Detecting Hate Speech and Fake Narratives in Code-Mixed Hinglish Social Media Text

Advaitha Vetagiri and **Partha Pakray**
Department of Computer Science and Engineering
National Institute of Technology Silchar
Email: (advaitha21_rs, partha)@cse.nits.ac.in

Abstract

The increasing prevalence of hate speech and fake narratives on social media platforms poses significant societal challenges. This study addresses these issues through the development of robust machine learning models for two tasks: (1) detecting hate speech and fake narratives (Task A) and (2) predicting the target and severity of hateful content (Task B) in code-mixed Hindi-English text. We propose four separate CNN-BiLSTM models tailored for each subtask. The models were evaluated using validation and 5-fold cross-validation datasets, achieving F1-scores of 74% and 79% for hate and fake detection, respectively, and 63% and 54% for target and severity prediction and achieved 65% and 57% for testing results. The results highlight the models' effectiveness in handling the nuances of code-mixed text while underscoring the challenges of underrepresented classes. This work contributes to the ongoing effort to develop automated tools for detecting and mitigating harmful content online, paving the way for safer and more inclusive digital spaces.

1 Introduction

The ICON 2024 Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)[1] (Biradar et al., 2024a) addresses a critical issue in the digital age: the intersection of fake narratives and hate speech. While enabling global connectivity, social media platforms have become hotspots for rapidly disseminating harmful and misleading content. This toxic combination exacerbates societal divisions, fosters animosity, and often leads to real-world consequences. The shared task emphasizes the need for innovative approaches to detect and counter this dual threat, particularly in code-mixed Hindi-English (Hinglish) text, which poses unique linguistic challenges.

Faux-Hate refers to the spread of hate speech rooted in fake narratives and misinformation deliberately constructed or shared to provoke hostility and divide communities. These instances often blend falsehoods with emotionally charged language, making them more potent and harder to detect. Addressing this problem requires a holistic approach to discerning hate speech and misinformation within diverse and noisy online environments.

We employed a CNN-BiLSTM (Vetagiri et al., 2024a) architecture to tackle the challenge for this shared task. This hybrid model combines the strengths of convolutional neural networks (CNNs) (Vetagiri et al., 2023) and bidirectional long short-term memory (BiLSTM) (Vetagiri et al., 2024b) networks. CNNs are adept at capturing local features and patterns within the text, while BiLSTMs excel in modelling sequential dependencies and contextual relationships. This combination makes the model well-suited for processing complex, code-mixed text, where both local nuances and global context play a critical role.

The Faux-Hate shared task is designed to:

- **Task A** Detect fake narratives and hate speech simultaneously.
  - *Fake*: Binary label indicating if the content is **fake** (1) or **real** (0).
  - *Hate*: Binary label indicating if the content is **hate speech** (1) or **not** (0).
- **Task B** Predict the target and severity of hateful content:
  - *Target*: Categorized as **Individual**, **Organization**, or **Religion**.
  - *Severity*: Classified into **Low**, **Medium**, or **High**.
- Encourage the development of methods robust enough to handle code-mixed Hindi-English text, a prevalent form of online communication in India.

[1] https://sai-kartheek-reddy.github.io/Shared-Task-on-Faux-Hate-Detection-at-ICON-2024

With the increasing prevalence of code-mixed content on social platforms, traditional NLP models often struggle to deliver accurate predictions due to linguistic diversity and noise. By integrating advanced architectures like CNN-BiLSTM, this task aims to push the boundaries of what's possible in multilingual and multimodal hate speech detection. This shared task contributes to a growing body of research to mitigate the harms caused by toxic and misleading online narratives, ensuring safer digital spaces for all.

2 Literature Survey

The detection of hate speech and fake narratives has advanced significantly with the adoption of machine learning (ML) and deep learning (DL) techniques. Early work predominantly relied on traditional ML approaches, such as Support Vector Machines (SVMs) and Naïve Bayes, combined with feature engineering techniques like term frequency-inverse document frequency (TF-IDF) and n-grams (Davidson et al., 2017; Waseem and Hovy, 2016). While effective for smaller datasets, these methods struggled with contextual understanding and often failed to generalize across languages and nuanced categories, such as implicit hate or sarcasm. Latent Dirichlet Allocation (LDA) has also been applied to uncover hidden topics within datasets, aiding in identifying hate-related themes (Waseem and Hovy, 2016).

Deep learning approaches have significantly improved performance in this domain. Convolutional Neural Networks (CNNs) are effective for capturing local patterns in text, while Long Short-Term Memory (LSTM) networks excel in modelling sequential data (Badjatiya et al., 2017; Putra et al., 2022). Hybrid architectures, such as CNN-BiLSTM, leverage the strengths of both frameworks, offering robust solutions for tasks involving code-mixed text, including Hindi-English datasets (Riyadi et al., 2024). Furthermore, transformer-based architectures like BERT and RoBERTa have set new benchmarks in detecting hate speech and fake narratives by leveraging pre-trained language models for better contextual understanding (Kenton and Toutanova, 2019). These models are particularly powerful in multilingual and low-resource settings, often fine-tuned on hate speech datasets to enhance domain-specific performance (Joshi et al., 2020; Mnassri et al., 2024).

Recent advancements include the use of generative adversarial networks (GANs) for data augmentation, which help mitigate class imbalance and improve robustness (Mnassri et al., 2024; Beddiar et al., 2021). Multimodal approaches, integrating textual and visual inputs, have been employed for tasks like detecting hateful memes and fake narratives involving images (Kiela et al., 2020). However, challenges persist in explainability, as deep learning models often act as "black boxes." This has motivated researchers to integrate explainable AI (XAI) techniques, ensuring more interpretable and ethical model predictions (Arrieta et al., 2020). Despite these challenges, the field continues progressing, with innovations targeting online content's nuanced and multilingual nature.

3 Dataset

The dataset for the Faux-Hate (Biradar et al., 2024b) shared task has been meticulously curated to reflect real-world challenges in detecting fake narratives and hate speech within code-mixed Hindi-English social media text. It consists of three files shared across the training, validation, and test phases, supporting both Task A: Binary Faux-Hate Detection and Task B: Target and Severity Prediction. For Task A, each text sample in the dataset is labelled with two binary annotations: *Fake*, indicating whether the content is **fake (1)** or **real (0)**, and *Hate*, indicating whether the content contains **hate speech (1)** or **not (0)**. Task B extends this with additional labels: *Target*, identifying the intended target of the hate speech as either **Individual (I)**, **Organization (O)**, or **Religion (R)**, and *Severity*, specifying the intensity of the hate speech as **Low (L)**, **Medium (M)**, or **High (H)**.

The dataset is released in three phases to facilitate model development and evaluation. Details of the dataset distribution are presented in Table 1, the training set comprising 6,396 samples. The validation set contains 800 samples. Finally, the test set consists of 800 samples.

Dataset	Number of Samples	Supports Task
Train	6,396	Task A & B
Validation	800	Task A & B
Test	800	Task A & B

Table 1: Dataset Distribution for Faux-Hate Tasks

Table 2, presents the statistics for Task A, with one table corresponding to the training set and the other to the validation set. These tables include

	0	1
Hate	2,295	4,101
Fake	3,110	3,286

(a) Train

	0	1
Hate	287	513
Fake	513	423

(b) Validation

Table 2: Statistics for Task A Train and Validation Datasets.

	N/A	O	I	R
Target	2,295	2,279	1,081	741
Severity	2,295	582	1,559	1,960

Table 3: Task B - Train Dataset Target and Severity Statistics

rows for the *Hate* and *Fake* labels, displaying their respective counts for both 0 (no) and 1 (yes). The subsequent two tables, 3 and 4, present the statistics for Task B, with separate tables for the training and validation sets. These tables display the distribution of the *Target* and *Severity* labels. The *Target* column refers to the categories: N/A (Not Applicable), O (Organization), I (Individual), and R (Religion), while the *Severity* column indicates the intensity levels: N/A, H (High), M (Medium), and L (Low). The dataset is provided in xlsx format, where each row corresponds to a text sample with its associated labels. For Task A, the files include the *Fake* and *Hate* columns, while for Task B, they also include the *Target* and *Severity* columns.

4 Methodology

In this work, we propose two separate deep learning models for addressing the tasks of detecting fake narratives and hate speech in code-mixed Hindi-English social media text. Both models are based on the CNN-BiLSTM architecture[2] as shown in figure 1, known for capturing local features and contextual dependencies in sequential data. The methodology for each task is described below.

4.1 Task A: Hate and Fake Detection

For Task A, which involves detecting whether a given text contains hate speech (`Hate = 1`) or not (`Hate = 0`) and whether the content is fake (`Fake = 1`) or real (`Fake = 0`), two separate models are developed. The architecture and workflow are as follows:

- **Text Representation:** Each input text sequence is tokenized and converted into dense vector embeddings using pre-trained word embedding techniques. These embeddings encode the semantic meaning of the words in a continuous vector space.

- **CNN Layer:** Convolutional layers are applied to capture local n-gram features in the text. The filters slide over the embeddings to extract significant patterns indicative of hate speech and fake narratives.

- **BiLSTM Layer:** The feature maps generated by the CNN are passed through a Bidirectional Long Short-Term Memory (BiLSTM) layer. The BiLSTM processes the sequence in both forward and backward directions, capturing the long-range dependencies and contextual relationships.

- **Classification Layers:**

	N/A	O	I	R
Target	287	274	140	99
Severity	287	74	182	257

Table 4: Task B - Validation Dataset Target and Severity Statistics

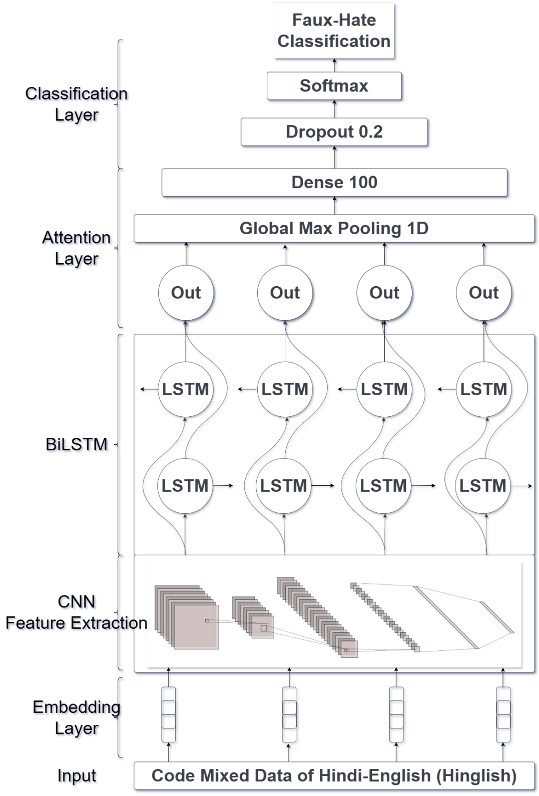


Figure 1: Proposed CNN-BiLSTM Model Architecture for Hate Speech and Fake Narrative Detection

[2] https://github.com/advaithavetagiri/CNLP-NITS-PP

  - The first model outputs a binary prediction for hate speech detection (`Hate = 1` or `Hate = 0`).
  - The second model outputs a binary prediction for fake narrative detection (`Fake = 1` or `Fake = 0`).

4.2 Task B: Target and Severity Prediction

Task B involves predicting the target (`N/A`, `O` for Organization, `I` for Individual, `R` for Religion) and severity (`N/A`, `H` for High, `M` for Medium, `L` for Low) of hateful content. Two models are developed for this task:

- **Text Representation:** As with Task A, word embeddings are used to convert the input text into a dense vector representation.
- **CNN Layer:** Convolutional layers are used to identify patterns and n-grams associated with different target categories and severity levels.
- **BiLSTM Layer:** The BiLSTM layer processes the output of the CNN, capturing the sequential and contextual dependencies within the text.
- **Classification Layers:**
  - The first model outputs a categorical prediction for the target of the hateful content (`N/A`, `O`, `I`, `R`).
  - The second model outputs a categorical prediction for the severity level (`N/A`, `H`, `M`, `L`).

4.3 Experimental Setup

The experiments are conducted using a Python-based deep learning framework. The training and validation datasets for each task are preprocessed to normalize text, tokenize words, and pad sequences to a fixed length. The following table summarizes the hyperparameters used for training both models:

4.4 Training and Optimization

Both models are trained separately using labelled datasets for their respective tasks. Each model minimizes a binary cross-entropy loss function during training, as shown in Tabel 5. The Adam optimizer is employed for parameter optimization, with a learning rate tuned for optimal performance. Dropout is applied to prevent overfitting during training.

Hyperparameter	Value
Embedding Dimension	300
Convolutional Filters	128
Kernel Size	3
LSTM Units	100
Dropout Rate	0.5
Batch Size	64
Optimizer	Adam
Learning Rate	0.001
Loss Function	Binary Cross-Entropy
Epochs	15

Table 5: Hyperparameters Used in Model Training

4.5 Evaluation Metrics

To evaluate the performance of the models, metrics such as precision, recall, and F1-score are computed for Task A. For Task B, weighted accuracy and macro F1-score are used to assess the models' ability to predict targets and severity levels. Results are reported separately for each task, ensuring a comprehensive analysis of all four models.

5 Results

The proposed CNN-BiLSTM models were evaluated on the validation dataset and through 5-fold cross-validation to assess their performance on the tasks of Hate and Fake detection (Task A). The results demonstrate the models' ability to effectively handle the challenges posed by code-mixed Hindi-English social media text.

5.1 Task A: Evaluation on Validation Data

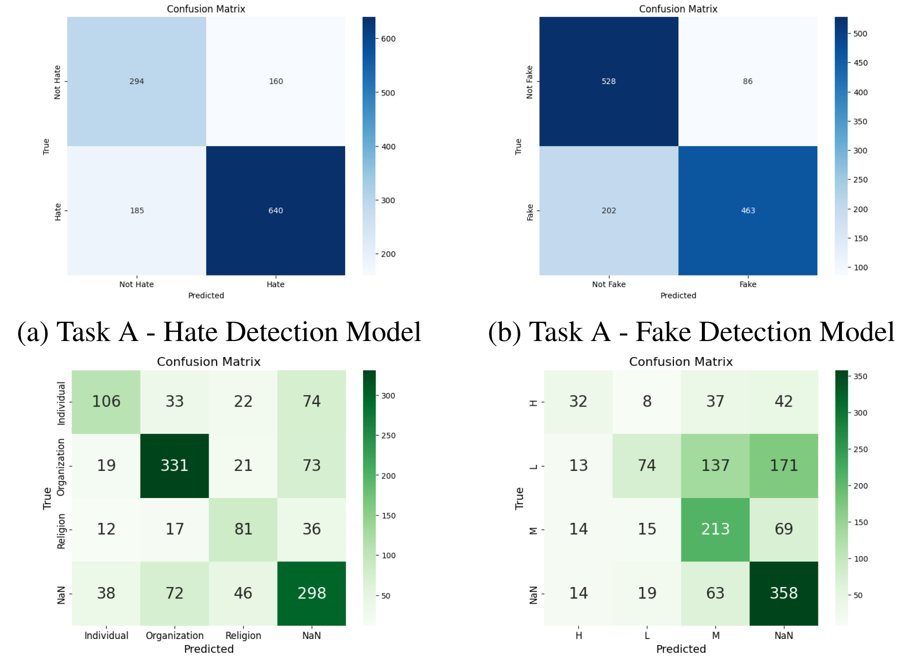
The performance of the models was first evaluated on the validation dataset after training on the training dataset. Table 6 presents the precision, recall, F1-score, and accuracy for the Hate detection and Fake detection models.

Model	Accuracy	Precision	Recall	F1-Score
Hate	74.0%	74.0%	74.0%	74.0%
Fake	77.6%	78.0%	77.0%	77.0%

Table 6: Task A Validation Dataset Results

5.2 Epoch-wise Training Performance

The epoch-wise training performance of the models was analyzed to track the progression of accuracy and loss during training and validation. Figure 2 depict the confusion matrix for the Hate and Fake detection models across the 5-folds.



(a) Task A - Hate Detection Model

(b) Task A - Fake Detection Model

(c) Task B - Target Prediction Model

(d) Task B - Severity Prediction Model

Figure 2: Confusion Matrices for All Models: (a) Hate Detection, (b) Fake Detection, (c) Target Prediction, (d) Severity Prediction.

5.3 Performance Insights

- **Hate Detection Model:** The Hate detection model achieved an average F1-score of 74% across 5-fold cross-validation. This reflects the model's ability to balance precision and recall for detecting hateful content.

- **Fake Detection Model:** The Fake detection model performed slightly better, with an average F1-score of 79%, indicating its robustness in detecting fake narratives.

- **Cross-Validation Consistency:** The minimal deviation in metrics across folds demonstrates the reliability and generalizability of the models.

- **Training Trends:** For both models, training accuracy steadily increased while validation accuracy plateaued around the 10th epoch, suggesting effective learning with no overfitting.

5.4 Task B: Target and Severity Models

The performance of the models for Task B, which involves predicting the target and severity of hateful content, was evaluated using both the validation dataset and 5-fold cross-validation. The results are detailed below.

5.4.1 Target Prediction Model

The Target prediction model classifies text into one of four categories: `Individual (I)`, `Organization (O)`, `Religion (R)`, and `N/A`. Table 8 shows the metrics for the validation dataset, while 5-fold cross-validation results are summarized in Table 7.

5.4.2 Severity Prediction Model

The Severity prediction model classifies the intensity of hate speech into one of four categories: `High (H)`, `Medium (M)`, `Low (L)`, and `N/A`. The validation dataset results are displayed in Table 8, and the 5-fold cross-validation results are presented in Table 7.

5.4.3 Performance Insights

- **Validation Set Results:**
  - For the Target prediction model, the best performance was observed for the `Organization (O)` and `N/A` classes, achieving F1-scores of 0.73 and 0.68, respectively.
  - For the Severity prediction model, the `N/A` and `Medium (M)` classes showed the highest F1-scores at 0.72 and 0.62, respectively.

Fold	Accuracy	Precision	Recall	F1-Score
Hate Detection				
Fold 1	72.0%	73.0%	72.0%	72.0%
Fold 2	77.0%	77.0%	74.0%	77.0%
Fold 3	75.0%	74.0%	75.0%	75.0%
Fold 4	74.0%	74.0%	70.0%	74.0%
Fold 5	73.0%	73.0%	73.0%	73.0%
Average	**74.2%**	**74.2%**	**72.8%**	**74.2%**
Fake Detection				
Fold 1	76.0%	77.0%	76.0%	76.0%
Fold 2	78.0%	79.0%	78.0%	79.0%
Fold 3	80.0%	80.0%	80.0%	80.0%
Fold 4	81.0%	81.0%	81.0%	81.0%
Fold 5	77.0%	78.0%	77.0%	77.0%
Average	**78.4%**	**79.0%**	**78.4%**	**78.6%**
Target Detection				
Fold 1	62.0%	60.0%	56.0%	56.0%
Fold 2	64.0%	61.0%	59.0%	60.0%
Fold 3	64.0%	61.0%	60.0%	60.0%
Fold 4	64.0%	61.0%	59.0%	60.0%
Fold 5	64.0%	65.0%	65.0%	64.0%
Average	**63.6%**	**61.6%**	**59.6%**	**60.0%**
Severity Detection				
Fold 1	54.0%	51.0%	50.0%	49.0%
Fold 2	55.0%	52.0%	53.0%	52.0%
Fold 3	55.0%	54.0%	50.0%	49.0%
Fold 4	55.0%	52.0%	53.0%	52.0%
Fold 5	53.0%	55.0%	55.0%	53.0%
Average	**54.4%**	**52.8%**	**52.8%**	**51.0%**

Table 7: Task A 5-Fold Cross-Validation Results

Model	Accuracy	Precision	Recall	F1-Score
Target	64.0%	65.0%	64.0%	64.0%
Severity	54.0%	54.0%	54.0%	52.0%

Table 8: Task B Validation Dataset Results

- **Cross-Validation Trends:**
  - The consistent results across folds indicate the robustness of the models. However, performance for underrepresented classes, such as `High (H)` in Severity prediction and `Religion (R)` in Target prediction, remains a challenge.
  - Improvements in model performance for low-support classes may require additional training data or specialized feature engineering.

5.5 Test Results

The performance of the proposed models was evaluated on the test dataset for both Task A and Task B. The results are presented in Table 9. For Task A, which involves detecting hate and fake narratives, the model achieved a Macro F1-score of 0.65. Similarly, for Task B, which predicts the target and severity of hateful content, the model achieved a Macro F1-score of 0.57. These results highlight the challenges of handling code-mixed text and imbalanced datasets, particularly in the context of nuanced classification tasks.

Task	Model Run	Macro F1-Score
Task A	CNLP-NITS-PP run1	0.65
Task B	CNLP-NITS-PP run1	0.57

Table 9: Test Results for Task A and Task B

6 Conclusion

This work addresses the critical challenge of detecting hate speech and fake narratives in code-mixed Hindi-English social media text and predicting the target and severity of hateful content. We proposed four separate CNN-BiLSTM models tailored for these tasks, demonstrating effective performance on validation and cross-validation datasets. For Task A, the Hate and Fake detection models achieved consistent results, with F1-scores of 74% and 79%, respectively, across 5-fold cross-validation. The robustness of the models was evident in their ability to generalize well on the validation set, with minimal performance drop. For Task B, the Target and Severity prediction models achieved F1-scores of 63% and 54%, respectively, highlighting the complexities involved in handling multi-class classification tasks, especially for underrepresented categories. The results underscore the challenges inherent in processing code-mixed text, particularly when dealing with nuanced tasks such as identifying fake narratives and varying levels of hate intensity. The models' performance demonstrates the potential of deep learning techniques to tackle such problems while also revealing areas for improvement, such as enhancing class representation for underrepresented categories like `Individual`, `Religion`, `High`, and `Low`.

Acknowledgements

We extend our heartfelt gratitude to Faux-Hate organisers and the National Institute of Technology Silchar for providing the opportunity and support to conduct our research and experimentation. We are particularly thankful for the resources and conducive research environment offered by the CNLP & AI Labs, NIT Silchar, which greatly facilitated the progress of our work.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Djamila Romaissa Beddiar, Md Saroar Jahan, and Mourad Oussalah. 2021. Data expansion using back translation and paraphrasing for hate speech detection. *Online Social Networks and Media*, 24:100153.

Shankar Biradar, Sai Kartheek Reddy Kasu, Sunil Saumya, and Md. Shad Akhtar, editors. 2024a. *Proceedings of the 21st International Conference on Natural Language Processing (ICON): Shared Task on Decoding Fake Narratives in Spreading Hateful Stories (Faux-Hate)*. Association for Computational Linguistics, AU-KBC Research Centre, MIT College, India.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2024b. Faux hate: unravelling the web of fake narratives in spreading hateful stories: a multi-label and multi-class dataset in cross-lingual hindi-english code-mixed text. *Language Resources and Evaluation*, pages 1–32.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Khouloud Mnassri, Reza Farahbakhsh, and Noel Crespi. 2024. Multilingual hate speech detection: A semi-supervised generative adversarial approach. *Entropy*, 26(4):344.

Bagas Prakoso Putra, Budhi Irawan, Casi Setianingsih, Annisa Rahmadani, Farradita Imanda, and Izzu Zantya Fawwas. 2022. Hate speech detection using convolutional neural network algorithm based on image. In *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, pages 207–212.

Slamet Riyadi, Annisa Divayu Andriyani, and Siti Noraini Sulaiman. 2024. Improving hate speech detection using double-layers hybrid cnn-rnn model on imbalanced dataset. *IEEE Access*.

Advaitha Vetagiri, Prottay Adhikary, Partha Pakray, and Amitava Das. 2023. CNLP-NITS at SemEval-2023 task 10: Online sexism prediction, PREDHATE! In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 815–822, Toronto, Canada. Association for Computational Linguistics.

Advaitha Vetagiri, Prateek Mogha, and Partha Pakray. 2024a. Multilate classifier: A novel ensemble of cnn-bilstm with resnet-based multimodal classifier for ai-generated hate speech detection. *SSRN*.

Advaitha Vetagiri, Partha Pakray, and Amitava Das. 2024b. A deep dive into automated sexism detection using fine-tuned deep learning and large language models. *SSRN*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.