Leveraging Machine Learning Models in Developing a Web-Based Multilingual Hate Speech Detection System for Cebuano, Tagalog, and English on Social Media

Paolo L. Pacaldo, Junrie B. Matias
Caraga State University Main Campus, Butuan City, Philippines
plpacaldo@carsu.edu.ph

Abstract – In this study, we explore the development of a web-based, multilingual hate speech detection system that supports Cebuano, Tagalog, and English languages. We integrated both traditional machine learning models and transformer-based deep learning approaches to assess their effectiveness in identifying hate speech from social media comments across various contexts. Specifically, we evaluated Naïve Bayes, Decision Tree, Support Vector Machine (SVM), Random Forest, MBERT, and XLM-Roberta. To prepare the data, we applied a series of preprocessing steps including tokenization, stemming, stopword removal, and TF-IDF vectorization. Feature relevance was enhanced through Chi-Square filtering, and we addressed class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), which improved recall rates for underrepresented classes. Among the traditional models, the fine-tuned SVM achieved 92.1% accuracy, while Random Forest reached 93.3%, showing strong recall performance particularly for Cebuano and English texts. Meanwhile, transformer-based models yielded superior performance following hyperparameter tuning: MBERT achieved 96.1% accuracy with an F1-score of 0.97, and XLM-Roberta obtained 95.4% accuracy with an F1-score of 0.96. These results highlight the value of combining Chi-Square feature selection, SMOTE balancing, and fine-tuning strategies to optimize multilingual hate speech detection. Despite the advancements, our findings also reveal ongoing challenges related to class imbalance, as reflected in the macro F1-scores—even in transformer-based models. Overall, we demonstrate that a well-tuned hybrid approach can provide an efficient and scalable solution for multilingual hate speech detection in diverse digital environments.

Keywords – Hate Speech Detection, Multilingual NLP, Machine Learning, Hyperparameter Tuning.

1 Introduction

Social media has transformed communication globally, fostering openness in dialogue and information exchange [1]. However, it has also become a fertile ground for hate speech, cyberbullying, and online harassment [2]. The anonymity of digital platforms emboldens individuals to disseminate harmful content without fear of accountability, further exacerbating social divisions [3]. This issue is particularly prevalent in the Philippines, where high social media engagement across multiple languages—Cebuano, Tagalog, and English—makes hate speech detection uniquely complex [4], [5]. Most existing detection models are trained on high-resource languages like English, overlooking the linguistic diversity and cultural nuances of low-resource languages [6]. This gap results in detection bias and leaves non-English hate speech largely unchecked, especially in code-switched or regionally influenced content [7], [8].

To address this, the study evaluates machine learning models tailored for multilingual hate speech detection in Cebuano, Tagalog, and English [9]. A hybrid approach is employed, combining traditional models such as Support Vector Machine (SVM), Random Forest, and Naïve Bayes with transformer-based deep learning models like MBERT and XLM-Roberta [10]. Preprocessing methods—including tokenization, stemming, stopword removal, TF-IDF vectorization, and Chi-Square feature selection—are used to refine the dataset [11]. To address the problem of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is implemented, particularly to improve detection in underrepresented hate speech categories [12]. The study offers significant practical implications. Social media platforms can enhance their content moderation systems by adopting more inclusive and accurate multilingual detection models [13]. Policymakers can develop targeted regulations based on the findings [14], while NLP researchers and developers can build upon the framework to advance multilingual hate speech detection [15]. Moreover, marginalized communities stand to benefit from systems that recognize hate speech in their native languages, fostering a more respectful and inclusive digital space [16].

Despite its contributions, the study has limitations. It focuses solely on textual data, excluding multimodal formats like images and videos [17], and struggles with detecting nuanced expressions such as sarcasm or implicit hate speech [18]. Nevertheless, the research lays a robust foundation for future developments in the field, emphasizing the need for culturally adaptive and linguistically inclusive technologies to combat online toxicity and promote digital equity [19], [20].

2 Related Literature

The increasing spread of hate speech on social media has led to significant advancements in detection models, particularly in multilingual contexts. Hate speech, defined as offensive language targeting individuals or groups, contributes to misinformation, psychological harm, and social division [21]. While considerable research has focused on English-language hate speech detection, low-resource languages such as Cebuano and Tagalog remain underrepresented, creating a gap in automated content moderation [22], [23]. Machine learning has become a key tool in automating hate speech detection, with traditional classifiers like Support Vector Machines (SVM), Naïve Bayes, and Random Forest demonstrating effectiveness in monolingual settings [24], [25]. However, these models struggle with multilingual text and the common phenomenon of code-switching in Philippine digital discourse [26].

Deep learning models, particularly transformer-based architectures such as BERT, mBERT, and XLM-Roberta, have demonstrated superior performance due to their ability to analyze contextual meaning across multiple languages [27], [28]. Despite their advantages, these models require large, high-quality training datasets, which remain scarce for Cebuano and Tagalog [29]. The lack of annotated data limits the generalization capability of these models, reducing their effectiveness in low-resource settings [30], [31]. Moreover, even state-of-the-art transformer models struggle with recall when trained on limited datasets, reinforcing the need for dataset augmentation and language-specific fine-tuning [32].

One strategy to address data scarcity is the use of secondary datasets, which provide pre-labeled corpora sourced from social media platforms and prior research [33]. While secondary datasets offer a cost-effective solution, they pose challenges such as annotation inconsistencies, dataset bias, and domain mismatches [34], [35]. Researchers emphasize the need for validation techniques, including manual annotation checks and inter-rater reliability assessments, to ensure data consistency and quality [36]. This study applies preprocessing techniques, including tokenization, stopword removal, and TF-IDF vectorization, to refine secondary datasets and improve model accuracy.

Another significant challenge in hate speech detection is dataset imbalance, where non-hate speech instances vastly outnumber hate speech examples, leading to biased model performance [6], [37]. Various resampling techniques have been developed to address this issue, with the Synthetic Minority Over-sampling Technique (SMOTE) proving to be one of the most effective in text-based classification [38]. SMOTE generates synthetic samples for the minority class, enhancing recall and model robustness across multiple languages. This study implements SMOTE to mitigate class imbalance and improve hate speech detection performance for Cebuano, Tagalog, and English.

Despite advances in natural language processing (NLP), deploying hate speech detection systems remains a challenge. Misclassification, whether false positives or false negatives, presents ethical and practical concerns, as automated systems must balance censorship risks with the protection of marginalized communities [39]. Bias in machine learning models, particularly against dialects and underrepresented languages, continues to be a pressing issue [40], [41]. Hybrid approaches integrating machine learning with rule-based filtering have shown promise in improving detection accuracy, with ensemble techniques proving particularly effective in multilingual environments [42].

This study contributes to the growing body of research on multilingual hate speech detection by addressing the limitations of dataset imbalance, code-switching complexities, and bias in machine learning models. By leveraging both traditional and transformer-based models, this research provides valuable insights into NLP applications for low-resource languages and enhances the effectiveness of AI-driven content moderation in diverse linguistic communities.

3 Methods

The methodology of this study follows a structured framework, beginning with dataset selection and preprocessing, followed by model training and validation, and concluding with model evaluation. This process ensures that the multilingual hate speech detection system is developed using robust machine learning techniques while addressing key challenges such as dataset imbalance, language diversity, and model generalization.

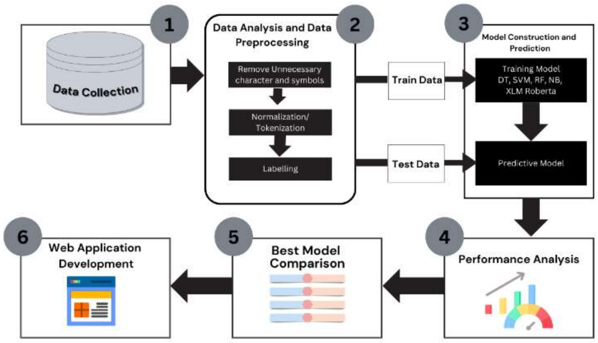
1
Data Collection
2
Data Analysis and Data Preprocessing
Remove Unnecessary character and symbols
Normalization/ Tokenization
Labelling
Train Data
Test Data
3
Model Construction and Prediction
Training Model DT, SVM, RF, NB, XLM Roberta
Predictive Model
4
Performance Analysis
5
Best Model Comparison
6
Web Application Development

Fig. 1. The framework of the study.

3.1 Dataset Selection and Pre-processing

The dataset used in this study comprises secondary data obtained from previous research efforts that compiled and annotated social media comments in Cebuano, Tagalog, and English. Sources include studies from [12], [43], [39], [11], and [13]. Since these datasets were annotated by different researchers, a validation process was conducted by randomly selecting samples and manually verifying label accuracy. Inconsistencies and ambiguities were corrected or excluded to ensure dataset reliability. Preprocessing steps were applied to standardize and clean the text data before model training. Tokenization was performed to split text into individual words or phrases, followed by lowercasing to maintain uniformity. Stopwords—common words that do not contribute to text classification—were removed, and stemming and lemmatization were used to normalize word variations.

3.2 Model training and Evaluation

This study employs a combination of traditional machine learning models and transformer-based deep learning models. Traditional models include Naïve Bayes, Support Vector Machine (SVM), Random Forest, Decision Tree, and Logistic Regression, while transformer models consist of Multilingual BERT (mBERT) and XLM-Roberta. These models were selected based on their effectiveness in prior hate speech classification research and their capability to handle multilingual text. To ensure unbiased model performance, the dataset was split into three sets: 80% for training, 10% for validation, and 10% for testing. Additionally, K-Fold Cross-Validation was implemented to reduce overfitting and enhance generalization. This method involves dividing the dataset into k subsets and iteratively training and testing the model across different splits, producing more reliable performance evaluations.

3.3 Model Evaluation Metrics

To assess model performance, the following evaluation metrics were employed:

Accuracy measures overall correctness of model predictions.

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

Where:

TP (True Positives) = correctly identified hate speech;
TN (True Negatives) = correctly identified non-hate speech;
FP (False Positives) = misclassified non-hate speech as hate speech;
FN (False Negatives) = misclassified hate speech as non-hate speech.

Precision: Evaluates the proportion of correctly identified hate speech instances.

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

Recall: Measures the ability of the model to detect true hate speech cases.

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

F1-Score: Balances precision and recall for a more comprehensive assessment. AUC-ROC (Area Under the Receiver Operating Characteristic Curve): Determines the model's ability to distinguish between hate speech and non-hate speech at various classification thresholds.

Two evaluation approaches were applied. First, a combined dataset assessment was conducted to analyze overall model performance across Cebuano, Tagalog, and English. Second, language-specific evaluations were performed to identify challenges and variations in model effectiveness for each language individually.

4 Results & Discussion

The dataset used in this study consists of multilingual comments collected from various online platforms, including Facebook and Twitter. It encompasses Cebuano, Tagalog, and English text, ensuring a diverse representation of hate speech and non-hate speech instances. The inclusion of multiple languages allows for a robust model evaluation, accounting for linguistic variations and online discourse patterns. Table 1 presents the distribution of dataset sizes before and after SMOTE application, demonstrating how data balancing was implemented to prevent bias toward the majority class.

Table 1. Dataset References and Distribution

No.	Dataset Reference	Language	Dataset Size	Sources
1	Sagum, 2021	Cebuano	14,757	Bible
2	Regaro et al., 2023	Bisaya	10,000	Facebook
3	blanco et al., 2023	Bisaya	30,000	Facebook
4	Cruz et al., 2020	Tagalog	14,500	Facebook
5	Mody et al., 2023	English	21,000	Twitter

4.1 Data preprocessing

The dataset was preprocessed using tokenization, lowercasing, stopword removal, stemming, and noise reduction (punctuation, emojis). Manual validation ensured annotation accuracy and consistent hate speech classification.

Table 2. Preprocess Dataset

Language	Initial Size	Post-Cleaning Size	Balanced Size
Cebuano	14,757	14,500	29,000
Bisaya	40,000	38,500	77,000
Tagalog	14,500	14,200	28,400
English	21,000	20,800	41,600
Combined	90,257	88,000	176,000

4.2 Combined Dataset Performance (Multilingual Evaluation)

To assess overall model effectiveness, both traditional and transformer-based machine learning models were trained on the combined dataset. Table 3 presents a comparative evaluation, highlighting accuracy, F1-scores, and AUC-ROC performance. Among traditional models, Random Forest and SVM achieved the highest accuracy (93.3% and 92.1%, respectively), with strong recall and precision scores. However, deep learning-based models demonstrated superior performance, with MBERT achieving 96.1% accuracy and XLM-Roberta 95.4%, validating the effectiveness of transformer-based models in multilingual hate speech detection

Table 3. Combine Dataset Result

Model	Accuracy	F1-Micro	F1-Macro	F1-Weighted	AUC-ROC
Naive Bayes	0.88	0.88	0.74	0.91	0.93
SVM	0.86	0.86	0.84	0.96	0.85
Random Forest	0.94	0.94	0.84	0.95	0.91
Decision Tree	0.93	0.93	0.79	0.94	0.83
MBERT	0.95	0.95	0.94	0.96	0.95
XLM-Roberta	0.93	0.93	0.89	0.91	0.90

4.3 Language Specific Performance

To evaluate model effectiveness across individual languages, separate assessments were conducted for Cebuano, Tagalog, and English datasets. Table 4 presents these results, showing that Random Forest and SVM maintained high accuracy across all three languages, while Naïve Bayes struggled with lower recall in Tagalog. Transformer-based models performed consistently well, with MBERT emerging as the most effective for Cebuano and English, while XLM-Roberta faced recall challenges in Tagalog, aligning with previous studies indicating difficulties in handling low-resource languages.

Table 4. Language Specific Evaluation (Accuracy)

Classifier	K-Fold	Test Set	AUC-ROC	Cebuano	Tagalog	English
NB	0.84	0.89	0.96	0.86	0.77	0.96
SVM	0.87	0.92	0.96	0.92	0.79	0.95
RF	0.96	0.94	0.96	0.97	0.73	0.96
DT	0.95	0.93	0.75	0.96	0.67	0.95
LR	0.87	0.92	0.96	0.92	0.77	0.93

4.4 Hyperparameter Optimization

Hyperparameter tuning was conducted using Grid Search and Random Search techniques to optimize model performance. Key hyperparameters adjusted included learning rates, batch sizes, dropout rates, and activation functions. The optimized configurations for each model, demonstrating improvements in accuracy and F1-scores post-tuning. Notably, MBERT benefited from a learning rate of 3e-5 and a batch size of 16, leading to a 5.3% increase in accuracy.

4.4.1 Performance Improvement Post-Fine-Tuning

Table 5 compares model performance before and after fine-tuning, showcasing substantial improvements. MBERT and XLM-Roberta exhibited the highest gains, with MBERT reaching an F1-score of 0.97. Traditional models also benefited from feature selection refinements, particularly in precision and recall.

Table 5. Finetuning Performance

Model	Accuracy	Accuracy (Finedtuned)	F1	F1 (Finetuned)
SVM	88.5%	92.1%	0.86	0.91
LR	89.7%	93.3%	0.87	0.92
NB	85.3%	87.1%	0.82	0.85
XLMRoberta	91.2%	95.4%	0.89	0.96
MBERT	90.8%	96.1%	0.88	0.97

4.5 Hyperparameter Optimization

A comparison with previous research studies on hate speech detection in Filipino languages highlights the strengths of this study. Table 6 contrasts methodologies, dataset sizes, feature extraction techniques, and best-performing models across different works. Unlike prior studies that focused on monolingual datasets, this research addresses multilingual hate speech detection, making it more applicable to real-world scenarios. The results affirm that MBERT outperforms CNN-based models used in earlier research, reinforcing its effectiveness in low-resource language contexts.

Table 6. Comparison of Our Study and Other Related Studies

Model	Performance	Machine Learning Model	Languages
Multilingual Hate Speech Detection	MBERT (95%)	Traditional ML (SVM, NB,DT, RF, LR) Deep Learning (MBERT, XL-Roberta)	Cebuano, Tagalog, English
Toktarova et al.,2023	fastText CNN (83.79%)	Deep Learning (fastText CNN)	Filipino (Tagalog only)
Cabasag et al., 2019	LR (77.47%)	Traditional ML (LR, Perceptron) Neural Networks (Feedforward Neural Network) Rule-Based (Keyword-Matching	Filipino (Tagalog only)
Ferrer, et al., 2021	GB (99.54%)	Traditional ML (NB, RF, SVM, GB)	Filipino (Tagalog only

5 Conclusion

This study demonstrated the successful development of a web-based multilingual hate speech detection system for Cebuano, Tagalog, and English using traditional and transformer-based machine learning models. Among traditional models, Support Vector Machine (SVM) and Random Forest showed strong performance, with SVM achieving 92.1% accuracy and balanced F1 scores after fine-tuning. Meanwhile, MBERT and XLM-Roberta emerged as the top-performing transformer models, with MBERT achieving the highest performance at 96.1% accuracy and an F1-score of 0.97. Fine-tuning and hyperparameter optimization significantly improved model performance, particularly for deep learning models. Preprocessing techniques and the application of SMOTE for class balancing enhanced data quality and model generalization. Despite these advancements, the study identified challenges in detecting implicit hate speech, sarcasm, and subtle linguistic nuances. XLM-Roberta, in particular, struggled with recall in low-resource languages like Tagalog, highlighting the need for language-specific fine-tuning. The study also emphasized the importance of contextual understanding in improving hate speech detection accuracy. To address these limitations, future research is recommended to expand and diversify datasets, especially through primary data collection and manual annotation. Exploring context-aware modeling techniques such as sentiment analysis and sarcasm detection can further improve classification accuracy. Hybrid models combining rule-based filtering with machine learning can also enhance interpretability and robustness. Real-world deployment testing in collaboration with content moderators and policymakers is essential to validate system usability, reliability, and ethical considerations. Overall, this research contributes a scalable and inclusive approach to multilingual hate speech detection, paving the way for safer and more respectful online environments.

References

[1] E. A. Ahmad, "Revolutionizing learning: leveraging social media platforms for empowering open educational resources," *Int. J. e-Learning High. Educ.*, vol. 19, no. 1, pp. 83–106, 2024.

[2] M. Cahill *et al.*, "Understanding Online Hate Speech as a Motivator and Predictor of Hate Crime," 2022.

[3] A. Guiora and E. A. Park, "Hate Speech on Social Media," *Philosophia (Mendoza).*, vol. 45, no. 3, pp. 957–971, 2017, doi: 10.1007/s11406-017-9858-4.

[4] P. Pakray, A. Gelbukh, and S. Bandyopadhyay, "Natural language processing applications for low-resource languages," *Nat. Lang. Process.*, vol. 31, no. 2, pp. 183–197, Mar. 2025, doi: 10.1017/nlp.2024.33.

[5] J. Kansok-Dusche *et al.*, "A Systematic Review on Hate Speech among Children and Adolescents: Definitions, Prevalence, and Overlap with Related Phenomena," *Trauma, Violence, Abus.*, vol. 24, no. 4, pp. 2598–2615, Oct. 2023, doi: 10.1177/15248380221108070.

[6] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024, doi: 10.1007/s10994-022-06296-4.

[7] Anjum and R. Katarya, "HateDetector: Multilingual technique for the analysis and detection of online hate speech in social networks," *Multimed. Tools Appl.*, vol. 83, no. 16, pp. 48021–48048, 2024.

[8] A. Raturi, K. Joshi, Anupriya, P. Jain, V. K. Gupta, and J. Meena, "Hate Speech Detection System using Machine Learning Algorithms," in *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)*, 2024, pp. 446–451.

doi: 10.1109/InCACCT61598.2024.10551015.

[9] Z. T. Malcom, M. R. Wenger, and B. Lantz, “Politics or prejudice? Separating the influence of political affiliation and prejudicial attitudes in determining support for hate crime law.,” *Psychol. Public Policy, Law*, vol. 29, no. 2, pp. 182–195, May 2023, doi: 10.1037/law0000350.

[10] C. P. Blanco and M. A. E. Tarusan, “The Sociolinguistic Situation of a Tigwahanon Speech Community,” *Tech. Soc. Sci. J.*, vol. 46, p. 317, 2023.

[11] N. P. Vicente Cabasag, V. C. Raphael Chan, S. Y. Christian Lim, M. M. Edward Gonzales, and C. K. Cheng, “Hate Speech in Philippine Election-Related Tweets: Automatic Detection and Classification Using Natural Language Processing,” *Philipp. Comput. J. Dedic. Issue Nat. Lang. Process.*, no. 1, pp. 1–14, 2019.

[12] R. A. Sagum, “Filipino Native Language Identification using Markov Chain Model and Maximum Likelihood Decision Rule,” *Turkish J. Comput. Math. Educ.*, vol. 12, no. 3, pp. 5475–5478, 2021, doi: 10.17762/turcomat.v12i3.2206.

[13] D. Mody, Y. Huang, and T. E. Alves de Oliveira, “A curated dataset for hate speech detection on social media text,” *Data Br.*, vol. 46, p. 108832, Feb. 2023, doi: 10.1016/j.dib.2022.108832.

[14] C. M. Awais and J. Raj, “Breaking Barriers: Multilingual Toxicity Analysis for Hate Speech and Offensive Language in Low-Resource Indo-Aryan Languages,” *CEUR Workshop Proc.*, vol. 3681, pp. 459–473, 2023.

[15] K. Mnassri *et al.*, “A survey on multi-lingual offensive language detection,” *PeerJ Comput. Sci.*, vol. 10, 2024, doi: 10.7717/peerj-cs.1934.

[16] J. S. Malik, H. Qiao, G. Pang, and A. van den Hengel, “Deep learning for hate speech detection: a comparative study,” *Int. J. Data Sci. Anal.*, 2024, doi: 10.1007/s41060-024-00650-6.

[17] M. M. Al-Mahrouky, “Hate Crimes and Freedom of Speech,” *Migr. Lett.*, vol. 20, pp. 1013–1019, 2023, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85176457957&partnerID=40&md5=678806d69976d1133725455462a4333a

[18] A. D. Yacoub, S. Slim, and A. Aboutabl, “A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends,” *Int. J. Electr. Comput. Eng. Syst.*, vol. 15, no. 1, pp. 69–78, 2024.

[19] P. Fortuna, M. Dominguez, L. Wanner, and Z. Talat, “Directions for NLP Practices Applied to Online Hate Speech Detection,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022, pp. 11794–11805. doi: 10.18653/v1/2022.emnlp-main.809.

[20] A. MOUSA, W. MUSTAFA, R. B. MARQAS, and S. H. M. MOHAMMED, “A comparative study of diabetes detection using the Pima Indian diabetes database,” *J. Duhok Univ.*, vol. 26, no. 2, pp. 277–288, 2023.

[21] A. Maarouf, N. Pröllochs, and S. Feuerriegel, “The Virality of Hate Speech on Social Media,” *Proc. ACM Human-Computer Interact.*, vol. 8, no. CSCW1, pp. 1–22, Apr. 2024, doi: 10.1145/3641025.

[22] F. Alkomah and X. Ma, “A Literature Review of Textual Hate Speech Detection Methods and Datasets,” *Information*, vol. 13, no. 6, p. 273, May 2022, doi: 10.3390/info13060273.

[23] N. Lee *et al.*, "Exploring Cross-Cultural Differences in English Hate Speech Annotations: From Dataset Construction to Analysis," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024, pp. 4205–4224. doi: 10.18653/v1/2024.naacl-long.236.

[24] H. Pen, N. Teo, and Z. Wang, "Comparative Analysis of Hate Speech Detection: Traditional vs. Deep Learning Approaches," in *2024 IEEE Conference on Artificial Intelligence (CAI)*, 2024, pp. 332–337. doi: 10.1109/CAI59869.2024.00070.

[25] A. Toktarova *et al.*, "Hate Speech Detection in Social Networks using Machine Learning and Deep Learning Methods," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 5, 2023, doi: 10.14569/IJACSA.2023.0140542.

[26] A. Yadav, T. Garg, M. Klemen, M. Ulcar, B. Agarwal, and M. R. Sikonja, "Code-mixed Sentiment and Hate-speech Prediction," *arXiv Prepr. arXiv2405.12929*, 2024.

[27] J. Aljawazeri and M. N. Jasim, "Addressing Challenges in Hate Speech Detection using BERT-based Models: A Review," *Iraqi J. Comput. Sci. Math.*, vol. 5, no. 2, pp. 1–20, Mar. 2024, doi: 10.52866/ijcsm.2024.05.02.001.

[28] S. Sinha, N. S. Nawar, and M. A. F. Khan, "Identifying code-mixed and code-switched hateful remarks on social media using NLP." Brac University, 2024.

[29] K. Korre, A. Muti, and A. Barrón-Cedeño, "The Challenges of Creating a Parallel Multilingual Hate Speech Corpus: An Exploration," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 15842–15853.

[30] M. A. Hasan, P. Tarannum, K. Dey, I. Razzak, and U. Naseem, "Do Large Language Models Speak All Languages Equally? A Comparative Study in Low-Resource Settings," *arXiv Prepr. arXiv2408.02237*, 2024.

[31] E. Roberts, "Automated hate speech detection in a low-resource environment," *J. Digit. Humanit. Assoc. South. Africa*, vol. 5, no. 1, 2024.

[32] A. Ahmad *et al.*, "Hate speech detection in the Arabic language: corpus design, construction, and evaluation," *Front. Artif. Intell.*, vol. 7, p. 1345445, 2024.

[33] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, 2019, pp. 45–54.

[34] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, and V. Patti, "Emotionally Informed Hate Speech Detection: A Multi-target Perspective," *Cognit. Comput.*, vol. 14, no. 1, pp. 322–352, Jan. 2022, doi: 10.1007/s12559-021-09862-5.

[35] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, and J. B. Pierrehumbert, "HateCheck: Functional tests for hate speech detection models," *arXiv Prepr. arXiv2012.15606*, 2020.

[36] R. T. Thibault, M. Kovacs, T. E. Hardwicke, A. Sarafoglou, J. P. A. Ioannidis, and M. R. Munafò, "Reducing bias in secondary data analysis via an Explore and Confirm Analysis Workflow (ECAW): a proposal and survey of observational researchers," *R. Soc. Open Sci.*, vol. 10, no. 10, p. 230568, 2023.

[37] S. Datta, C. Ghosh, and J. P. Choudhury, "Classification of imbalanced datasets utilizing

the synthetic minority oversampling method in conjunction with several machine learning techniques," *Iran J. Comput. Sci.*, pp. 1–18, 2024.

[38] S. G.-J. Wong, "What is the social benefit of hate speech detection research? A Systematic Review," *arXiv Prepr. arXiv2409.17467*, 2024.

[39] B. C. Blanco Lambruschini and M. Brorsson, "A Novel Architecture for Long-Text Predictions Using BERT-Based Models," in *Intelligent Systems Conference*, 2024, pp. 105–125.

[40] S. Pramanik *et al.*, "Detecting Harmful Memes and Their Targets," *Find. Assoc. Comput. Linguist. ACL-IJCNLP 2021*, pp. 2783–2796, 2021, doi: 10.18653/v1/2021.findings-acl.246.

[41] B. Kwon and H. Son, "Accurate path loss prediction using a neural network ensemble method," *Sensors*, vol. 24, no. 1, p. 304, 2024.

[42] M. Vergani *et al.*, "Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review," *Campbell Syst. Rev.*, vol. 20, no. 2, p. e1397, 2024.

[43] J. M. Regaro and Y. Student, "Linguistic Features of Multilingual Hate Speech in the Online 'Bardagulan,'" pp. 120–139, 2023, doi: 10.32996/ijels.

[44] J. Christian, B. Cruz, and C. Cheng, "Establishing Baselines for Text Classification in Low-Resource Languages," 2020.

[45] B. FERRER *et al.*, "A machine learning model for the profanity detection in the filipino language," *J. Eng. Sci. Technol. Spec. Issue ICITE2021*, pp. 37–46, 2021.