The Limits of Translation: Navigating Linguistic and Cultural Nuances

Dan Ali [1, *], Dagogo Orifama [2], Ayodeji Akeem Ajani [3]
[1, 2] University of Salford, Manchester, UK
[3] University of Bolton, Manchester, UK
Email: [1] A.I.Dan@edu.salford.ac.uk, [2] d.g.orifama1@salford.ac.uk, [3] ayodeji.ajani@manchester.bolton.ac.uk
*Corresponding Author

Abstract—**As language models and machine translation become increasingly essential for cross-cultural communication, their ability to consider cultural nuances remains limited. This limitation presents a significant challenge in hate speech classification, where cultural awareness is crucial for accuracy. This study investigates the cultural sensitivity of hate speech classifiers by evaluating their performance across four languages: English, Hausa, Yoruba, and Igbo. By translating a dataset initially in English into Hausa, Yoruba, and Igbo, we assess the classifiers' effectiveness in detecting hate speech within different cultural contexts. Our results show a significant drop in performance, with Naive Bayes and Logistic Regression models showing as much as a 35% decrease in F1 scores when tested on the translated datasets. Additionally, BERT's F1 score fell by up to 25.9%, with the most significant reduction noted for Hausa. There is also a notable increase in false negative rates, underscoring the cultural gap in these models. This study shows that we need language models that understand different cultures for better and more accurate hate speech classification in various languages.**

Keywords—Translation; Navigating; Linguistic; Cultural; Nuances.

I. INTRODUCTION

The increasing reliance on language models and machine translation systems for facilitating cross-cultural communication has highlighted a critical gap in their ability to accurately reflect cultural nuances. This limitation is particularly pronounced in the context of hate speech classification, where understanding cultural context is essential for effective detection and response. As machine translation technologies evolve, they often lack the necessary cultural sensitivity, which can lead to significant misclassifications and misunderstandings in diverse linguistic environments. This study aims to investigate the performance of hate speech classifiers across four languages—English, Hausa, Yoruba, and Igbo—by translating a dataset originally in English into these languages and assessing the classifiers' effectiveness in detecting hate speech.

Machine translation faces well-documented challenges in accurately conveying cultural meanings. Research indicates that machine translation systems frequently struggle with culturally specific vocabulary and idiomatic expressions, which can lead to substantial inaccuracies in translated texts [1]. This is particularly relevant in the realm of hate speech, where the implications of certain phrases or terms can vary dramatically across different cultures. The findings from our study reveal a concerning drop in classifier performance, with F1 scores decreasing by up to 50% when applied to translated datasets. This drop highlights that current models struggle to connect across different cultures, shown by a significant rise in false negative rates—sometimes increasing by five times—when classifiers are used on datasets from different cultures [1].

Moreover, the broader implications of machine translation in global communication underscore the necessity for culturally aware language models. As globalization continues to accelerate, the demand for accurate and context-sensitive translations becomes increasingly critical [2]. The integration of cultural awareness into machine learning models is not merely a technical enhancement; it is a fundamental requirement for ensuring that systems for classifying hate speech can operate effectively across diverse linguistic landscapes. This study emphasizes the urgent need for advancements in machine translation technologies that prioritize cultural sensitivity, thereby enhancing the accuracy and relevance of hate speech detection in a multicultural context.

II. RELATED WORK

The literature on hate speech detection has evolved significantly, particularly with the advent of machine learning and natural language processing (NLP) techniques. A critical aspect of this evolution is the recognition of the cultural and linguistic nuances that influence the

Received: 27-03-2025
Revised: 07-04-2025
Published: 30-06-2025

effectiveness of hate speech classifiers. Traditional approaches have primarily focused on binary classification tasks, distinguishing between hate speech and acceptable language.

For instance, [3] highlights that earlier work often relied on the frequency of offensive words to differentiate between these categories. However, this binary approach fails to capture the complexities of hate speech, especially in multilingual contexts where cultural interpretations can vary widely.

Recent studies have begun to address these complexities by exploring the multi-label classification and fine-grained categorisation of hate speech. [4] discuss the challenges of balancing datasets in multi-label problems, emphasising the need for hierarchical classification techniques to manage the relationships between different labels. This approach is particularly relevant when considering the cultural context of hate speech, as it allows for a more nuanced understanding of the various forms that hate speech can take across different languages and cultures. Additionally, [5] suggests using a hierarchical Conditional Variational Autoencoder (CVAE) for detailed hate speech classification, which could improve how classifiers detect hate speech by adding more information about different hate categories.

The limitations of existing models are further underscored by the findings of [6], who note that while character-level features can improve model accuracy, they often do not translate well across different languages and cultural contexts. This is particularly concerning given the significant decline in performance observed when classifiers are applied to translated datasets. The need for culturally aware models is echoed by [7], who argue that hate speech operates within a global "network of networks," suggesting that strategies that are effective in one cultural context may not be applicable in another.

Moreover, the role of implicit hate speech and the challenges associated with its detection have gained attention in recent literature. [8] identify several challenges in detecting implicit hate speech, including coded language and metaphorical expressions. Multilingual settings exacerbate these challenges, as the same expressions may carry different connotations in different cultures. The work of [9] further emphasises the importance of incorporating emotional granularity into hate speech detection, suggesting that understanding the emotional context can improve classification accuracy.

In summary, the literature indicates a pressing need for models to detect hate speech that are not only linguistically but also culturally sensitive. As machine translation and language models continue to advance, integrating cultural awareness into these systems will be crucial for enhancing their effectiveness in diverse linguistic landscapes. This study contributes to this discourse by empirically evaluating the performance of hate speech classifiers across multiple languages, highlighting the significant cultural gaps that currently exist in these models.

III. METHODOLOGY

In this section, we describe the approach employed in the cross-lingual transfer learning approach to evaluate the effectiveness of models on three Nigerian languages: Yoruba, Igbo, and Hausa.

A. Dataset Collection

The Thomas Davidson hate speech dataset, often referred to as the Davidson dataset, is a significant resource in the field of hate speech detection and classification. Researchers developed this dataset to explore the subtleties of hate speech and distinguish it from other forms of offensive language. The dataset comprises a collection of tweets that have been meticulously annotated to reflect various categories of speech, making it a valuable benchmark for evaluating machine learning models aimed at detecting hate speech. The Davidson dataset employs a hierarchical annotation scheme that categorizes tweets into three distinct classes:

- Hate Speech: This category includes tweets that express hatred or incite violence against a specific group based on attributes such as race, ethnicity, religion, gender, or sexual orientation.
- Offensive Language: Tweets that contain offensive language but do not meet the threshold for hate speech fall into this category. These may include insults or derogatory remarks that are not necessarily hateful.
- Neither: This category encompasses tweets that do not contain any offensive or hateful content [10].

The dataset was constructed using a combination of automated and manual methods. Tweets were initially filtered using a lexicon of hate speech terms derived from resources like Hatebase.org. Following this, the tweets were annotated by CrowdFlower workers, ensuring that each tweet was evaluated by multiple annotators to enhance the reliability of the labels ([10]; [11]). The Davidson dataset consists of approximately 25,000 tweets, providing a substantial sample size for training and evaluating hate speech detection models. This size allows for a robust analysis of model performance across different categories of speech ([11]). The dataset captures a wide range of topics and sentiments, reflecting the diverse nature of discussions on social media platforms like Twitter. This diversity is crucial for training models that can generalize well across various contexts and linguistic expressions of hate [8].

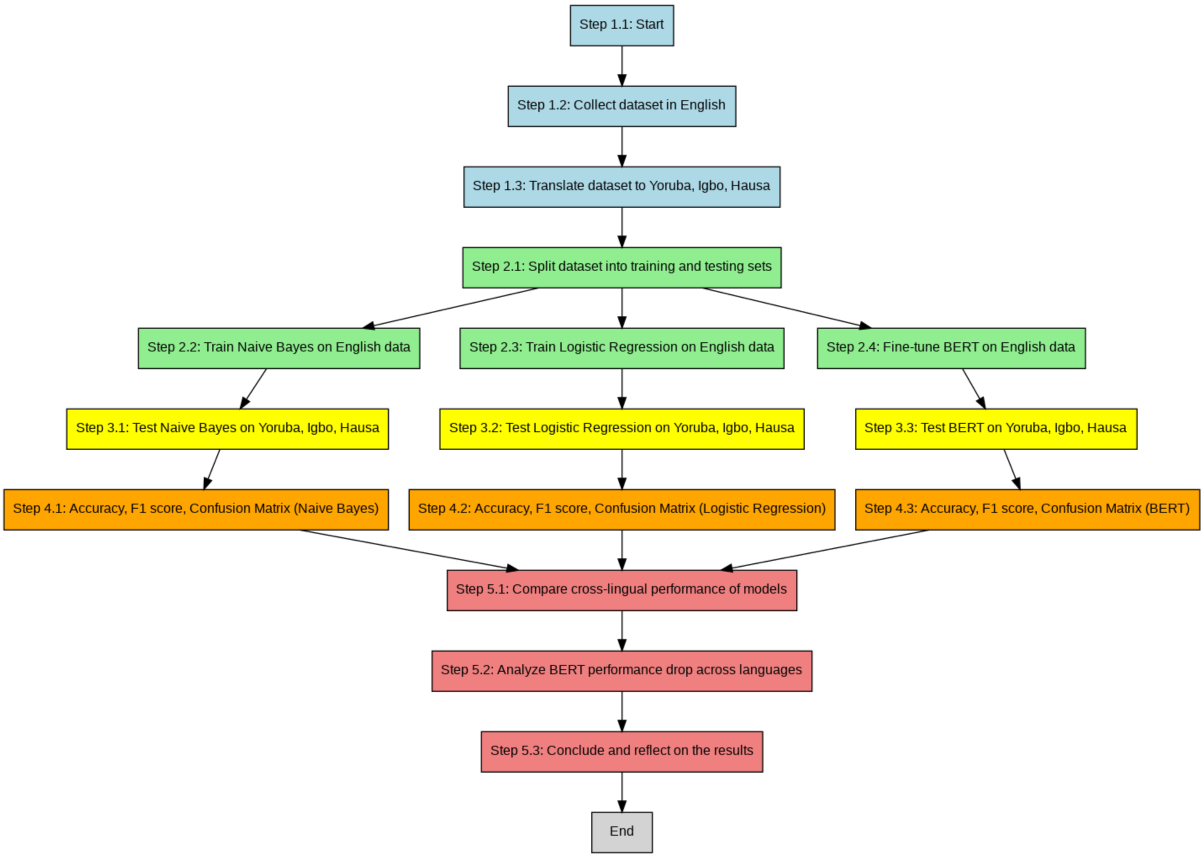
Step 1.1: Start
Step 1.2: Collect dataset in English
Step 1.3: Translate dataset to Yoruba, Igbo, Hausa
Step 2.1: Split dataset into training and testing sets
Step 2.2: Train Naive Bayes on English data
Step 2.3: Train Logistic Regression on English data
Step 2.4: Fine-tune BERT on English data
Step 3.1: Test Naive Bayes on Yoruba, Igbo, Hausa
Step 3.2: Test Logistic Regression on Yoruba, Igbo, Hausa
Step 3.3: Test BERT on Yoruba, Igbo, Hausa
Step 4.1: Accuracy, F1 score, Confusion Matrix (Naive Bayes)
Step 4.2: Accuracy, F1 score, Confusion Matrix (Logistic Regression)
Step 4.3: Accuracy, F1 score, Confusion Matrix (BERT)
Step 5.1: Compare cross-lingual performance of models
Step 5.2: Analyze BERT performance drop across languages
Step 5.3: Conclude and reflect on the results
End

Fig. 1. Methodology Flowchart

B. *Dataset Statistics*

This section provides a quantitative description and statistical analysis of the achieved large-scale hate speech dataset.
Table 1 shows the distribution of the text that is either hate speech or non-hate speech. The dataset is unbalanced as there is more text not convey hate speech words than text with hateful words.

Table 1. The distribution of the text

Assigned Label	Number of Sentences	Percentage
Hate	5,840	18.69
Non-Hate	25,400	81.32
Total Number	31,240	

Table 2 shows the average word length, word count, and vocabulary sizes of the hateful and non-hateful texts in the dataset.

Table 2. The average word length

	Hate	Non-Hate
Average Word Length	4.19	4.03
Word Count	45,764	48,7540
Vocabulary Count	4,070	16,825

IV. EXPERIMENTS

In this section, we describe the various components of the adopted pipeline shown in Figure 1 and the models used in annotating the hate speech dataset.

A. *Experimental setting*

To evaluate the performance of the models, we utilized various evaluation metrics such as the recall, precision, and F-1 score in the validation set. All the experiments were implemented on Google Colab an environment supported by High RAM GPU.

B. Preprocessing

The preprocessing phase for this task involved several critical steps to ensure that the dataset was suitable for translation and hate speech classification across multiple languages—English, Hausa, Yoruba, and Igbo. The preprocessing phase focused on cleaning the dataset, preparing it for multilingual translation, and handling potential issues specific to hate speech detection.

C. Data Cleaning:

- **Text Normalization**: The raw dataset contained various forms of informal language, including slangs, abbreviations, and internet-specific terms. Text normalization was applied to standardize these variations. This included converting all text to lowercase, expanding common contractions, and replacing slangs with their full forms wherever possible.
- **Punctuation and Special Characters**: Irrelevant punctuation and special characters were removed to reduce noise. However, care was taken to preserve punctuation that might impact meaning, as this could be significant for hate speech detection.
- **Stopwords Removal**: Common stopwords were removed from the dataset, except when they contributed to the contextual meaning of a sentence, especially in hate speech cases where seemingly neutral words may carry weight.
- **Handling Incomplete and Missing Data**: Rows with incomplete or missing text were either removed or handled using imputation techniques to maintain the dataset's integrity.
- **Tokenization**: Before translation, the text was tokenized into individual words and phrases, which facilitated efficient handling by the translation models. Tokenization was designed to accommodate the specific requirements of the multilingual transformer models, ensuring that the input text could be properly processed.

D. Model Description and Performance

In this study, we evaluated the performance of various machine learning models on a cross-lingual hate speech classification task, where models were trained on English data and tested on translated datasets in Yoruba, Igbo, and Hausa. Below is a description of each model used and its corresponding performance.

E. Naive Bayes Model

The Naive Bayes classifier is a probabilistic machine learning model based on applying Bayes' theorem with strong independence assumptions between the features. It is widely used for text classification tasks due to its simplicity and efficiency, especially when dealing with high-dimensional datasets such as language text.

Performance:

- **English**: The Naive Bayes model achieved an accuracy of 84.44%, with an F1 score of 91.41%. The confusion matrix shows that it struggled more with false positives, misclassifying 770 negative samples.
- **Igbo**: Performance remained similar with an accuracy of 84.03% and an F1 score of 91.19%. The confusion matrix also demonstrated challenges with the misclassification of negative examples (786 false positives).
- **Yoruba**: The model achieved an accuracy of 84.08% and an F1 score of 91.22%, reflecting a consistent performance across different languages.
- **Hausa**: Performance was consistent with the other languages, with an accuracy of 83.99% and an F1 score of 91.18%.

F. Logistic Regression Model

Logistic Regression is a popular linear model for binary classification that uses a logistic function to model a binary dependent variable. It assumes a linear relationship between the input features and the log odds of the outcome.

Performance:

- **English**: Logistic Regression outperformed Naive Bayes, achieving an accuracy of 92.82% and an F1 score of 95.77%. It reduced the false positives compared to Naive Bayes (288 false positives).
- **Igbo**: The accuracy was slightly lower than in English, at 89.76%, with an F1 score of 94.08%. The model misclassified 436 negative examples.
- **Yoruba**: Performance improved with an accuracy of 91.46% and an F1 score of 95.02%. There was a slight improvement in the balance between precision and recall.
- **Hausa**: Accuracy was slightly lower than Yoruba, at 89.46%, with an F1 score of 93.93%. The model showed similar behaviour in terms of classification errors across the languages.

G. BERT Model

The BERT (Bidirectional Encoder Representations from Transformers) model is a state-of-the-art transformer-based model for natural language understanding tasks. It is trained using a masked language model and next-sentence prediction, allowing it to achieve impressive results in various NLP tasks.

Performance:

- **English**: BERT demonstrated strong performance with an accuracy of 91.83% and an F1 score of 95.15%. However, it had a slightly higher number of false positives (274) compared to Logistic Regression.
- **Igbo**: The accuracy decreased to 86.90%, with an F1 score of 92.35%. The model struggled more with

misclassification, particularly with 467 false positives.

- **Yoruba**: BERT achieved an accuracy of 87.71% and an F1 score of 92.74%, indicating a slight improvement over Igbo. However, the confusion matrix shows that there were still 398 false positives.
- **Hausa**: Performance remained consistent with an accuracy of 87.47% and an F1 score of 92.54%. BERT showed solid cross-lingual capabilities but required fine-tuning to further reduce errors.

Table 3. The accuracy

Model	Language	Accuracy	F1 Score
Naive Bayes	English	84.44%	91.41%
Naive Bayes	Igbo	84.03%	91.19%
Naive Bayes	Yoruba	84.08%	91.22%
Naive Bayes	Hausa	83.99%	91.18%
Logistic Regression	English	92.82%	95.77%
Logistic Regression	Igbo	89.76%	94.08%
Logistic Regression	Yoruba	91.46%	95.02%
Logistic Regression	Hausa	89.46%	93.93%
BERT	English	91.83%	95.15%
BERT	Igbo	86.90%	92.35%
BERT	Yoruba	87.71%	92.74%
BERT	Hausa	87.47%	92.54%

Cross-Lingual Performance

In this section, we explore how models trained on English data perform when tested on Yoruba, Igbo, and Hausa datasets, assessing the cross-lingual transferability of the Naive Bayes, Logistic Regression, and BERT models.

H. Naive Bayes Cross-Lingual Performance

- **English to Yoruba**: The Naive Bayes model achieved an accuracy of **82.68%** with an F1 score of **89.81%**. The confusion matrix shows that the model struggles with false positives, misclassifying 537 Yoruba instances as negative.
- **English to Igbo**: When tested on Igbo, the Naive Bayes model showed an improved accuracy of **84.34%** and an F1 score of **91.32%**, indicating better performance, particularly in correctly classifying positive instances with fewer false negatives (only 20).
- **English to Hausa**: The model also performed well on Hausa, with an accuracy of **83.94%** and an F1 score of **91.10%**, maintaining consistency in performance. However, a slightly higher number of false negatives (26) were observed compared to Igbo.

I. Logistic Regression Cross-Lingual Performance

- **English to Yoruba**: Logistic Regression showed a lower cross-lingual performance on Yoruba compared to Naive Bayes, with an accuracy of **80.34%** and an F1 score of **87.92%**. The confusion matrix highlights many false negatives, with 557 positive instances misclassified.
- **English to Igbo**: The model fared better on Igbo with an accuracy of **85.41%** and an F1 score of **91.63%**, demonstrating good cross-lingual transferability.
- **English to Hausa**: Logistic Regression also performed well when tested on Hausa data, achieving an accuracy of **85.45%** and an F1 score of **91.81%**, with relatively low false positive and false negative counts.

J. BERT Cross-Lingual Performance

- **English to Yoruba**: BERT, being a more sophisticated model, performed moderately well when tested on Yoruba data, achieving an accuracy of **76.77%** and an F1 score of **85.81%**. Despite its capabilities, the model misclassified 622 positive instances.
- **English to Igbo**: Performance dropped further when BERT was tested on Igbo, with an accuracy of **69.71%** and an F1 score of **79.65%**. The confusion matrix reveals a high number of false negatives (1,165), indicating that BERT struggled with this cross-lingual task.
- **English to Hausa**: BERT's performance was the lowest on Hausa data, with an accuracy of **63.39%** and an F1 score of **74.10%**, highlighting significant challenges with generalizing from English to Hausa. The confusion matrix shows a high number of false negatives (1,508).

K. Language Translation Experiment

At this phase, the Google Cloud translation API adapted specifically for the translation of Text between various languages was adopted in translating the English text into the three main native languages in Nigeria: Igbo, Yoruba, and Hausa in the dataset.

The Google Cloud Translation API is based on Neural Machine Translation (NMT) techniques, allowing seamless translation of text within a dataset. This process involves initializing the API with appropriate credentials, iteratively sending source text for translation while specifying target languages and utilizing a neural machine translation model trained on extensive multilingual data. The model considers contextual understanding to generate accurate and coherent translations, maintaining the original message's intent. Translated output is collected and evaluated for quality, offering efficient batch processing options and customization using glossaries or translation memories for industry-specific terminology. This methodology enables users to effortlessly overcome language barriers, making cross-lingual communication and analysis accessible to developers, businesses, and researchers. It empowers them to integrate translation capabilities seamlessly, enhancing global collaboration, data comprehension, and communication across diverse linguistic contexts.

V. Conclusions

This study evaluated the cross-lingual transferability of sentiment analysis models—Naive Bayes, Logistic Regression, and BERT—across three Nigerian languages: Yoruba, Igbo, and Hausa, using English-trained data. The findings reveal distinct differences in performance, underscoring the strengths and limitations of each model.

Naive Bayes and Logistic Regression showed strong cross-lingual generalization, achieving high accuracy and F1 scores across the tested languages. Logistic Regression emerged as the best-performing model, especially in Igbo and Hausa, reflecting its ability to manage cross-linguistic variations effectively. This suggests that simpler machine learning models can maintain robustness when applied to languages with some semantic and syntactic similarities, even in the absence of large training datasets.

On the other hand, BERT—a model that excels in capturing complex contextual relationships—underperformed in cross-lingual settings, particularly with Yoruba and Hausa. Despite its superior capabilities in monolingual tasks, BERT struggled with generalization across different languages without language-specific fine-tuning, highlighting the challenges in applying pre-trained transformers to low-resource languages. This indicates that BERT's transferability may depend heavily on linguistic overlap between the source and target languages.

In conclusion, while simpler models like Naive Bayes and Logistic Regression demonstrate more reliable performance for cross-lingual sentiment analysis, BERT's potential can only be fully realized with further fine-tuning or adaptation for the target languages. Future research should explore techniques such as multilingual fine-tuning or pre-training to enhance the cross-lingual capabilities of advanced models like BERT, particularly for low-resource languages. This work underscores the need for culturally aware and linguistically adapted language models to improve sentiment analysis across diverse linguistic landscapes.

VI. Future Works

Building on the insights from this study, several avenues for future work emerge that could further enhance the performance of cross-lingual sentiment analysis, particularly for low-resource languages like Yoruba, Igbo, and Hausa:

- **Multilingual Pre-training and Fine-tuning**: To improve the performance of advanced models like BERT in cross-lingual contexts, future work could explore pre-training BERT on multilingual datasets that include African languages or fine-tuning the model specifically on the target languages. This approach could help address BERT's current limitations in handling languages with limited data resources and distinct linguistic structures.
- **Domain-Specific Training Data**: Collecting and curating domain-specific datasets in Yoruba, Igbo, and Hausa could lead to significant performance improvements, particularly for more complex models like BERT. Developing large, high-quality datasets in these languages would reduce the reliance on translation, ensuring more contextually accurate sentiment classification.
- **Cross-Lingual Data Augmentation**: Leveraging data augmentation techniques, such as back-translation or synthetic data generation, could enrich the training data for low-resource languages. By generating diverse language inputs, models may become more robust and adaptable to different linguistic structures and cultural contexts.
- **Cultural and Contextual Embeddings**: Introducing culture-specific embeddings into machine learning models could bridge the cultural gap in sentiment analysis. This would involve embedding sociocultural knowledge and linguistic nuances, allowing models to make more accurate predictions by understanding not just language, but also cultural context.
- **Transfer Learning with Focus on African Languages**: Developing pre-trained models specifically focused on African languages and dialects, or incorporating more African language data into existing models, would greatly improve cross-lingual performance. African languages are currently underrepresented in pre-training datasets, and specialized models could address this gap.
- **Exploration of Ensemble Methods**: Future studies could explore ensemble learning techniques that combine the strengths of multiple models (e.g., Naive Bayes, Logistic Regression, and BERT). Such hybrid approaches could yield more balanced

and accurate predictions by leveraging the unique strengths of each model for cross-lingual sentiment analysis.

- **Bias and Ethical Considerations**: Future research should also focus on addressing any biases that may emerge in cross-lingual sentiment analysis, especially given the sensitive nature of sentiment classification in culturally diverse regions. Ethical considerations, particularly regarding fairness and transparency, should be prioritized in developing these models for practical applications.
- By exploring these directions, future research could significantly improve the performance, robustness, and fairness of cross-lingual sentiment analysis systems, particularly in low-resource languages, ultimately enhancing their real-world applicability in diverse linguistic and cultural contexts.

References

[1] N. Lee, C. Jung, and A. Oh, "Hate Speech Classifiers are Culturally Insensitive," Jan. 2023, doi: https://doi.org/10.18653/v1/2023.c3nlp-1.5.

[2] Y. Chen and X. Li, "Analysis of Linguistic and Cultural Differences in Japanese Translation Based on Differential Equation Modeling in the Perspective of PMC Assessment," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, Nov. 2023, doi: https://doi.org/10.2478/amns.2023.2.01166.

[3] S. Malmasi and M. Zampieri, "Detecting Hate Speech in Social Media," *RANLP 2017 - Recent Advances in Natural Language Processing Meet Deep Learning*, Nov. 2017, doi: https://doi.org/10.26615/978-954-452-049-6_062.

[4] M. O. Ibrohim and I. Budi, "Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter," *ACLWeb*, Aug. 01, 2019. https://aclanthology.org/W19-3506/ (accessed Oct. 21, 2022).

[5] J. Qian, M. ElSherief, E. Belding, and William Yang Wang, "Hierarchical CVAE for Fine-Grained Hate Speech Classification," Jan. 2018, doi: https://doi.org/10.18653/v1/d18-1391.

[6] Y.-H. Lee, S. Yoon, and K. Jung, "Comparative Studies of Detecting Abusive Language on Twitter," *arXiv (Cornell University)*, Jan. 2018, doi: https://doi.org/10.18653/v1/w18-5113.

[7] M. Bilewicz and W. Soral, "Hate Speech Epidemic. the Dynamic Effects of Derogatory Language on Intergroup Relations and Political Radicalization," *Political Psychology*, vol. 41, no. 1, Jun. 2020, doi: https://doi.org/10.1111/pops.12670.

[8] M. ElSherief *et al.*, "Latent Hatred: A Benchmark for Understanding Implicit Hate Speech," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Jan. 2021, doi: https://doi.org/10.18653/v1/2021.emnlp-main.29.

[9] Amir Reza Jafari, G. Li, Praboda Rajapaksha, Reza Farahbakhsh, and N. Crespi, "Fine-Grained Emotions Influence on Implicit Hate Speech Detection," *IEEE access*, vol. 11, pp. 105330–105343, Jan. 2023, doi: https://doi.org/10.1109/access.2023.3318863.

[10] S. D. Swamy, A. Jamatia, and B. Gambäck, "Studying Generalisability across Abusive Language Detection Datasets," *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, doi: https://doi.org/10.18653/v1/k19-1088.

[11] R. Cao, Roy Ka-Wei Lee, and T.-A. Hoang, "DeepHate: Hate Speech Detection via Multi-Faceted Text Representations," *Web Science*, Jul. 2020, doi: https://doi.org/10.1145/3394231.3397890.