Journal of Information & Optimization Sciences
ISSN 0252-2667 (Print), ISSN 2169-0103 (Online)
Vol. 45 (2024), No. 4, pp. 885–896
DOI : 10.47974/JIOS-1613

Detection of twitter hate tweets leading to crime using multiseries BERT model

Tenzin Choesang *
Boppuru Rudra Prathap †
Department of Computer Science and Engineering
Christ (Deemed to be University)
Bangalore 560060
Karnataka
India

Abstract

Addressing hate speech on social media platforms is a pressing concern, particularly on platforms like Twitter and Facebook. To date, extensive research and numerous projects on the subject of hate speech detection and prevention have been conducted, as well as for crime detection and its prevention. However, there is a major void in the current research terrain whereby there exists no specific effort to study how hate speech and crime are related globally. This research work proposes a multi-series methodology to investigate the complex linkages between online hate speech and real world acts of violence. Multi series BERT model from transformer was used for detecting the hate speech in the tweets of popular politicians, celebrities, peoples tweets on popular cases and showing the linkage of these to the real world crime happenings. When compared with the other available popular models like CNN. KNN, SVM, RF, LR, NB, DT which are generally used for detecting hate speech and crime detection, the proposed BERT model gave great results with 93% in accuracy as well as 92% in F1 Score outperforming all the other models. Furthermore the results showed that states like Delhi, Maharashtra and Tamil Nadu have the higher chances of crime happening due to hate speech based on the prediction done. However the data may vary for different social media platforms as the current research work is only based on twitter.

Subject Classification: *68T45, 68T50, 68T20.*

Keywords: *Hate speech, Deep learning, Artificial neural network, Transformers.*

* *E-mail:* **tenzin.choesang@mtech.christuniversity.in (Corresponding Author)**
† *E-mail:* **boppuru.prathap@christuniversity.in**

1. Introduction

Hate speech is a prime concern in various popular online social networks. There is a thin line that separates it from aggressive or abusive text; however, all these have one thing in common, which is offensiveness. The impacts of such forms of expression are not uniform and can result in different forms of crimes like the blockade on the road with mass protests stopping all vehicles as well as religious clashes leading to deaths, cyber bullying among others. Hence, shedding light on the correlation between hate speech and tangible offenses can enhance public consciousness about the gravity of hate speech. It can prompt individuals to be more accountable in their online interactions and report hate speech when they come across it. It can aid in identifying patterns and early indicators, enabling proactive intervention to avert potential acts of violence or bias. Acknowledging the consequences of hate comments and speech on tangible victims of hate crimes can lead to amplified support and resources for those affected. It can also nurture a sense of unity and communal response to hate crimes. Demonstrating the association between hate speech tweets and tangible offenses is crucial for comprehending the severity of hate speech, taking legal measures when necessary, heightening public consciousness, shaping policies, averting future incidents, and supporting victims.

This research proposes a multi-series methodology to investigate the complex linkages between online hate speech and real world acts of violence. The approach consists of various key elements, each tailored to fulfill specific research objectives. It encompasses data collection, preprocessing, model development, and analysis, culminating in the exploration of potential links between hate speech and criminal incidents. The project utilizes a powerful transformer model, particularly BERT models, to ensure the accuracy and context-awareness necessary for a comprehensive investigation. Through meticulous data gathering, annotation, and training, it aims to detect hate speech within tweets and identify tweets associated with criminal activities. The findings from both these processes will be intricately examined to establish correlations and patterns, offering valuable insights into how online hate speech may impact or manifest itself in the perpetration of real-world crimes.

2. Literature Review

Hate speech on social media contributes to worldwide crime. It targets women, minorities, and public servants, according to studies like

one that examined tweets from UK lawmakers [1]. To combat it, researchers are creating instruments [2]. Nevertheless, difficulties arise from a lack of data, models that are biased and unable to handle context, and the discrepancy between automated detection and human comprehension [3].

Researchers use Machine Learning (ML) and Natural Language Processing (NLP) to automatically detect hate speech on Twitter. Studies explore combining linguistic features with user and network data [4, 5]. While ML shows promise [6, 7, 8], challenges include context, sarcasm, and skewed data [6]. Specific algorithms like J48graft with tailored parameters perform well, and some suggest merging hate speech and offensive content categories for better accuracy [8, 9]. Sentiment analysis techniques also prove effective in identifying cyberbullying in tweets [10]. Overall, these studies showcase the potential of ML and NLP for tackling harmful content on social media platforms.

NLP techniques are being used to deal with hate speech on the internet, though they have limitations due to imprecise definitions and context [11]. The finest algorithms for hate speech identification are determined by comparing algorithms such as Random Forests, SVMs, and RNNs [12]. Gradient Boosting, Random Forests, and SVMs are effective for multi-language detection, according to research utilizing multilingual annotators [13]. While existing approaches are effective for certain hate categories, their wider applicability is hampered by a lack of data [14]. Opportunities emerge when diversified data and context-aware analysis are combined with deep learning algorithms. Deep learning experiments on the Arabic language detect hate speech with high precision [15].

The accuracy of hate speech identification is increased when text and picture analysis are combined (multimodal technique) as opposed to text analysis alone [16]. Better identification rates result from the efficient capture of context and local information in text by deep learning models such as CNNs and Bi-GRUs [16, 17, 18]. These studies underscore the need for continued study in this field and emphasize the significance of taking context into account for reliable hate speech identification [19].

It appears promising to combine sentiment analysis and hate speech detection to potentially improve model performance [20]. Novel approaches address issues such as sparse data in other languages [21] and separating hate speech from foul language [23]. Hate speech detection is a strong suit for deep learning models, particularly pre-trained models like BERT [22, 24]. Recent advances in NLP have led to the development of the Transformer architecture, which bases future language modeling research on self-attention processes for a range of tasks [25].

Study investigates online social networks through a variety of perspectives. One study developed a restaurant recommendation engine based on user visit patterns and reviews [27]. Another study examined public worries about cybersecurity during COVID-19 by sentiment analysis of Twitter tweets [28]. This survey discovered rising awareness and optimism about combating cybercrime. Finally, research looked into investor sentiment in the Shanghai stock market during the pandemic's early phases [29]. It discovered a negative relationship between investor sentiment and the stock market index, demonstrating the impact of behavioral finance on markets during crises.

3. Methodology

This section deals on the Model Implementation in different steps 1: Data Collection - Scrape Data from Twitter 2: Data Preprocessing 3: First BERT Model - Hate Speech Detection 4: Filter Out Hate-Labeled Tweets 5: Second BERT Model - Crime-Related Tweet Classification 6: Output - Crime Labeled Tweets. Figure 1. Shows the Model architecture of the proposed research methodology. Figure 2 shows the transformer model and Figure 3 shows the Bertbase mode.

In each of the attention heads several steps happen as shown in the Figure 4 where the input is first embedded with values and the corresponding queries, keys and values are generated by multiplying the input to a matrix. Then the score is calculated by multiplying queries and keys. Finally the sum is calculated by combining the multiplied values of softmax and value(v).

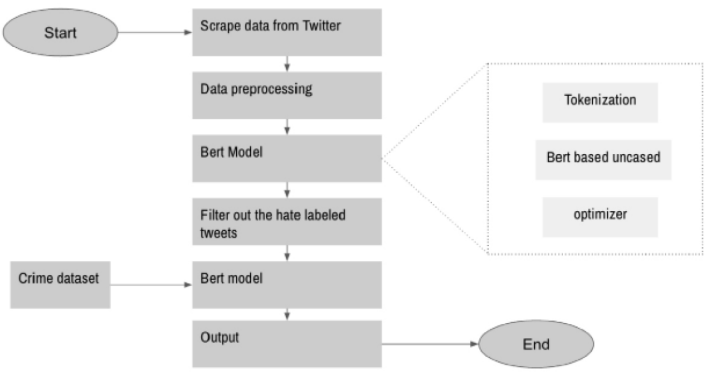

Figure 1

Model Architecture

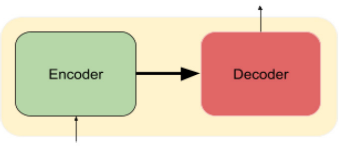

Figure 2

Transformer model

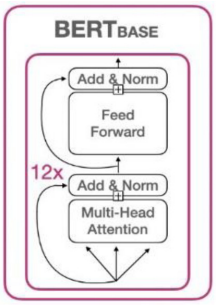

Figure 3

Bertbase model

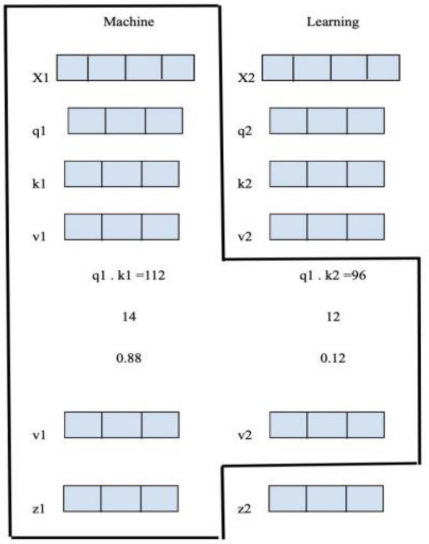

Figure 4

Attention Head

Model Explanation: To begin with, we use the processed data that has been cleansed, tokenized and put into an appropriate form before feeding it into the system/model for further analysis. The dataset includes tweets captured in line with certain conditions such as alignment with the project's goals, difference of opinion, and the most up-to-date information. Later, these purified tweets are fed into a ready-for-training BERT module. One significant feature about architecture of BERT is its ability to go through the text bidirectionally that helps capture contextual information from both sides of language which allows for understanding the intricacies of language.

After the first BERT model makes predictions, you will filter out tweets labeled as "hate speech" from the dataset. This creates a subset of tweets that are likely to contain hate speech. The second model works the same as the previous model. In this model the input will be the labeled hate tweets that were the output from the first model. This model is prepared on the crime dataset and subsequently will group the information tweets as crime-related tweets and non-crime-related tweets. This model is important as through this model we will be able to tell the connection between the hate speech tweets and the crime happening in India.

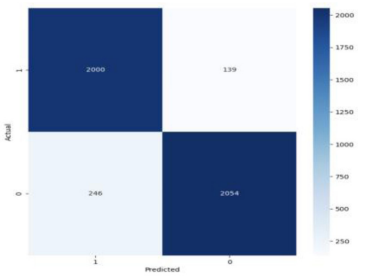

Figure 5
Model Confusion Matrix

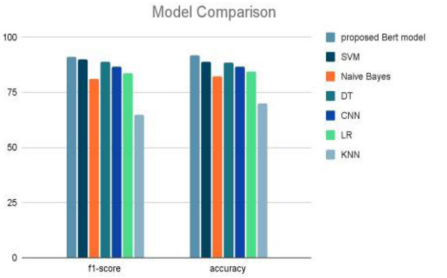

Figure 6
Model Comparison

5. Result

In this paper, we have taken a collection of data sets and also scraped the fresh data/tweets of popular Indian politicians and celebrities and used a multi series BERT model to give out the correlation between the hate speech and crime detection. Comparison was done with other different models as shown in Figure 6 such as CNN. KNN, SVM, RF, LR, NB, DT which are popularly used for detecting hate speech and crime. The multiseries BERT model has shown a promising result in terms of both accuracy and recall with 92% and 93% respectively. The confusion matrix shown in Figure 5 gives the corresponding values of true positive, true negative, false positive and false negative. The accuracy, precision, recall and f1 score can be calculated based on this data which are coming off as 92, 89, 93 and 92 respectively.

As illustrated in Figure 7 and 8, when the model was passed with the dataset collected/scrapped from twitter, it can be seen that states like Delhi, Maharashtra and Tamil Nadu have the higher percentage of chances of crime happening due to hate tweets with Delhi being 20.7%, Maharashtra as 16.3% and Tamil Nadu as 7.4% while states like Arunachal Pradesh, Andaman & Nicobar Island and Meghalaya have lower rate with less the 1% respectively. The crime count is illustrated in Table 2 featuring "State Crime Count Data Distribution" where the highest count is in Delhi with 1394 count and least in Meghalaya with 24 count. Table 1, 2 and 3 represent the examples of crime tweets from the dataset and the predicted tweets respectively. Heat map in Figure 9 and 10 drawn from the results shows the concentration of possible crime happenings in individual states based on the hate speech detected from the people's tweet in India map geographically.

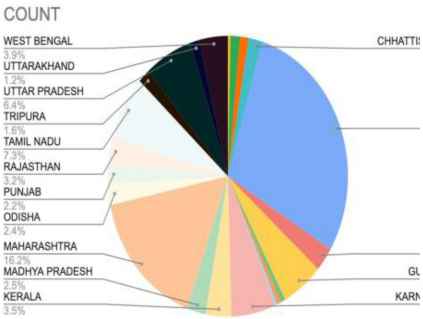

Figure 7

Crime Data distribution of different States of India

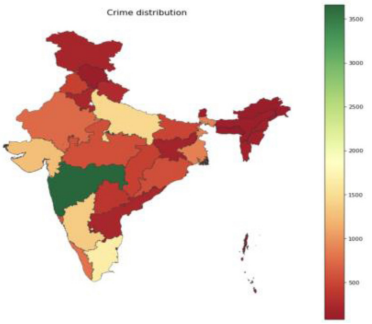

Figure 8

Indian state Crime distribution Heatma

Table 1

Crime Tweets Example

Sr. No	Tweets	Location
1	#Newsunday: 15 -year -old girl shot in Patna |1 minute, see 1 minute in news segment, important news in detail… https://t.co/8cjqckjrul	New Delhi, India
2	Buldana Accident Video: If you have looked back before you hit the uterus ..?CCTV imprisoned the thrilling accident in Buldana #Buldana… https://t.co/mkzcznlu	Maharashtra, India
3	Through conflict with men between women in Kudumbasree; 10 injured #Crime https://t.co/zieg6ippfk	Trivandrum
4	A young man died in suspicious condition after falling from roof in Rohini, police started investigating murder case	New Delhi, India

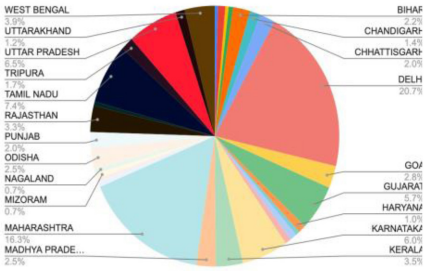

Figure 9

Individual State Crime Count Data Based on the Celebrity Tweets

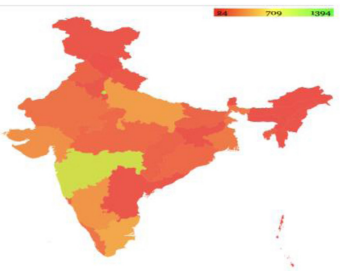

Figure 10

Celebrity tweets Crime Distribution Heatmap

Table 2

State Crime Count Data distribution

Sr. no	Location	Count
1	Andaman & Nicobar Island	28
2	Andra Predesh	64
3	Arunanchal Pradesh	29
4	Assam	39
5	Bihar	149
6	Chandigarh	92
7	Chhattisgarh	134
8	Delhi	1394
9	Goa	186
10	Gujarat	380
11	Haryana	68
12	Himachal Pradesh	35
13	Jammu & Kashmir	59
14	Jharkhand	54
15	Karnataka	404
16	Kerala	236
17	Madhya Pradesh	168
18	Maharashtra	1093
19	Manipur	44
20	Meghalaya	24

Contd...

21	Mizoram	46
22	Nagaland	45
23	Odisha	166
24	Punjab	137
25	Rajasthan	219
26	Sikkim	34
27	Tamil Nadu	500
28	Tripura	111
29	Uttar Pradesh	439
30	Uttarakhand	81
31	West Bengal	262

Table 3

Predicted labeled tweets Example

Sr. No	Tweets	labeled	Location
1	"My father served the nation for 25 years as an Army officer and my family is being treated like we are terrorists": Rhea Chakraborty	1	Delhi
2	Sanjay Raut Shiv Sena leader has given me an open threat and asked me not to come back to Mumbai, after Aazadi graffitis	2	Maharashtra
3	Shah Rukh Khan · @iamsrk. Done more than a few songs with. @deepikapadukone … but nothing like a love song done the Atlee way!!! Happy Birthday my friend.	0	none

7. Conclusion

The research aims to address the complex relationship between online hate speech and real-world crimes by implementing a comprehensive approach of Data Collection, Hate Speech Detection, Crime-Related Tweet Classification and Establishing Connections. Through these comprehensive objectives, the project has contributed valuable knowledge about the intersection of hate speech and crime in the digital age, ultimately aiding in the development of more effective strategies for addressing and mitigating the harmful consequences of online hate speech. This research paper proposes a multi-series methodology to investigate the complex

linkages between online hate speech and real world acts of violence. Furthermore, the paper has considered comparing with other different models which are popularly used for hate speech detection and crime detection. Many of these models are specifically used for particular one aspect either hate speech or crime detection. Until this point in time, broad examination and various undertakings have been led regarding the matter of disdain discourse location and counteraction, as well as crime detection and prevention. The multiseries BERT model used for this research for showing the correlation between these two has shown a promising result in terms of both accuracy and recall with 92% and 93% respectively.

References

[1] Agarwal, Pushkal, et al. "Hate speech in political discourse: A case study of UK MPs on Twitter." Proceedings of the 32nd ACM conference on hypertext and social media (2021).

[2] Wang, Chih-Chien, Min-Yuh Day, and Chun-Lian Wu. "Political Hate Speech Detection and Lexicon Building: A Study in Taiwan." IEEE Access 10 : 44337-44346 (2022).

[3] Velankar, Abhishek, Hrushikesh Patil, and Raviraj Joshi. "A review of challenges in machine learning based automated hate speech detection" (2022). arXiv preprint arXiv:2209.05294.

[4] Koushik, Garima, K. Rajeswari, and Suresh Kannan Muthusamy. "Automated hate speech detection on Twitter." 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA). IEEE (2019).

[5] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." Proceedings of the NAACL student research workshop (2016).

[6] Mullah, Nanlir Sallau, and Wan Mohd Nazmee Wan Zainon. "Advances in machine learning algorithms for hate speech detection in social media: a review." IEEE Access 9 : 88364-88376 (2021).

[7] Abro, Sindhu, et al. "Automatic hate speech detection using machine learning: A comparative study." *International Journal of Advanced Computer Science and Applications* 11.8 (2020).

[8] Elisabeth, Damayanti, Indra Budi, and Muhammad Okky Ibrohim. "Hate code detection in Indonesian tweets using machine learning approach: a dataset and preliminary study." 2020 8th Interna-

tional Conference on Information and Communication Technology (ICoICT). IEEE (2020).

[9] Watanabe, Hajime, Mondher Bouazizi, and Tomoaki Ohtsuki. "Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection." IEEE access 6 : 13825-13835 (2018).

[10] Mody, Akankshi, et al. "Identification of potential cyber bullying tweets using hybrid approach in sentiment analysis." 2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT). IEEE (2018).

[11] Alrehili, Ahlam. "Automatic hate speech detection on social media: A brief survey." 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA). IEEE (2019).

[12] Putri, T. T. A., et al. "A comparison of classification algorithms for hate speech detection." Iop conference series: Materials science and engineering. Vol. 830. No. 3. IOP Publishing (2020).

[13] Oriola, Oluwafemi, and Eduan Kotzé. "Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets." IEEE Access 8 : 21496-21509 (2020).

[14] Alkomah, Fatimah, and Xiaogang Ma. "A literature review of textual hate speech detection methods and datasets." Information 13.6 : 273 (2022).

[15] Alshalan, Raghad, and Hend Al-Khalifa. "A deep learning approach for automatic hate speech detection in the saudi twittersphere." *Applied Sciences* 10.23 : 8614 (2020).

[16] Boishakhi, Fariha Tahosin, Ponkoj Chandra Shill, and Md Golam Rabiul Alam. "Multi-modal hate speech detection using machine learning." 2021 IEEE International Conference on Big Data (Big Data). IEEE (2021).

[17] Khan, Shakir, et al. "HCovBi-caps: hate speech detection using convolutional and Bi-directional gated recurrent unit with Capsule network." IEEE Access 10 : 7881-7894 (2022).

[18] Roy, Pradeep Kumar, et al. "A framework for hate speech detection using deep convolutional neural network." IEEE Access 8 : 204951-204962 (2020).

[19] Pérez, Juan Manuel, et al. "Assessing the impact of contextual information in hate speech detection." IEEE Access 11 : 30575-30590 (2023).

[20] Pérez, Juan Manuel, et al. "Assessing the impact of contextual information in hate speech detection." IEEE Access 11 : 30575-30590 (2023).

[21] Mozafari, Marzieh, Reza Farahbakhsh, and Noel Crespi. "Cross-lingual few-shot hate speech and offensive language detection using meta learning." IEEE Access 10 : 14880-14896 (2022).

[22] Sharmila, P., et al. "PDHS: Pattern-Based Deep Hate Speech Detection With Improved Tweet Representation." IEEE Access 10 : 105366-105376 (2022).

[23] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." Proceedings of the international AAAI conference on web and social media. Vol. 11. No. 1 (2017).

[24] Malik, Jitendra Singh, Guansong Pang, and Anton van den Hengel. "Deep learning for hate speech detection: a comparative study" (2022). arXiv preprint arXiv:2202.09517.

[25] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[26] Vivek, Meghashyam, and Boppuru Rudra Prathap. "Crime Analysis and Forecasting using Twitter Data in the Indian Context." 2023 International Conference on Artificial Intelligence and Smart Communication (AISC). IEEE (2023).

[27] Nayak, S. R., Arora, V., Sinha, U., & Poonia, R. C. A statistical analysis of COVID-19 using Gaussian and probabilistic model. *Journal of Interdisciplinary Mathematics*, 24(1), 19–32 (2021).

[28] Khandelwal, Sonal & Chaudhary, Aanyaa. COVID-19 pandemic & cyber security issues: Sentiment analysis and topic modeling approach, *Journal of Discrete Mathematical Sciences and Cryptography,* 25:4, 987-997 (2022).

[29] Teng, Yin-Pei & Lan, Li-Zhen. Investor sentiment at the initial stage of public emergency: Evidence from Shanghai stock market, *Journal of Statistics and Management Systems,* 27:1, 1–8 (2024).