Decoding Online Hate in the United States: A BERT-CNN Analysis of 36 Million Tweets from 2020 to 2022

1st Shasank Sekhar Pandey
DACS
University of Twente
Enschede, The Netherlands
shasanksekharpandey@gmail.com

2nd Alberto Garcia-Robledo
Unidad Querétaro
Conahcyt-CentroGeo
Querétaro, México
agarcia@centrogeo.edu.mx

3rd Mahboobeh Zangiabady
DACS
University of Twente
Enschede, The Netherlands
m.zangiabady@utwente.nl

***Abstract*—Since its inception, social media has enabled people worldwide to connect with like-minded individuals and freely express their thoughts and opinions. However, its widespread nature has not only had an immeasurable impact on society but also presented significant challenges. One such challenge is online hate speech. Consequently, the identification of hate speech has recently gained considerable attention, ranging from reactive methods, such as classifying individual posts, to proactive strategies that utilize contextual information to decipher the complex lexicon of online discussions. Despite these efforts, current research lacks a comprehensive analysis of hate speech on Twitter during the crucial 2020-2022 period, marked by significant events such as the COVID-19 pandemic. In this paper, we present a BERT-based model for classifying hate speech. To this end, we collected 36 million tweets posted in the United States on Twitter during this period. We developed, trained, and tested a BERT-based Convolutional Neural Network (BERT-CNN), using it to classify the collected tweets. The classification of this dataset revealed a high incidence of targets motivated by ethnicity, with gender and nationality as other prominent categories. This work provides insightful data on the sentiments of individuals across the United States during the events of 2020-2022.**

***Index Terms*—Hate Speech Detection, Sentiment Analysis, Social Network Analysis, BERT, Convolutional Neural Networks, Twitter**

I. Introduction

With about 59% of the world using social media, for an average of 2 hr 31 mins per day [1], the impact of social media on our lives is immeasurable. However, the impact of social media is not always good.

One such negative impact is the publishing of hateful comments i.e., comments targeted at individuals or groups based on ethnicity, national background, gender identity, sexual orientation, societal class, or disability on social media platforms [2]. Moreover, hate speech has been shown to have substantial negative effects on victims' mental health, for example, in a survey focused on understanding the effects of online and offline hate speech on the LGBTQ+ community in Ukraine and Moldova, it has been shown that hate speech can cause emotional distress, depression, sleep disturbances, exhaustion, panic attacks, and feelings of social isolation [3].

These alarming trends have motivated social media platforms to deploy automated and manual detection and moderation systems, to prevent further harm. While comprehensive global statistics on online hate speech are currently lacking, it is evident that both social networking platforms and organizations dedicated to countering hate speech acknowledge the necessity of preventive measures to tackle this detrimental online phenomenon [4], [5].

This domain has also sparked a lot of attention from researchers, who have experimented with different machine learning methods such as SVMs, Random Forests, and Deep Neural Networks for identifying abusive and offensive content. Originally designed for visual pattern recognition, Convolutional Neural Networks (CNNs) now also aid in diverse fields like Natural Language Processing (NLP). Google's BERT model enables deep initial learning of text for further machine learning applications. When used together, CNNs and BERT have the potential to improve the extraction of local text information [6].

In this paper, we create and train an NLP classifier using BERT and CNNs to classify hate speech. We then apply it to a large dataset of social media posts. We collected a 36M tweets dataset from Twitter (recently re-branded as 'X'), as it is one of the few 'data-light' [7] social media sources, allowing the easy collection and storage of large datasets. As detailed throughout the paper, the classification of the collected dataset showed a high percentage of targets being motivated by ethnicity, with gender and nationality being the other dominant categories.

To the best of our knowledge, our work is the first one that studies hate speech on Twitter exploiting a large dataset, focusing on the United States and the 2020-2022 period. We believe this work offers insightful data into the sentiments held by individuals across the United States during the investigated period.

The remainder of this paper is structured as follows: In Section II, we provide background about the BERT-CNN model for hate speech detection. Section III delves into a discussion of related research. Section IV offers details of our workflow, encompassing aspects such as data collection, pre-processing,

and model training. Section V presents an analysis of the results acquired during model training and the classification of the US dataset. Lastly, Section VI furnishes conclusive remarks and outlines potential avenues for improvements.

II. Background

A. CNNs and BERT

Convolutional Neural Networks (CNNs) were first introduced as a mechanism of visual pattern recognition [8], but since have been used in various application areas, including but not limited to, Activity Recognition, Text Recognition, Face Recognition, and NLP [9]. The basic design of a CNN consists of an input layer, an output layer, and multiple hidden layers that may or may not include convolutional layers, pooling layers, fully-connected layers, and various normalization layers [9].

BERT is a linguistic model developed by a group of scientists from Google. It allows for deep preliminary learning of bidirectional text representation for subsequent use in machine learning models [10]. For our research, we aim to use this bidirectional ability of BERT to extract contextual information [11], before passing it to the CNN for classification. BERT is trained on plain text for masked word prediction and next-sentence prediction tasks [12]. Therefore, to apply the capabilities of BERT for the text classification task, it must be fine-tuned using task-specific training data [6]. Furthermore, additional task-specific layers can be applied in combination with the pre-trained BERT model to further improve its capabilities [6].

For our research, we aim to use CNN to learn features from word vectors produced using BERT and classify them. Using a CNN in combination with BERT allows for obtaining local information in text more effectively [6].

B. Hate speech detection using BERT

BERT provides a transfer-learning approach to hate speech detection, as it can be fine-tuned and applied in combination with other deep learning models for hate speech detection [13]. This transfer-learning approach has been utilized by various researchers. The authors of [14], have fine-tuned BERT with Masked Rationale Prediction (MRP) to increase the model's explainability and have obtained a macro F1 score of 0.699. In [2], the authors present dictNN, where they combined BERT with a 3-layer CNN along with a dictionary approach in the preprocessing stage, to obtain a macro F1 score of 0.61. Lastly, [13] explores different combinations of BERT and DLMS presenting models such as BERT + Non-linear layers with an F1-score of 0.92, BERT+LSTM with an F1 score of 0.88 and BERT+CNN with an F1-score of 0.92.

III. Related work

Authors in [4] present the outcomes of the "Italian Hate Map" project, which utilized a lexicon-based approach involving semantic content analysis. The project extracted a total of 2,659,879 tweets from 879,428 Twitter profiles over 7 months. Among these, 412,716 tweets contained negative language targeting six distinct groups. Among geolocated tweets, women were the most targeted group, subjected to 71,006 hateful tweets (accounting for 60.4% of the negative geolocated tweets). They were followed by immigrants (12,281 tweets, 10.4%), gay and lesbian individuals (12,140 tweets, 10.3%), Muslims (7,465 tweets, 6.4%), Jews (7,465 tweets, 6.4%), and disabled individuals (7,230 tweets, 6.1%).

Authors in [15] employ a descriptive qualitative approach, gathering data from Social Media Analysts through a query targeting hate speech terminology linked to the interplay between religion and the State. The primary emphasis lies on the Twitter platform due to its capacity for conveying intricate messages. This social networking medium is frequently utilized across diverse societal strata to articulate viewpoints, reactions, and responses, particularly concerning topics entwined with religion, politics, and governance.

Asian Americans have been subjected to both verbal and physical violence driven by individual-level racism and xenophobia Since their arrival in America in the late 1700s. This unfortunate trend has persisted through the present day. Furthermore, at an institutional level, the state has often indirectly supported and perpetuated such violence by endorsing prejudiced rhetoric and exclusionary policies [16].

The emergence of COVID-19 has exacerbated this issue, fostering the proliferation of racism and generating a sense of national insecurity. This has led to increased apprehension towards foreigners and generalized fear, ultimately fueling xenophobia. Notably, this surge in xenophobic attitudes appears to be connected to the rise in anti-Asian hate crimes during the pandemic.

In this context, the authors in [16] delve into the dynamics of these hate crimes. They explore how these acts, rooted in deep-seated historical and interconnected forms of individual-level and institutional-level racism and xenophobia, have contributed to the marginalization of Asian Americans, perpetuating inequality.

The authors in [17] provide a comprehensive context for understanding the risk posed by transphobic hate crimes. They achieved this by investigating the lived experiences, various forms, and factors associated with transphobic hate crimes encountered by transgender women in the San Francisco Bay Area, a location designated as the SFBA site within the longitudinal cohort study known as Trans*National. Additionally, the authors explored whether these instances of hate crimes were officially reported to law enforcement and delved into the factors influencing the decision to report. In their pursuit of more nuanced analysis, the authors employed a stratified approach based on race and ethnicity. This methodology allowed them to discern potential disparities tied to racial and ethnic factors in the ways trans women experienced hate crimes.

In [18] the author's primary objective is to explore the phenomenon of online harassment within the context of press censorship in today's digital society. The central argument put forth is that the act of online harassment can be understood as a form of mob censorship. This concept is defined as a

grassroots, citizen-driven form of vigilantism to discipline and effectively silence journalists.

The implications of mob censorship are portrayed as multifaceted, posing significant threats to journalists' safety and freedom of expression rights. The overarching goal of the article is to propose a conceptual framework that acknowledges the global impact of online harassment on journalists while simultaneously recognizing the specific nuances at local and national levels.

For analysis, the article focuses primarily on the United States. The author draws upon a range of sources, including academic studies, surveys conducted with journalists, news stories, and background interviews. Notably, the author conducted interviews with thirty reporters and editors employed by mainstream news organizations in 2019 to gather firsthand insights and perspectives on the issue. In essence, the article delves into the intricate connection between online harassment, the implications it has for press censorship, and the distinct characteristics of this phenomenon as observed in the context of the United States.

Other works related to hate speech analysis for tweets posted by people in the US are the following. Authors in [19] study the relationship between temperature and hate speech by analyzing 4 billion tweets from 773 cities across the United States between 2014 and 2020. They found a quasi-quadratic relationship between temperature and hate speech, and found an increase in the number of hate tweets at hotter and colder temperatures. Authors in [20] study, hate speech crimes committed by Americans against Asians on Twitter. They analyze tweets related to both "COVID-19" and "Asianhatecrimes". From a 10M dataset, 3 thousand tweets are annotated by four Asian and Asian-American annotators. Authors in [21], identify topics related to race, ethnicity, and racism, on a dataset of tweets posted after the Atlanta spa shootings of March 16, 2021. They examine patterns in expressions of hate speech and solidarity before and after the incident. The dataset included 708 thousand tweets using race-related keywords.

To the best of our knowledge, our work is the first to investigate hate speech as expressed on the Twitter platform in the United States while specifically covering the 2020-2022 time frame, involving the classification of a large and all-encompassing dataset (i.e. not focusing on specific terms, keywords or hashtags), consisting of tens of millions of tweets.

IV. Methodology

A. Data collection

For training our models we used a hate speech and offensive language Twitter dataset from [22], containing 25,000 tweets that were manually labeled using the crowd-sourcing platform CrowdFlower (labels: hate speech, offensive language, neither). The distributions of the labels are shown in Table I.

For applying the model, we collected 36,790,672 tweets made in the USA between 01-01-2020 to 31-12-2022 using the public advanced search API of Twitter, with 1400 tweets being collected for every hour of every day.

TABLE I
Distribution of classes in the dataset

label	class	No. of Instances
Hate Speech	0	1430
Offensive	1	19190
Neither	2	4163

B. Pre-processing

For normalization of tweets, we used Ekphrasis library[1] as it allows for pre-processing text from social networks and performs functions such as tokenization, word normalization, word segmentation, and spell correction.

Specifically, we normalized each tweet to remove usernames, and URLs and correct spelling errors using the ekphrasis library.

C. BERT-CNN model architecture and implementation

The architecture of the implemented BERT-CNN model is shown in Fig. 1. The model is structured in two parts:

1) *Pre-trained BERT base model.* It is used for the conversion of words in the tweet into contextualized vector representations [23]. In other words, the BERT model is used to convert input text i.e., a tweet into word vectors and to create a primary input matrix.
2) *CNN classifier.*. The input matrix is fed to two convolutional layers, which create feature maps. The feature maps are then converted to max-value feature vectors using the Global Max Pooling layer to downsample the inputs. The results from the pooling layer are passed onto the fully connected dense layer for dimensionality reduction. Next, a dropout layer is used to reduce overfitting by dropping the forward and backward connections of certain neurons, thus preventing co-adaptation. Lastly, a final fully connected output layer is used for classification. This architecture is similar to the CNN architecture used in [15].

There have been different methods of designing the CNN classifier, from using 3 convolutional layers with increasing output channels [2], to using 4 parallel convolutional filters of different sizes [23]. However, due to the increased complexity of using parallel convolutional filters, we chose to use the structure described in Kupi et al. [2], but with only two convolutional layers.

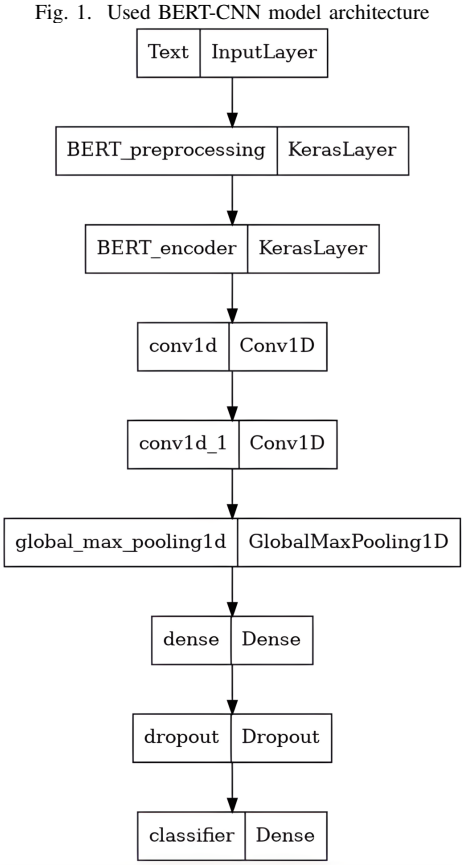
We used TensorFlow[2] to create the BERT-CNN model. We used the pre-trained Small-BERT model, consisting of 4 Hidden layers with hidden layer size = 512 and 8 Attention Heads. We also utilized the available pre-processing model for Small-BERT, to provide it with the desired inputs.

All layers, except for the last Dense layer (classifier), were implemented with the Rectified Linear Unit (ReLU) activation

[1]https://pypi.org/project/ekphrasis/. ekphrasis PyPI. Visited on April 24, 2023.

[2]https://www.tensorflow.org/. TensorFlow. Visited on May 19, 2023.

Fig. 1. Used BERT-CNN model architecture



function. The final Dense layer implemented a sigmoid activation function. We also experimented with a SoftMax activation function for the last layer, but it led to overfitting.

D. Model training

To train the model, we divided the dataset by [22] into 90% Train and 10% Test splits. We use a larger train dataset to counter the unbalanced nature of our dataset, allowing for more samples of each class to be available during training. Moreover, we stratified the splits based on classes to ensure the availability of all classes in both datasets. We batched the dataset into batches of size 32, to allow for GPU utilization for training.

The final model was built using Adam Optimization and trained with the following parameters: epochs = 120 and learning rate = 3 x 10^{-7}. Learning rates of 3 x 10^{-6}, 3 x 10^{-5}, 0.1, 0.03 were also experimented with, along with various epoch lengths such as 10, 20, 25, 40 and 80. However, the model overfitted the training dataset with higher learning rates and shorter epochs. Lastly, the final model used unprocessed tweets and only relied on BERT pre-processing, as that one provided the best accuracy and F1-score.

V. Results

A. Model training performance

To determine the model's classification abilities, we used the values of precision, recall, and F1-score per class. Precision allows for visualizing the reliability of the model and is calculated by the following formula:

$$Precision = \frac{TP}{TP + FP}. \tag{1}$$

Recall allows for measuring the ability of the model to detect positive samples and is calculated by the following formula:

$$Recall = \frac{TP}{TP + FN}. \tag{2}$$

Lastly, we also calculated the F1-score because it provided us with the accuracy of the model by combining precision and recall of the model in the following formula:

$$F1\ score = \frac{2 * Precision * Recall}{Precision + Recall}, \tag{3}$$

where TP denotes True Positive, FP denotes False Positive, TN denotes True Negative and FN denotes False Negative. The resulting values of Precision, Recall, and F1-score of the two models can be seen in Table II.

We also calculated the accuracy and macro F1-score of the model as shown in Table II. Accuracy allows us to get a general understanding of how many labels are correctly classified by the model. Accuracy is calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4}$$

Macro F1-score is an arithmetic mean of the F1 scores of all the labels and allows us to measure the model's performance, specifically when trained on imbalanced datasets like ours (see Table III).

The results show that the model has a high accuracy of 0.9 as seen in Table III and it performs the best when classifying "Offensive" tweets as can be seen by the F1-score of 0.95 in Table II. The model has a lower recall for "Hate speech" tweets, resulting in a lower F1-score of 0.65. As a reference, the work in [2] reports F1-scores between 0.54 and 0.74 when classifying hateful, abusive, and normal tweets. Thus, the score shown by our model is not surprising. The lower F1-score

TABLE II
Per class Precision, Recall, and F1 score

label	Precision	Recall	F1-score
Hate-Speech	0.76	0.57	0.65
Offensive	0.94	0.96	0.95
Neither	0.78	0.86	0.82

TABLE III
Accuracy, Macro F1, and Testing Loss

Model	Accuracy	Macro F1-score	Test Loss
BERT-CNN	0.90	0.81	0.291

Fig. 2. Hate speech by target category (%) over the 2020-2022 period.

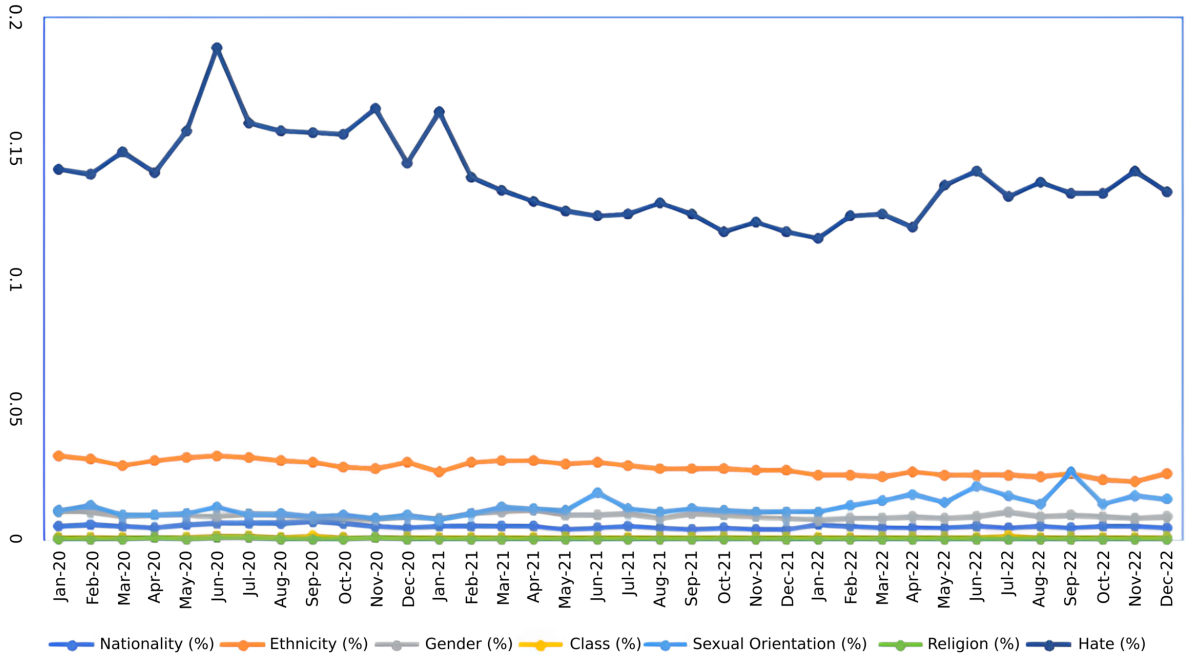


TABLE IV
CLASSIFICATION RESULTS WITH BERT-CNN

class	Total samples	Percentage (%)
Hate Speech (0)	5,090,064	13.84
Offensive (1)	4,231,252	11.5
Neither (2)	27,469,356	74.66

for the hate speech can be attributed to the skewness in the training dataset, where around 75% of the data was labeled as Offensive, with less than 6% of data labeled as "Hate speech".

B. US dataset classification

We use the developed BERT-CNN model to classify the dataset collected from Twitter between 2020 and 2023. The process of classification required 62.04 hours with the classification results shown in Table IV.

We used 1,550 terms filtered for the English language on HateBase. We searched for each of these terms in the tweets classified as hate speech, using the regular expression 5 given below:

$$RegexPattern = \backslash b\{word\}\backslash b \tag{5}$$

The above regex allows one to search for each term as a word and forego false hits with a simple substring search. For example, a substring search would result in True when searching for "and" in "Anderson has 2 kids", while the above regex would result in False, which is the correct result.

We utilized the above regex pattern, replacing the word variable with the offensive term, and searching through all hateful tweets for each offensive term resulted in 75 offensive terms with a percentage of occurrence greater than 0.01%. The top 10 terms are shown in Table V.

TABLE V
TOP 10 HATE TARGETS

Term	Occurrence	Percentage (%)
niggas	42,933	0.84
nigga	33,090	0.65
Girl	20,094	0.39
gay	17,898	0.35
af	9,756	0.19
queen	6,982	0.14
property	6,719	0.13
chief	6,187	0.12
trash	5,670	0.11
queer	3,691	0.07

TABLE VI
TARGETS GROUPED BY CATEGORY.

Category	Percentage (%)
Nationality	0.54
Ethnicity	2.80
Gender	0.94
Class	0.12
Sexual Orientation	0.12
Religion	0.06

We also categorized all identified terms into groups/categories based on the labels provided by HateBase with the results shown in Table VI. The results show a high percentage of targets being motivated by ethnicity, with

gender and nationality being the other dominant categories.

We charted the hate speech (%) along with the hate speech by category (%) in Fig. 2. A decrease in hate speech can be seen from January 2020 to December 2022, with the highest peak in May 2020 and the lowest peak in January 2022, as observed by the dark blue line. However, the percentage of hate tweets targeted based on ethnicity has remained almost constant in the period as can be seen by the orange line. The other target categories have a way lower percentage than ethnicity, but there has been a gradual increase in hate speech targeted based on sexual orientation as seen by the light blue line.

VI. Conclusions

In this paper, we developed and trained a BERT-CNN model for hate speech detection on a large all-encompassing dataset of tweets issued in the US during the 2020-2022 period. The results revealed a high percentage of targets being motivated by ethnicity, gender, and nationality. On the other side, we observed a decrease in hate speech from January 2020 to December 2022. However, we also observed a steady trend in hate speech related to ethnicity and an increase in hate speech related to sexual orientation.

The model performed well during training for the offensive label, with an F1-score of 0.95, confirming the abilities of Machine Learning in the domain of hate speech analysis. Although the model showed a score of 0.65 for the case of the hate speech label, it falls not far from our reference model in the literature. A more balanced dataset should enable better classification performance of the model. Another improvement to the model would be the addition of more convolutional layers in the CNN architecture, as it would allow for learning of more high-level textual features, and better classification.

Given the size of the classified dataset and the analyzed period, we believe that this work can be potentially useful for better understanding the sentiments held by individuals across the United States during the investigated period, which coincides with the span of the COVID-19 pandemic if considered together with works from other disciplines such as journalism and social sciences.

References

[1] S. Kemp, "Digital 2023: Global overview report — datareportal – global digital insights."

[2] M. Kupi, M. Bodnar, N. Schmidt, and C. E. Posada, "dictnn: A dictionary-enhanced cnn approach for classifying hate speech on twitter," 2021.

[3] O. Ștefăniță and D.-M. Buf, "Hate speech in social media and its effects on the lgbt community: A review of the current research," *Romanian Journal of Communication and Public Relations*, vol. 23, no. 1, pp. 47–55, 2021.

[4] V. Lingiardi, N. Carone, G. Semeraro, C. Musto, M. D'Amico, and S. Brena, "Mapping twitter hate speech towards social and sexual minorities: A lexicon-based approach to semantic content analysis," *Behaviour & Information Technology*, vol. 39, no. 7, pp. 711–721, 2020.

[5] O. Ștefăniță and D.-M. Buf, "Hate speech in social media and its effects on the lgbt community: A review of the current research," *Romanian Journal of Communication and Public Relations*, vol. 23, no. 1, pp. 47–55, 2021.

[6] S. Zheng and M. Yang, "A new method of improving bert for text classification," in *Intelligence Science and Big Data Engineering. Big Data and Machine Learning: 9th International Conference, IScIDE 2019, Nanjing, China, October 17–20, 2019, Proceedings, Part II 9*, pp. 442–452, Springer, 2019.

[7] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," *PloS one*, vol. 10, no. 3, p. e0115545, 2015.

[8] K. Fukushima, "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," *Biological cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[9] A. Dhillon and G. K. Verma, "Convolutional neural network: a review of models, methodologies and applications to object detection," *Progress in Artificial Intelligence*, vol. 9, no. 2, pp. 85–112, 2020.

[10] M. Koroteev, "Bert: a review of applications in natural language processing and understanding," *arXiv preprint arXiv:2103.11943*, 2021.

[11] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, pp. 1–41, 2021.

[12] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, vol. 1, p. 2, 2019.

[13] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pp. 928–940, Springer, 2020.

[14] J. Kim, B. Lee, and K.-A. Sohn, "Why is it hate speech? masked rationale prediction for explainable hate speech detection," *arXiv preprint arXiv:2211.00243*, 2022.

[15] H. Sazali, U. A. R. SM, R. F. Marta, *et al.*, "Mapping hate speech relationships indonesia's religion and state in social media," *Communicatus: Jurnal Ilmu komunikasi*, vol. 6, no. 2, pp. 189–208, 2022.

[16] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 1668–1678, 2019.

[17] A. O. Gyamerah, G. Baguso, E. Santiago-Rodriguez, A. Sa'id, S. Arayasirikul, J. Lin, C. M. Turner, K. D. Taylor, W. McFarland, E. C. Wilson, *et al.*, "Experiences and factors associated with transphobic hate crimes among transgender women in the san francisco bay area: comparisons across race," *BMC public health*, vol. 21, no. 1, pp. 1–15, 2021.

[18] S. Waisbord, "Mob censorship: Online harassment of us journalists in times of digital hate and populism," *Digital Journalism*, vol. 8, no. 8, pp. 1030–1046, 2020.

[19] A. Stechemesser, A. Levermann, and L. Wenz, "Temperature impacts on hate speech online: evidence from 4 billion geolocated tweets from the usa," *The Lancet Planetary Health*, vol. 6, no. 9, pp. e714–e725, 2022.

[20] A. Toliyat, S. I. Levitan, Z. Peng, and R. Etemadpour, "Asian hate speech detection on twitter during covid-19," *Frontiers in Artificial Intelligence*, vol. 5, p. 932381, 2022.

[21] S. Criss, T. T. Nguyen, E. K. Michaels, G. C. Gee, M. V. Kiang, Q. C. Nguyen, S. Norton, E. Titherington, L. Nguyen, I. Yardi, *et al.*, "Solidarity and strife after the atlanta spa shootings: A mixed methods study characterizing twitter discussions by qualitative analysis and machine learning," *Frontiers in Public Health*, vol. 11, p. 952069, 2023.

[22] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, pp. 512–515, 2017.

[23] A. Safaya, M. Abdullatif, and D. Yuret, "Kuisail at semeval-2020 task 12: Bert-cnn for offensive speech identification in social media," *arXiv preprint arXiv:2007.13184*, 2020.