ainu population declined. The chapter also mentions that Japanese government refused to acknowledge the Ainu as a indigenous population until 1997 when the Ainu Cultural Promotion Act was passed. However, this act has been contested as measure taken by the government to appear proactive in tackling the issue while actually hindering indigenous rights movements according to Siddle.

Neary [3] discusses how the Burakumin population are those who are believed to be descended from the outcaste communities of the Tokugawa period (1600-1876). In the chapter, it is mentioned how these outcaste groups often consisted of those who worked in the leather industry, profession beggars, or entertainers. Even after the abolition of the feudal system in Japan, those from the Burakumin population faced discrimination in housing, employment, and marriage according to Neary.

Finally, in Lie's article [4], he talks about how the Zainichi Korean population are Koreans who came to Japan and settled during the time of the Japanese Empire. The article mentions that these Koreans came to Japan in search of employment, and this was further facilitated by the fact that Koreans under the Japanese Empire were citizens of Japan. However, Lie also says that after Japan's defeat in World War II, this citizenship was rescinded, placing Koreans who chose to stay in Japan in a precarious position. Being foreigners in Japan, the Zainichi Korean population found it hard to assimilate in Japan but many of them were unable or unwilling to go back to Korea either according to the article. A Time article [5] comments that the Zainichi Koreans are one most targeted groups of abuse on social media, and Hate Speech Act passed by the Japanese government in 2016 has been ineffective due to the lack of penalties associated with it.

B. Technical Background

The family of machine learning model used in this paper is the Bidirectional Encoder Representations from Transformers (BERT) model from Google. This model is a natural language model that can be used for a wide variety of applications. Training a BERT model can be divided into two different tasks: pre-training and fine-tuning [6]. During pre-training, the model is taught using unlabeled data. Fine-tuning takes this pre-trained model and trains it further to identify labeled data [6]. This process of taking a pre-trained model and fine-tuning it is known as transfer learning.

BERT internally uses an architecture known as the Transformer architecture [6]. This architecture uses an attention mechanism to understand the relationship between all the words in a sentence irrespective of the words' position in the sentence [7]. Additionally, the architecture uses a feed forward neural network[1] to make learn and make predictions [7].

II. Approach

In order to scrape data from Twitter, the `snscrape`[2] tool was used. All search queries are limited to no further back than January 1st, 2018 to only scrape recent data and prevent excessive scraping. The filters `-filter:replies lang:ja`

[1]A neural network is a type of artificial intelligence inspired by human brains. It is made up of many interconnected nodes that process input and pass along data to each other. The feed forward component means that the neural net does not have any cycles between its units.

[2]https://github.com/JustAnotherArchivist/snscrape

were used in the search query to remove replies to tweets and restrict queries to only tweets from Japan. Additionally, all the search queries used included `OR @a1s2d3f4g5h6j7k8l` to prevent scraping tweets from users with the search term as their username. Finally, each search query consisted of both the Japanese and Latin representation of the term. So the three queries are `アイヌ OR ainu`, `部落民 OR burakumin`, and `在日 OR zainichi`.

The specific pre-trained model used for classification is the `bert-japanese-model`[3]. This model used Japanese Wikipedia as a corpus[4] for pre-training and uses the same architecture as the base BERT model of 2 layers, 768 dimensions of hidden states, and 12 attention heads. This model was then fine-tuned by training on a dataset of 1000 Japanese messages from Surge AI[5]. 500 of the entries in the dataset were labeled as nontoxic, and the other 500 were labeled as toxic. Training at this stage was done for 10 epochs[6]. The dataset was split in a stratified[7] manner into a training and testing dataset. This is done at a 4:1 ratio such that the training dataset has 800 entries and the testing dataset has 200 entries. The training dataset is fed as input to the model to learn which the testing dataset was used to check the performance of the model after each epoch of training.

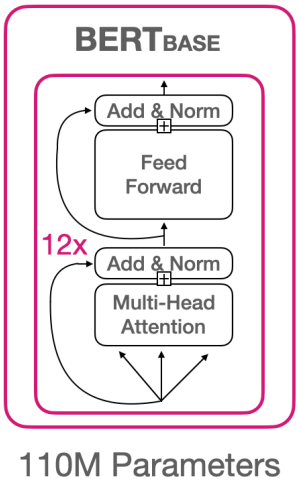

Fig. 1: BERT Architecture

Source: Adapted from [8]

III. Results

A. Model Fine-Tuning

After the end of epoch 10 of fine-tuning, the model achieved an evaluation accuracy of 98%. In addition to evaluation accuracy, there are a few other metrics that can be used to determine the performance of the model after fine-tuning. These metrics are evaluation loss, precision, recall, and F1.

Loss indicates how well the model is fitting to the data. The lower the loss value, the better this fit of the model. Accuracy tells us how well the model is correct overall, while precision tells us how well the model is performing at differentiating between categories. Recall is a metric that tells us how well a model is at identifying results within a category, and F1 is a combination of both precision and recall. Accuracy, precision, recall, and F1 can all be represented as some fraction of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The toxic entries in this paper's dataset were labeled as 1 (positive) and the nontoxic entries were labeled as 0 (negative).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

The results of the aforementioned fine-tuning metrics across the 10 epochs is summarized in Fig. 2.

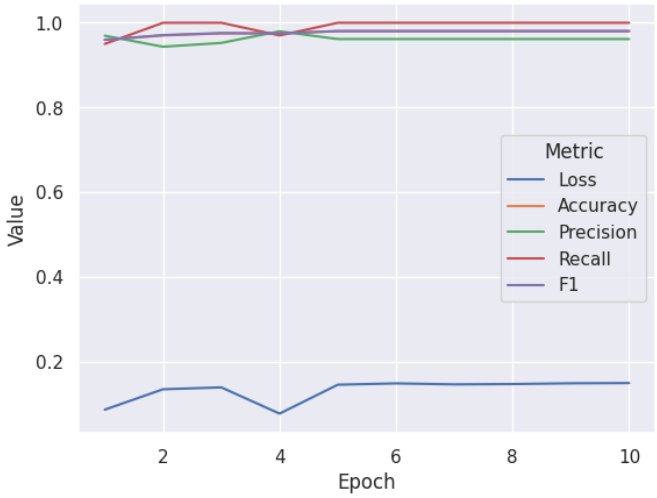

Fig. 2: Evaluation Metrics

B. Data Scraping

The number of tweets scraped after running the three different search queries can be seen in Fig. 3. We see that There are over 10 times more tweets pertaining to the Zainichi Korean minority group than the Ainu. Additionally, there are approximately 700 times more tweets pertaining to Ainu than Burakumin, and there are over 14000 times more tweets pertaining to Zainichi Koreans than Burakumin.

[3]https://github.com/cl-tohoku/bert-japanese/tree/v1.0

[4]A collection of written texts.

[5]https://app.surgehq.ai/datasets/japanese-toxicity

[6]An epoch means that the model has seen the whole dataset exactly once.

[7]Both the training and testing dataset contain an equal proportion of nontoxic and toxic entries.

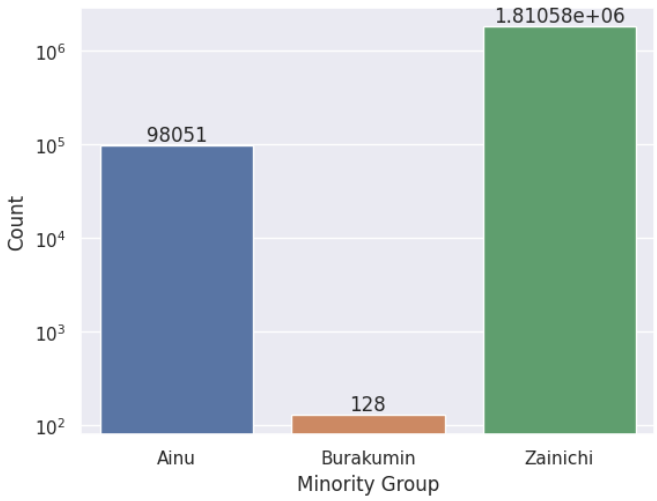

Fig. 3: Number of Tweets Scraped (Log Scale)

C. Data Classification

The results of applying the fine-tuned model to the scraped data can be seen in Fig. 4. To see the overall trends in the data, locally estimated scatterplot smoothing (LOESS) is used. This smoothing method uses local regressions to create a smooth graph from the datasets [9]. Smoothing didn't affect the Burakumin graph as there were so few data points in the first place. The graphs generally show that for Ainu, there are generally around two times as much nontoxic tweets as there are toxic tweets. The amount of tweets about Ainu also seem to be increasing in recent years, except for an abnormal drop midway into 2022. There is not much difference between the amount of nontoxic and toxic Burakumin tweets. However, there was an abnormally large spike midway in 2018. As for Zainichi tweets, there are constantly much more toxic tweets than nontoxic tweets. The graphs also indicates that time frames with decreased nontoxic tweets correlate with decreased toxic tweets and vice versa. This means that nontoxic tweets and toxic tweets decrease and increase about proportionally with the total amount tweets.

IV. Analysis

The results of this paper reinforce how Japanese homogeneity is a myth. Despite the prevailing notion that Japan is monoethnic state and the Japanese government's claims that no minorities exist in country, the existence of the Ainu, Burakumin, Zainichi Koreans, and other groups clearly show that Japan is full of diverse people. However, as the bar graph in Fig. 3 shows, the Japanese population can be reluctant to acknowledge these groups. Over a five to six year timespan, there were only 128 tweets containing the term Burakumin. Comparatively, the population of Burakumin in Japan is estimated to be around 1.2 million [10]. The disparity between the two numbers could be accounted to the ignorance and prejudice of non-Burakumin Japanese against Burakumin, as well as a mindset by the Burakumin population to distance themselves from the label to escape discrimination.

The results also show how many more tweets were classified as toxic against Zainichi Koreans than were classified as nontoxic. This demonstrates how the Korean population residing in Japan faces a great amount of prejudice and discrimination not only in the real world, but on social media as well. The data also supports the ineffectiveness of the Hate Speech Act in protecting minorities online.

Lastly, the results also show how there are more nontoxic tweets containing the search term Ainu than there are toxic tweets. This shows that social media can be used in a positive, non-hateful manner. It can be applied in ways to promote understanding and fight ignorance about these issues relating to discrimination.

V. Threats to Validity

One major threat to the validity of this paper is the fact that the model is trained with such a small dataset of only 1000 entries. Due to the small size of training data, there is a risk of misclassification by the model. This risk is partially mitigated by the usage of BERT, a state-of-the-art natural language model. The results from Fig. 2 also provides a good indication as to how well transfer learning worked on the model with nearly 1.0 accuracy, precision, recall, and F1 scores. However, there may be nuances such as sarcasm that may be difficult for the model to detect. Additionally, any biases in the training dataset would have transferred to the model, leading to possible biased classifications.

It is also hard to generalize these results to other social media platforms as all the data analyzed is pulled from Twitter. However, as Twitter is one of the most used social media platforms in Japan, it is a reasonable representation of the state of social media in the country.

The way the search query is constructed may also affect the results. With different search parameters, there may have been more or less tweets for each minority group. For example, there may be other words used to reference a specific minority group that was not included in the search query. Additionally, although the term Zainichi is commonly associated with the Japanese-residing Korean population, the term itself means "a foreign citizen staying in Japan," so using `在日 OR zainichi` may have grabbed tweets not strictly relating to Zainichi Koreans.

VI. Conclusion

In this paper, machine learning methods were used to analyze data from Twitter. The results weaken the Japanese myth of homogeneity, showing how social media has been used in a negative manner towards minority groups such as the Zainichi Koreans. It shows how in some instances, there is lack of discussion about the issues surrounding minority group as in the case of the Burakumin. However, despite these drawbacks to social media, there are instances of social media being used in a more positive manner as with the Ainu. This paper provides insight into how further research can be done on how

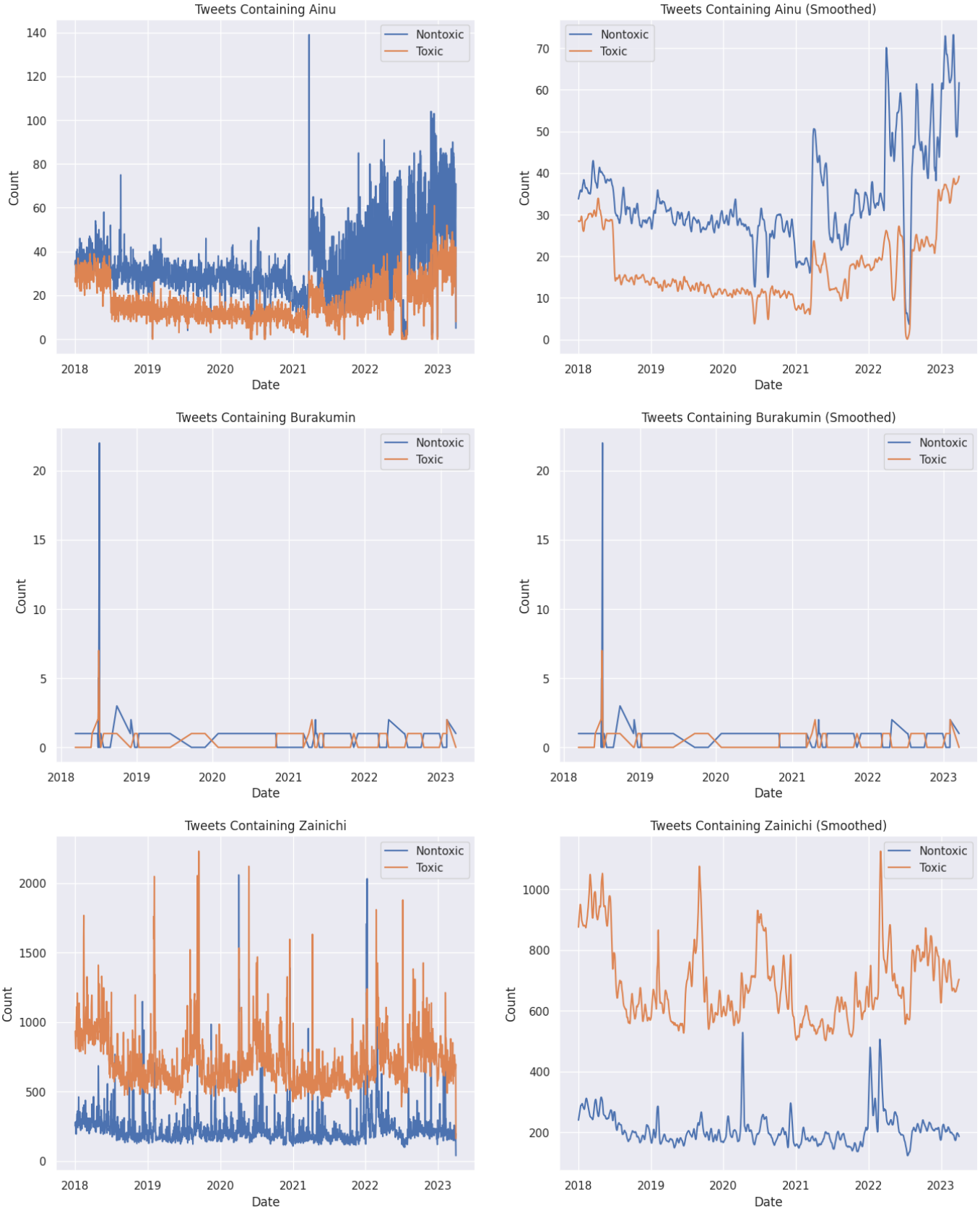

Fig. 4: Results of Classification

social media affects minority groups. Future work could be done into exploring toxicity on other social media platforms, how search queries should be refined further, improving the model with larger training data, and classifying the data into more specific categories.

REFERENCES

[1] B. J. McVeigh, *Interpreting Japan: Approaches and Applications for the Classroom*. Routledge/Taylor & Francis Group, 2014.

[2] R. M. Siddle, *The Ainu*, 2nd ed. Routledge, 2009, pp. 21–39.

[3] I. J. Neary, *Burakumin in contemporary Japan*, 2nd ed. Routledge, 2009, pp. 59–83.

[4] J. Lie, "Zainichi: The korean diaspora in japan," *Education About Asia*, vol. 14, no. 2, pp. 16–21, 2009.

[5] Y. Sato, "Online platforms like twitter are missing a brutal wave of hate speech in japan," *Time*, Sep 2022. [Online]. Available: https://time.com/6210117/hate-speech-social-media-zainichi-japan/

[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[8] B. Muller, "Bert 101 - state of the art nlp model explained," Mar 2022. [Online]. Available: https://huggingface.co/blog/bert-101

[9] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481038

[10] A. Kobayakawa, "Japan's modernization and discrimination: What are buraku and burakumin?" *Critical Sociology*, vol. 47, no. 1, pp. 111–132, 2020.