VoCL: Text-Based Multilingual Hate Speech Profiling using NLP Speaker Embeddings

Iber Joseph P. Bonilla
College of Computing and Information Technologies
National University
Sampaloc, Manila
bonillaip@students.national-u.edu.ph

Andrew Laurence T. Fat
College of Computing and Information Technologies
National University
Sampaloc, Manila
fatat@students.national-u.edu.ph

***Abstract*—As our world grows increasingly global and digital, the detection of hate speech needs to be increasingly contextualized across languages, cultures, and settings. Thus, we contribute to the field of study with VoCL (Voice of Community Language), a form of text-based multilingual hate speech profiling based on NLP-derived speaker embeddings for community-level linguistic patterns and preferences. Where most approaches to hate speech detection default to a classification problem, VoCL addresses it more as a profiling problem involving semantic, syntactic and contextual speaker characteristics in multiple languages, including but not limited to Arabic, Chinese and Indonesian. Using powerful, pre-determined transformer-based architectures and embedding techniques, VoCL achieves increased generalization for cross-lingual detection for its multilingual aspects but also fosters equity and fairness through user-level contextualization. Ultimately, a series of experiments upon a multimodal hate speech corpus in multiple languages demonstrate the efficacy of our approaches in terms of accuracy and cross-lingual results, fostering a new direction for auto-moderation systems to address multicultural realms in scalable, fair fashions.**

***Index Terms*—HateSpeech, NLP, CNN, Classification**

I. Introduction

With online content more accessible than ever through global networks and web-based digital libraries, the ability to fight back against hate speech in a cross-lingual, cross-cultural arena is more prominent than ever. Although many advances have been made for in-language hate speech detection systems much does not transfer, generalize or scale from translation, dialects or cross-cultural applications. This poses an issue for content moderation infrastructure which must engage with multilingual sites and cross-culturally appreciate how hateful sentiment manifests differently in structure, semantics and socio-linguistic domains.

Yet historically, the creation of hate speech detection systems have emerged from the basis of keyword-based prevention filters or language-dependent classifiers that fail to render the perspective of speaker context and linguistic variation. This is to say that systems are bound to detect hate speech due to task classification and assessment without the inclusion of speaker-dependent features that might offer useful secondary context. This not only fails to leverage the capabilities of low-resource languages and cross-linguals, but also fails to render a detection system that works successfully on a global scale in online forums.

In this study, we proposed VoCL (Voice of Community Language), a text-based multilingual hate speech profiling framework that seeks to utilize NLP-based speaker embeddings for context relevance and user-level profiling as we do not want to simply detect but instead, seek to profile speakers through acknowledgment of their unique linguistic habits, writing styles and contextual behaviors so that analysis of hate speech detection can occur with the proper multilingual assessments and speaker relevance.

VoCL detects by creating hate speech detection from speaker embeddings relying upon pre-trained multilingual transformer models from semantic and syntactic representations existing in multilingual means, and further blurring with speaker embedding approaches that encode representative context for speaker-specific information over time.

II. Related Work

Hate speech detection has advanced rapidly, particularly with the growth of multilingual social media content. Traditional methods relying on monolingual feature engineering and keyword lists have been largely replaced by deep learning models capable of extracting richer semantic representations. In this chapter, we review recent studies (2020 onward) on multilingual and cross-lingual hate speech detection, focusing on key modeling approaches and contributions.

A. Cross-Lingual Transfer Learning

Cross-lingual transfer learning has become one of the primary methods to address the challenges of little labeled data for many languages. For example, according to Ranasinghe and Zampieri [1], cross-lingual transfer by fine-tuning a pre-existing multilingual transformer model resulted in success with classifying hate speech in the low-resourced languages of Hindi and Bengali with substantial F1-score improvements when compared to existing baseline classifiers.

Another common method is the use of zero-shot models. For example, according to Pelicon et al. [2], a multilingual transformer developed on English hate speech was used on other European languages with no additional training. The findings for news articles and tweets for cross-lingual generalization found strong results, suggesting some multilingual

shared representations are generalizable; however, results were not consistent across all applied languages.

Yet Nozza [3] challenged the idea that zero-shot models perform the same across cultural boundaries. For example, replicating the same research into other cultures does not yield similar results. Her findings suggest that sociolinguistic contextualization is key as when trained models in English are used on the Italian and Spanish-created corpora, errors arise due to semantics mismatching.

On the other hand, Eronen et al. [4] tried to determine if languages closer together fared better in transfer results. Their findings show that transfer is best with typological similarities, at the language family level, meaning trained zero-shot detections work better with related high-resource language trains and their applied components.

Finally, Monnar et al. [5] created multilingual task-specific embeddings trained exclusively on hate speech across languages, which outperformed more generic multilingual models such as mBERT in zero-shot testing, suggesting that embeddings work best when made for their intended purpose.

B. Joint Multilingual and Ensemble Models

Some investigations work on creating universal architectures from multilingual data. Vashistha and Zubiaga [6] trained a joint CNN-BiLSTM model over a merged corpus from Hindi and English made of six datasets and achieved the highest accuracy over all languages, showing that one architecture could work for multilingual hate speech detection.

Mahajan et al. [7] created an ensemble called EnsMulHateCyb where CNNs, LSTMs, and BiGRUs operate as base learners with a meta-learner. The ensemble was validated over nine datasets across languages; the ensemble operation achieved state-of-the-art performance in every instance, signifying that the ensemble operation combines the best of various models' parts for more efficient classification across multiple languages.

Lastly, Hashmi et al. [8] used a multilingual approach with the ability to validate over 13 languages using transformer architectures like XLM-RoBERTa. Their F1-scores exceeded previously claimed bests by more than 10%, signaling that transformer-based architectures scale well for multilingual hate speech detection.

III. Methodology

The multilingual pipeline for hate speech detection utilizes a convolutional neural network (CNN) with attention layers. This solution aims to achieve cross-lingual generalization with a single architecture and a standard preprocessing approach.

A. Data Collection and Preprocessing

The datasets utilized in this study were sourced from two primary repositories. The first dataset is the *Hate Speech and Offensive Language Dataset* available on Kaggle, compiled by Andrii Samoshyn. This dataset contains labeled social media posts categorized as hateful, offensive, or neither, providing a diverse collection of text for training and evaluation purposes. The second dataset is the *Filipino Hate Speech Text Dataset from Twitter*, compiled by Rommel Urbano Jr., Jeffrey Uy, Angelic Angeles, Maria Nikki Quintos, Joseph Marvin Imperial, and Ramon Rodriguez. This dataset contains hate speech in text form transcribed from Filipino Twitter, primarily related to political discourse. The data from these sources were preprocessed and integrated to form a comprehensive multilingual corpus for the hate speech detection model.

The multilingual social media posts comprise the test set. The languages used are English, German, Spanish, French, Italian, Korean, Indonesian, Chinese, and Filipino. Each post is either labeled as hateful (1) or non-hateful (0).

To ensure data quality, the core classifier is a multi-channel CNN with an attention mechanism. The architecture includes an embedding layer that maps word indices to dense 300-dimensional vectors. These embeddings can be trained from scratch or initialized with random values, or vertical multilingual embeddings can be extended, such as FastText or mBERT, as referenced in [1], [5]. The convolution layers consist of three 1D convolutional layers, each with kernels of size 3, 4, and 5, respectively, which learn to identify n-grams. Each convolutional layer has 200 channels and is followed by batch normalization and ReLU activation, along with adaptive max pooling. A linear attention layer is applied to the concatenated output from the convolutional layers, which has 600 dimensions, allowing the model to prioritize semantically relevant features. Dropout regularization with a dropout rate of 0.3 combats overfitting, and the final dense layer produces one logit for binary classification. This preprocessing pipeline is language-agnostic and similar to multilingual approaches in [2], [5], and [8].

B. Model Architecture

The primary classifier is a multi-channel CNN with attention. The architecture includes an embedding layer that maps word indices to 300 dense embeddings. These embeddings are either randomly initialized or, for improved performance, expanded via pretrained universal embedding options like FastText or mBERT, as referenced in [1], [5]. The convolution layers consist of three 1D convolutional layers with filters of kernel sizes 3, 4, and 5 for n-gram detection. Each convolutional layer results in 200 channels followed by batch normalization, ReLU activation, and adaptive max pooling. A linear attention layer is applied to the concatenated features from the convolutional layers, which have 600 dimensions, to choose semantically relevant features. A dropout layer at 0.3 is used to avoid overfitting, and the final dense layer is fully connected with one output logit for binary classification.

C. Training Configuration

The model is trained using Binary Cross-Entropy with Logits BCE with Logits Loss with label smoothing set to 0.1 to avoid overconfidence and encourage better generalization, as detailed in [6], [7]. The optimizer used is Adam with weight decay regularization. A cyclic learning rate scheduler

is employed for faster convergence and to prevent the model from falling into local minima. Gradient clipping is applied with a max_norm of 1.0 to prevent exploding gradients. Early stopping is enabled so that if the validation loss does not improve over seven epochs, the training stops to prevent overfitting. A batch size of 256 is used for GPU operating efficiency, and the model is trained for a maximum of 20 epochs, with early stopping occurring beforehand.

D. Evaluation Metrics

The model is evaluated using several metrics including accuracy, precision, recall, and F1-Score. Additionally, per-language accuracy is also measured. These metrics are standard for hate speech detection and allow for comparison with existing models as discussed in [1], [3], [6].

E. Inference and Testing

After training, the model is tested on unseen CSV-formatted test files, each representing a different language. Tokenization and encoding follow the same approach as during training. Predictions are made in batches, and the sigmoid outputs are compared to the ground truth labels to compute precision. Languages such as Korean and German performed poorly, likely due to tokenization mismatches or vocabulary sparsity, as noted in [3], [4], and [8].

F. Data Sources

The datasets utilized in this study were sourced from two primary repositories. The first dataset is the *Hate Speech and Offensive Language Dataset* available on Kaggle, compiled by Andrii Samoshyn. This dataset contains labeled social media posts categorized as hateful, offensive, or neither, providing a diverse collection of text for training and evaluation purposes. The second dataset is the *Filipino Hate Speech Text Dataset from Twitter*, compiled by Rommel Urbano Jr., Jeffrey Uy, Angelic Angeles, Maria Nikki Quintos, Joseph Marvin Imperial, and Ramon Rodriguez. This dataset contains hate speech in text form transcribed from Filipino Twitter, primarily related to political discourse. The data from these sources were preprocessed and integrated to form a comprehensive multilingual corpus for the hate speech detection model.

IV. Results

A. Training Performance

The model was trained for 20 epochs, with loss and accuracy monitored at each epoch. The following graphs display the key performance metrics. Loss over Epochs

Training Loss decreases steadily, reflecting the model's learning progress.Validation Loss decreases and stabilizes after a few epochs, indicating successful generalization.

1) Accuracy over Epochs: [Graph Placeholder: Accuracy over Epochs] ((Plot showing train accuracy and validation accuracy over the 20 epochs.)) Train Accuracy reaches 94.24% by epoch 20.Validation Accuracy stabilizes after epoch 6 at 90.37%, showing good generalization and no overfitting.

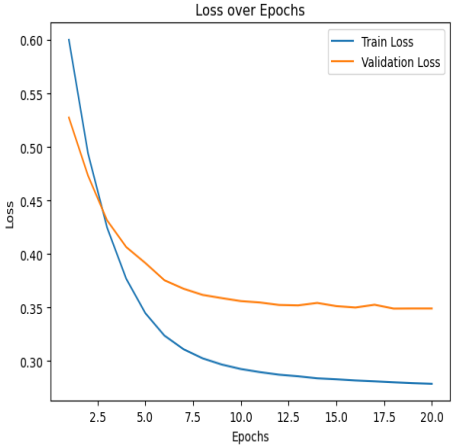

Fig. 1. Loss vs Epochs

2) F1-Score over Epochs: [Graph Placeholder: F1-Score over Epochs] ¡Plot showing train F1-score and validation F1-score over the 20 epochs.¿

F1-Score improves steadily, with the final validation F1 reaching 0.87, indicating a balanced model performance on both classes.

B. Test Results

After training, the model was tested on several multilingual datasets. The following graph shows the test accuracy for each language.

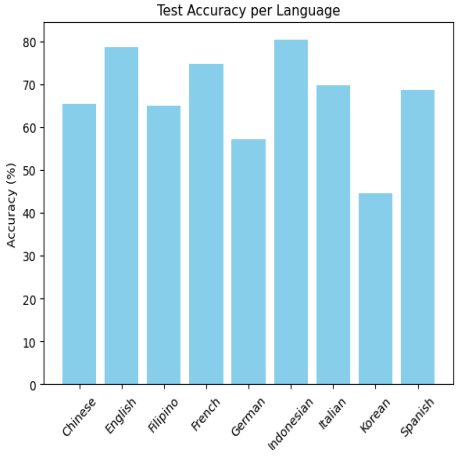

Fig. 2. Test Accuracy per Language

Indonesian achieved the highest accuracy at 80.41%, likely due to its rich online presence. Korean had the lowest accuracy at 44.43%, likely due to tokenization challenges and language-specific issues.

C. Accuracy Over Epochs

This graph shows the improvement over time and provides insights into whether the model is overfitting.The accuracy trends, both train and validation accuracy increase, indicating effective learning and generalization.

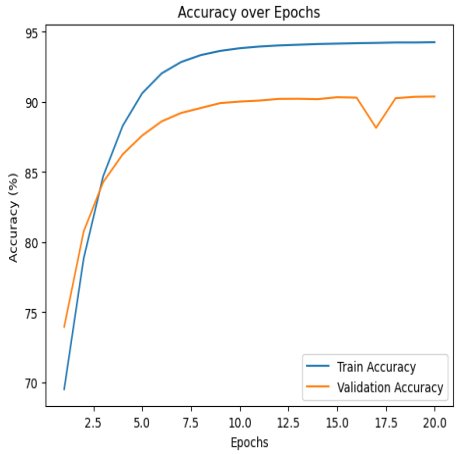

Fig. 3. Accuracy over Epochs

D. Discussion of Results

The model performed well in several languages, with Indonesian achieving the highest accuracy. This may be better due to better tokenization for languages with simpler structures or more available resources. Indonesian likely has more online text data, which aids the model in learning better representations.

However, the model struggled with Korean, achieving only 44.43% accuracy. This could be due to tokenization challenges with non-Latin scripts. Language-specific nuances not captured by the model, likely due to a lack of domain-specific features.

In general, the model demonstrated strong generalization for languages like English and Spanish. However, there are challenges in low-resource languages like Korean and German. Further improvements could be achieved by integrating pre-trained multilingual embeddings and refining tokenization techniques for non-Latin scripts.

V. Discussion

A. Interpretation of Results

The experiments show that the proposed CNN-based model with attention performs well on several multilingual datasets, with particularly strong performance on languages with rich online resources such as Indonesian and English. The validation accuracy of 90.37% at epoch 20 demonstrates that the model can generalize effectively across different datasets.However, test accuracy reveals significant performance disparities across languages. Indonesian (80.41%) achieves the highest accuracy, which is likely due to the availability of more abundant data and the model's alignment with the language's characteristics. On the other hand, languages like Korean (44.43%) and German (57.08%) see much lower accuracy. This discrepancy highlights the challenges of multilingual hate speech detection, especially for languages with fewer resources or more complex grammatical structures.

B. Challenges and Limitations

Several challenges became apparent during the testing phase:

- Tokenization Issues: The model uses a simple regular expression tokenizer, which is not ideal for non-Latin scripts or languages with complex morphology, like Korean. This likely contributed to the model's poor performance on languages such as Korean and Chinese.
- Language-Specific Nuances: The model's performance on Korean and German highlights the difficulties of applying a single model to diverse linguistic structures. For example, the Korean language, with its agglutinative structure, may require more sophisticated tokenization methods or domain-specific embeddings to capture the nuances of hate speech expressions.
- Resource Constraints for Low-Resource Languages: The model performed less effectively on languages with relatively less online data or fewer resources, such as German and Filipino. In these cases, the model struggled to achieve high accuracy, indicating that data scarcity is a major limitation.

C. Future Directions

Several avenues can be explored to further improve the performance of multilingual hate speech detection:

- Pretrained Multilingual Embeddings: Integrating pretrained models such as mBERT, XLM-R, or FastText would likely improve the handling of non-Latin scripts and low-resource languages by providing richer word representations.
- Advanced Tokenization: A more robust tokenization technique, such as subword tokenization, would help address the issues faced by languages with complex morphology, like Korean and Chinese.
- Fine-Tuning on Language-Specific Data: For languages with poor performance, fine-tuning the model on additional data from those languages could improve accuracy and generalization.
- Ensemble Models: Combining the CNN-based model with other architectures, such as LSTM or Transformer-based models, might help capture different linguistic features and improve performance across diverse languages.

D. Conclusion

This study demonstrates the potential of deep learning, specifically CNNs with attention mechanisms, for multilingual hate speech detection. While the model shows strong generalization capabilities for certain languages, the performance gap for languages like Korean and German highlights the need for language-specific improvements and richer language representations. The findings contribute valuable insights into the challenges of multilingual hate speech detection and provide a foundation for future research aimed at improving the robustness of these models across diverse languages and contexts.

REFERENCES

[1] T. Ranasinghe and M. Zampieri, "Multilingual offensive language identification with cross-lingual embeddings," in Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval), Barcelona, Spain (Online), Dec. 2020, pp. 127–134.

[2] B. Pelicon, P. Robnik-Šikonja, and B. Mozetič, "Investigating cross-lingual transfer for offensive language detection," Information Processing & Management, vol. 58, no. 3, 2021.

[3] D. Nozza, "Exposing the limits of zero-shot cross-lingual hate speech detection," in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Punta Cana, Dominican Republic, Nov. 2021, pp. 3652–3662.

[4] A. Eronen, D. Alharthi, and B. Plank, "Better transfer with linguistically similar languages: Revisiting zero-shot hate speech detection," in Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 2022, pp. 1862–1873.

[5] M. Monnar, L. Padró, and C. Saggion, "Domain-specific multilingual embeddings for hate speech detection," ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), vol. 23, no. 1, pp. 1–25, Jan. 2024.

[6] S. Vashistha and A. Zubiaga, "Fine-grained multilingual hate speech classification using multi-input CNN-BiLSTM networks," IEEE Transactions on Affective Computing, vol. 12, no. 2, pp. 477–490, Apr.–Jun. 2021.

[7] A. Mahajan, M. Tomar, and D. Rani, "EnsMulHateCyb: Ensemble-based multilingual hate speech and cyberbullying detection model," Applied Soft Computing, vol. 136, Jan. 2024.

[8] A. Hashmi, K. Gupta, and H. Awan, "A multilingual framework for hate speech detection using transformer models and data augmentation," Expert Systems with Applications, vol. 223, Apr. 2024.

[9] A. Samoshyn, "Hate Speech and Offensive Language Dataset," Kaggle, 2017. [Online]. Available: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset

[10] R. Urbano Jr., J. Uy, A. Angeles, M. N. Quintos, J. M. Imperial, and R. Rodriguez, "Filipino Hate Speech Text Dataset from Twitter," GitHub, 2021. [Online]. Available: https://github.com/imperialite/filipino-tiktok-hatespeech