Multilingual Hate Speech and Offensive Language Detection

Sapthak Mohajon Turjya
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
sapthakmohajon6@gmail.com

Rina Kumari
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
rina.kumarifcs@kiit.ac.in

Sujata Swain
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
sujata.swainfcs@kiit.ac.in

Anjan Bandyopadhyay
School of Computer Engineering
Kalinga Institute of Industrial Technology
Bhubaneswar, India
anjan.bandyopadhyayfcs@kiit.ac.in

2023 OITS International Conference on Information Technology (OCIT) | 979-8-3503-5823-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/OCIT59427.2023.10431222

***Abstract*—Internet and social media usage has skyrocketed over the past two decades, changing how people communicate with one another on a basic level. Numerous favourable results have resulted from this. The risks and harms that come with it are also there. It is impossible for humans to control the amount of damaging content, such as hate speech, that is available online. Researching automated methods for hate speech identification has drawn more attention from academics. Through the creation of a single homogeneous dataset, we investigate various publicly accessible datasets in this work. We establish a baseline model and enhance model performance scores using various optimisation strategies after classifying them into two categories: hate or non-hate. After achieving a competitive performance score, we develop a tool that, using the same feedback, quickly locates and evaluates a page with an effective measure. This tool then retrains our model using the new data. In three languages: English, German, and Spanish. We demonstrate the superior performance of our multilingual approach. In comparison to most monolingual models, this results in performance that is equal to or better.**

***Index Terms*—Hate speech, Non-hate speech, Label classification, BERT**

I. Introduction

In today's interconnected world, online platforms have become breeding grounds for hate speech, posing significant threats to individuals and communities. Hate speech, characterized by discriminatory, offensive, or harmful language targeted at specific groups based on attributes such as race, religion, ethnicity, gender, or sexual orientation, has been recognized as a pervasive issue across various digital spaces [1], [2].

Detecting and addressing hate speech is crucial to maintaining a safe and inclusive online environment. While hate speech detection has received considerable attention in recent years, most research has primarily focused on English-centric approaches. However, with the rapid globalization of the digital landscape, it is essential to develop effective detection methods that can identify and combat hate speech in a multilingual context [2]–[4]. Such an approach recognizes the diversity of languages used in online communication and acknowledges the unique linguistic characteristics and cultural nuances associated with hate speech across different languages. Nowadays, it generally comes in code-mixed languages. We list some of the hate speech examples below:

Examples of hate speech

Text-instance-1:"Main kal movie dekhne jaa rahi thi but I missed the bus."
Translation:I was going to watch a movie yesterday but I missed the bus.
Hate Speech Label:"Non Hate"

Text-instance-2:" Look ye politicians suvar jaise baithe rehte hain sirf money ke liye kaam karte hain. They don't care about the public ."
Translation: Look these politicians are sitting like pigs and they only work for money. They don't care about the public.;
Hate Speech Label:" Hate"

The rise of social media platforms, discussion forums, and online communities has facilitated the spread of hate speech beyond linguistic boundaries. Hate speech can be found in various languages, posing significant challenges for automated detection systems. Multilingual hate speech detection involves the development of robust models that can accurately identify and classify hateful content across a wide range of languages.

The primary objective of this research paper is to address the need for effective multilingual hate speech detection methods. By extending the focus beyond English and considering multiple languages, we aim to develop models that are capable of detecting hate speech in a diverse linguistic landscape. The proposed models should be sensitive to the unique linguistic patterns, cultural references, and contextual cues present in different languages. This paper provides a comprehensive analysis of the existing literature on hate speech detection, with a particular emphasis on multilingual approaches. It ex-

979-8-3503-5823-0/23/$31.00 © 2023 IEEE

Authorized licensed use limited to: University of Plymouth. Downloaded on August 28,2024 at 07:37:00 UTC from IEEE Xplore. Restrictions apply.

plores the challenges associated with detecting hate speech in multiple languages and investigates various machine learning and deep learning techniques that have been employed in this domain. Furthermore, we present an evaluation of our proposed multilingual hate speech detection model and discuss its performance on representative datasets.

By addressing the research gap in multilingual hate speech detection, this study aims to contribute to the development of more inclusive and effective mechanisms for detecting and combating hate speech across languages. By fostering safer and more respectful online environments, we can protect the rights and dignity of individuals, promote healthy dialogue, and work towards a more tolerant and inclusive society.

In this research study, our focus is to examine the current strategies employed in addressing the issue of hate speech. To gather relevant information, we selected Twitter as our primary source, considering its widespread usage as a popular social media platform.

The structure of this research paper is organized as follows: Section II provides a comprehensive review, highlighting relevant studies and approaches. In Section III, we present the methodology employed for our multilingual hate speech detection task.Section IV discusses the dataset in detail. Moving forward, Section V details the experimental setup, encompassing the selection of evaluation datasets, training procedures, and model performance evaluation. In Section VI, error analysis is discusses. Finally, in Sections VII, VIII, we conclude our work by summarizing the main contributions of this research, discussing its implications, and suggesting potential avenues for future exploration, respectively.

We formalize our research by responding to the following questions:

RQ1: In a context without regard to language, how is hate speech different from everyday communication?

RQ2: How effective are the various model architectures in detecting hate speech when trained on a variety of languages?

RQ3: In comparison to a model trained on the complete set of languages, how well does a model that was trained on a particular language family perform on a language that is a member of that family?

II. Related Work

The utilization of machine learning and natural language processing (NLP) techniques has greatly advanced the identification of hate speech on digital platforms. The scientific community has extensively researched machine learning and deep learning approaches for automatically detecting hate speech and offensive material [5]–[8]. Traditional machine learning methods often use features like word and character n-grams. Supervised machine learning approaches utilize various text mining-based features, including surface features, sentiment analysis, lexical resources, linguistic features, knowledge-based features, and metadata related to users and platforms. However, these approaches heavily rely on well-defined feature extraction methods [9]–[11]. In recent years, text representation and deep learning techniques, such as convolutional neural networks (CNNs), bidirectional long short-term memory networks (Bi-LSTMs), and BERT, have significantly improved the performance of models designed to detect hate speech and offensive content. These methods have the ability to capture the underlying linguistic patterns associated with hate speech and offensive content, leading to more accurate identification of such content. As a result, they have become the state-of-the-art approaches for detecting hate speech and offensive content [12]–[14]. The advances in machine learning and NLP for hate speech detection have the potential to significantly reduce the amount of hate speech and offensive content on digital platforms. This is important because hate speech and offensive content can have a negative impact on individuals and society as a whole. By reducing the amount of hate speech and offensive content, these techniques can help to create a more positive and inclusive online environment. [15]–[17] Here are some of the benefits of using machine learning and NLP for hate speech detection:

- Increased accuracy: Machine learning and NLP techniques can learn the underlying linguistic patterns of hate speech and offensive content, which allows them to more accurately identify these types of content.
- Reduced human bias: Machine learning and NLP techniques are not susceptible to human bias, which can be a problem with traditional methods of hate speech detection.
- Scalability: Machine learning and NLP techniques can be scaled to handle large volumes of data, which is important for detecting hate speech on large social media platforms.

The use of machine learning and NLP for hate speech detection is still a developing field, but it has the potential to significantly reduce the amount of hate speech and offensive content on digital platforms [18]. These techniques have the potential to create a more positive and inclusive online environment.

III. Methodology

In our research, we employed a pre-trained BERT transformer model to effectively identify inflammatory language and hate speech. Fig. 1 illustrates the utilization of BERT in detecting objectionable language and hate speech. BERT, which stands for Bidirectional Encoder Representations from Transformers, is a transformer encoder stack that has undergone extensive training on a large English corpus. There are two variants available: $\text{BERT}_{\text{base}}$ and $\text{BERT}_{\text{large}}$, distinguished by the number of transformer layers. $\text{BERT}_{\text{base}}$ comprises 12 layers, while $\text{BERT}_{\text{large}}$ comprises 24 layers. Additionally, $\text{BERT}_{\text{large}}$ possesses larger feed-forward networks with hidden representations of 1024 and 16 attention heads, compared to $\text{BERT}_{\text{base's}}$ 768 hidden representations and 12 attention heads.

Similar to the standard transformer model, BERT takes word sequences as input and undergoes a series of operations in each layer, including self-attention and passing the results through a feedforward network before transferring them to the

Authorized licensed use limited to: University of Plymouth. Downloaded on August 28,2024 at 07:37:00 UTC from IEEE Xplore. Restrictions apply.

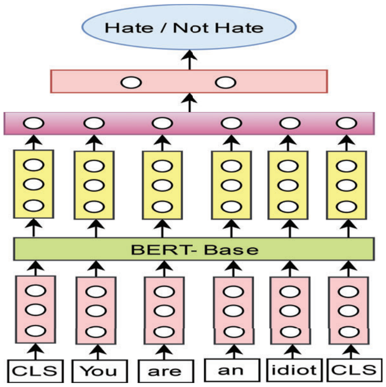


Fig. 1: Utilizing BERT for classifying sequences in Hate Speech Data.

subsequent encoder. The embeddings from BERTbase have 768 hidden units. The BERT model processes word/token sequences, accommodating a maximum length of 512, and generates encoded representations with a dimensionality of 768. The pre-trained BERT models excel in word representation due to their extensive training on diverse sources such as Wikipedia and books. To adapt the general pre-trained BERT model to specific tasks, fine-tuning becomes necessary. During the fine-tuning process, the parameters of the pre-trained BERT model are updated using a labeled dataset comprising hate speech and offensive content. When fine-tuning for downstream sentence classification tasks, minimal modifications are made to the BERTbase configuration. In this particular architecture, only the output of the [CLS] token is utilized. The [CLS] token's output is derived from the 12th transformer encoder and comprises 768 dimensions. It is then passed through a fully connected neural network, and for sentence classification, the softmax activation function is applied. This enables BERT to determine whether a given sentence contains inappropriate or hateful content. In our study, which encompasses data in English, German, and Spanish, we employed a pre-trained multilingual BERTbase model. The architecture of the multilingual BERT model aligns with the standard BERT model but is pre-trained on diverse multilingual Wikipedia language sources, enabling it to accommodate multiple languages. Overall, our methodology, which involves leveraging a pre-trained BERT transformer model, fine-tuning it on a labeled dataset, and utilizing a fully connected neural network for sentence classification, proved highly effective in identifying inflammatory language and hate speech across multiple languages. This approach capitalizes on BERT's strength in word representation and the adaptability of the multilingual model to handle diverse languages. Consequently, it provides a robust and versatile method for detecting objectionable content.

The models created for this study are designed to be applied in a virtual environment. The pursuit of a cutting-edge model that can be retrained in a certain amount of time in a resource-constrained environment becomes crucial. Analysing both the prediction time and the training time of the model is essential. Our model takes very less time to train compared to the existing models.

TABLE I: The detailed data description

Columns	Description
tweet_reference	A unique identifier for each tweet.
tweet_text	The complete text content of the tweets.
task	A label indicating whether the tweet is categorized as HOF or NOT.

TABLE II: Language Dataset Statistics: English, German, and Spanish Training Data

Language	**Hate**	**Non Hate**
English	2426	2570
German	1522	1510
Spanish	324	830

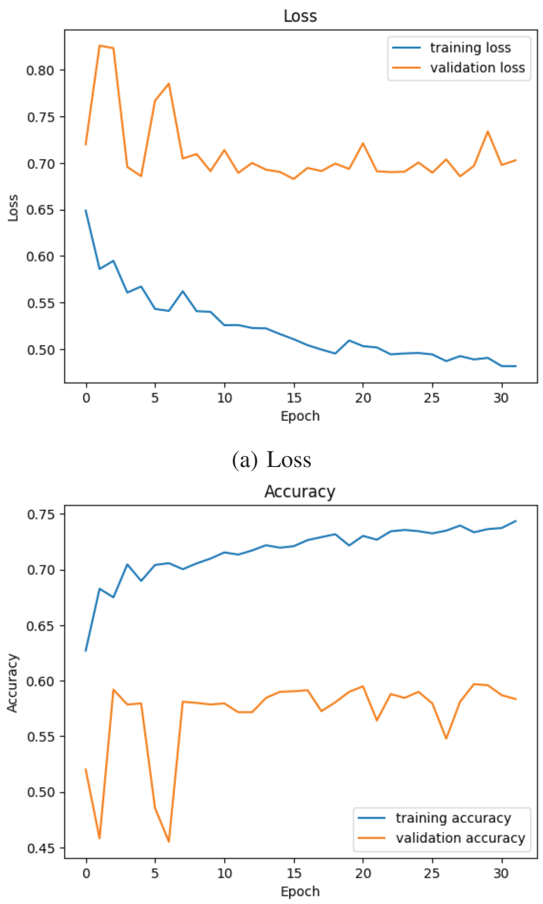
IV. Dataset

Additionally, we integrated pre-existing datasets that have been utilized by other researchers to target various instances of hate speech. While these datasets may individually be modest in size, their consolidation allows us to create a substantial corpus of hate speech sequences. By combining offensive material from five distinct categories, namely sexual orientation, religion, nationality, gender, and ethnicity, we aim to develop models capable of categorizing text content as either hateful, abusive, or neither. The dataset used in our study consists of approximately ten thousand samples, and in order to maintain high variance and low bias, we refrain from sub-categorizing hate classes. Table I depicts the dataset attributes and Table II depicts the dataset distributions in detail.

V. Results and Analysis

Our hate speech detection project achieved an overall accuracy of 72.67%, with a precision of 72.77%, recall of 72.90%, and F1 score of 72.64%. These results demonstrate that our model is able to detect hate speech with reasonable accuracy across all three languages. Upon analysing the results specifically for English, we found that the accuracy was 83.59%, precision was 83.71%, recall was 83.68%, and F1 score was 83.59%. These results are significantly higher compared to the overall results, indicating that our model is particularly effective in detecting hate speech in English. In the confusion matrix, the true negative score is leading with 2102 points, followed by 2074 in true positive. However, the results for German were comparatively lower, with an accuracy of 62.81%, precision of 62.97%, recall of 62.83%, and F1 score of 62.72%. The confusion matrix for German shows that the true positive score has 1031 points, followed by true negative with 876 points. This indicates that our model has difficulty in accurately detecting hate speech in German. Similarly, for Spanish, our model achieved an accuracy of 76.04%, precision of 74.79%, recall of 59.83%, and F1 score of 60.11%. The confusion matrix for Spanish shows that the true positive score has the highest score with 802 points, followed by false negative with 249 points. This indicates that our model is

Authorized licensed use limited to: University of Plymouth. Downloaded on August 28,2024 at 07:37:00 UTC from IEEE Xplore. Restrictions apply.



(a) Loss

(b) Accuracy

Fig. 2: Loss and Accuracy graph

better at detecting true positives in Spanish, but struggles with false negatives.

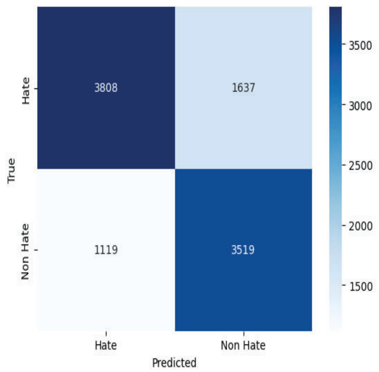


Fig. 3: Confusion Matrix for all languages

Proof. We have generated two graphs to analyze the performance of our model. We show these graphs in Fig. 2. The first graph represents the loss versus the number of training iterations, while the second graph illustrates the accuracy versus the number of iterations. In the loss graph, we observe a consistent decrease in the loss during the training phase. This indicates that our model is effectively learning and optimizing its parameters to minimize the discrepancy between predicted and actual values. However, in the validation phase, we notice some fluctuations in the loss values, with occasional peaks and dips. Towards the end of the graph, there is a slight upward trend in the validation loss. These fluctuations suggest that the model's performance on unseen data may vary during different stages of training. Despite these fluctuations, the overall trend of the validation loss indicates that the model is learning and generalizing reasonably well. Turning to the accuracy graph, we can observe a positive trend in both the training and validation phases. As the training progresses, the accuracy steadily increases, indicating that the model is becoming more proficient at correctly classifying instances within the training dataset. Similarly, the validation accuracy also shows a consistent improvement over time, suggesting that the model is successfully generalizing its learned patterns to unseen data. The increasing accuracy throughout both phases is an encouraging sign, demonstrating the effectiveness of our model in accurately classifying instances. In summary, the loss graph highlights the learning process of our model, with the training loss decreasing steadily. The validation loss, while exhibiting some fluctuations, demonstrates an overall decreasing trend, indicating satisfactory generalization. Furthermore, the accuracy graph demonstrates a continuous improvement in both training and validation accuracy, affirming the model's ability to correctly classify instances. These results provide evidence of the effectiveness and learning capabilities of our model. Overall, our project has achieved reasonable success in detecting hate speech across the three languages using machine learning and NLP. While the results for English were particularly strong, there is room for improvement in the accuracy of hate speech detection in German and Spanish. Further analysis can be done to explore the reasons behind the lower accuracy for German and Spanish and identify ways to improve the model's performance in those languages. □

VI. ERROR ANALYSIS

Fig. 3 presents the confusion matrix for BERT model's performance on English, German, and Spanish datasets. The binary classification for English was the best-performing model. However, the German model misclassified a significant number of hate-speech labels. In terms of offensive content evaluation, the English model had an accuracy of 83%, while the Spanish model had mediocre results with an accuracy of 70%. The multilingual-BERT model struggled to identify hate speech and offensive content for the German language, misclassifying most of the labels as "NOT HATE" and exhibiting poor performance on those datasets.

VII. CONCLUSION

In this study, we utilized pre-existing bi-directional encoder representations trained using transformers, specifically BERT and multilingualBERT, to effectively detect hate speech in English, German, and Spanish languages. Our analysis involved a

Authorized licensed use limited to: University of Plymouth. Downloaded on August 28,2024 at 07:37:00 UTC from IEEE Xplore. Restrictions apply.

comprehensive comparison between BERT and other machine learning and neural network classification techniques. The results of our investigation demonstrated that fine-tuning pre-trained BERT and multilingualBERT models for hate-speech text classification tasks resulted in significant enhancements in macro F1 score and accuracy metrics when contrasted with conventional word-based machine learning approaches. Furthermore, the provided dataset included labels for both hate speech and offensive content within the same sentence, highlighting the interconnectedness between these two tasks. To establish a robust relationship between these tasks, joint learning models can be employed. By adopting a deep joint classification model, a better understanding of the provided datasets can be achieved, thereby aiding in the accurate identification and classification of hate speech and offensive content. In conclusion, our study showcases the effectiveness of leveraging pre-trained BERT and multilingualBERT models for hate speech detection across multiple languages. The superior performance of these models, in conjunction with the potential of joint learning approaches, contributes to the advancement of hate speech detection techniques and the mitigation of harmful online content.

VIII. Future Scope

Improving accuracy for German and Spanish: While our multilingual BERT model achieved an accuracy of approximately 83% for English, the accuracy for German and Spanish was lower, at 54% and 72%, respectively. To address this, we can explore ways to improve the model's performance on these languages, such as by fine-tuning the model or collecting more training data. This can help ensure that our hate speech detection system is effective across different languages.

Expanding language coverage: While our project covered three major languages, there are many other languages spoken around the world. To address this, we can explore the possibility of extending our hate speech detection system to other languages, which can help address hate speech in a global context. This can involve building models for new languages or adapting our existing model to support additional languages.

Real-time detection: Currently, our project involves analysing text data after it has been created. However, hate speech can have real-world consequences, and detecting it in real-time can help prevent harm. To address this, we can explore the possibility of developing a real-time hate speech detection system, which can be used by social media platforms or other online communities to monitor and address hate speech as it occurs.

References

[1] Aluru, S. S., Mathew, B., Saha, P., & Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. arXiv preprint arXiv:2004.06465.

[2] Corazza, M., Menini, S., Cabrio, E., Tonelli, S., & Villata, S. (2020). A multilingual evaluation for online hate speech detection. ACM Transactions on Internet Technology (TOIT), 20(2), 1-22.

[3] Vashistha, N., & Zubiaga, A. (2020). Online multilingual hate speech detection: experimenting with Hindi and English social media. Information, 12(1), 5.

[4] Röttger, P., Seelawi, H., Nozza, D., Talat, Z., & Vidgen, B. (2022). MULTILINGUAL HATECHECK: Functional Tests for Multilingual Hate Speech Detection Models. arXiv preprint arXiv:2206.09917.

[5] Markov, I., Ljubešić, N., Fišer, D., & Daelemans, W. (2021, April). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (pp. 149-159).

[6] Al-Hassan, A., & Al-Dossari, H. (2019, February). Detection of hate speech in social networks: a survey on multilingual corpus. In the 6th international conference on computer science and information technology (Vol. 10, pp. 10-5121).

[7] Chakravarthi, B. R. (2020, December). HopeEDI: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (pp. 41-53)

[8] Chiril, P., Benamara, F., Moriceau, V., Coulomb-Gully, M., & Kumar, A. (2019, July). Multilingual and multitarget hate speech detection in tweets. In Conférence sur le Traitement Automatique des Langues Naturelles (TALN-PFIA 2019) (pp. 351-360). ATALA.

[9] Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). Information Systems, 105, 101584.

[10] Kovács, György, Pedro Alonso, and Rajkumar Saini. "Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources." SN Computer Science 2 (2021): 1-15.

[11] Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. Expert Systems with Applications, 166, 114120.

[12] Kapil, P., & Ekbal, A. (2020). A deep neural network based multi-task learning approach to hate speech detection. Knowledge-Based Systems, 210, 106458.

[13] Kshirsagar, R., Cukuvac, T., McKeown, K., & McGregor, S. (2018). Predictive embeddings for hate speech detection on twitter. arXiv preprint arXiv:1809.10644.

[14] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). 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 International Publishing.

[15] Pitsilis, G. K., Ramampiaro, H., & Langseth, H. (2018). Effective hate-speech detection in Twitter data using recurrent neural networks. Applied Intelligence, 48, 4730-4742.

[16] Mosca, E., Wich, M., & Groh, G. (2021, June). Understanding and interpreting the impact of user context in hate speech detection. In Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media (pp. 91-102).

[17] Arango, A., Pérez, J., & Poblete, B. (2022). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). Information Systems, 105, 101584.

[18] Sohn, H., & Lee, H. (2019, November). Mc-bert4hate: Hate speech detection using multi-channel bert for different languages and translations. In 2019 International Conference on Data Mining Workshops (ICDMW) (pp. 551-559). IEEE.

Authorized licensed use limited to: University of Plymouth. Downloaded on August 28,2024 at 07:37:00 UTC from IEEE Xplore. Restrictions apply.