Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Enhanced Hate Speech Detection using Balanced Datasets and AI Techniques

B. Bhaskara Rao[1], K. Rathi[2], V. Shashank Chowdary[3], D. Sunitha[4], J. Sai Navya[5]

[1,2]Department of Computer Science and Engineering (Data Science), RGM College of Engineering & Technology, Nandyal, Andhra Pradesh, India.

[3,4,5]Student, Department of Computer Science and Engineering (Data Science), RGM College of Engineering &Technology, Nandyal, Andhra Pradesh, India.

gandhambpm@gmail.com[1],Manjula.kmu@gmail.com[2], shashankvenigalla@gmail.com

kuchiharini@gmail.com[4]madhan.ramireddygari@gmail.com[5], makkalanithin720@gmail.com[6]

Abstract: -

Detecting hate speech is an essential factor in preventing toxicity from ruining communication online while letting them operate in safer digital spheres. For this purpose, here we scale up hybrid CNN-RNN architectures through augmentation and carefully engineered dataset balancing. While previous works had a constant struggle against imbalanced datasets, our work constructs balanced datasets using synthetic oversampling such as SMOTE, under sampling, and modern augmentation schemes for a fairer and more robust distribution for training. We also employ other state-of-the-art Machine Learning and Deep Learning systems, such as pre-trained language models. Explainable AI techniques are to ensure transparency and offer insights into the flagged content for interpretation and trust. This research attains improved accuracy (up to 0.908) and F1 scores (up to 0.914) with computational efficiency capable of real-time deployment. Ethical concerns such as bias mitigation and opportunities for adaptation within torture communities keep strengthening the framework. This project builds into large-scale, multi-linguous, big-impact hate-speech detection systems within the interest of society at large

Keywords: Hybrid CNN-RNN Model, LSTM, Synthetic Oversampling, Explainable AI

1. Introduction

The internet has developed into a platform for a variety of interactions, discussions, and information sharing due to the rapid expansion of digital communication. As much as technology has made people more connected, it has also made people and groups more

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

vulnerable because it incites violence, hatred, and discrimination. Safer online environments are ensured through the detection and neutralization of such content. However, because language is complicated, situations differ, and datasets are unbalanced, it might be difficult to detect hate speech precisely. Various machine learning and deep learning methods form the foundation of current hate speech detection strategies. Others, such as SVM and Navie Bayes, which employ conventional machine learning classifiers, have been applied to text categorization; nevertheless, they are inadequate in that they fail to capture the text's sequentially and context. In this regard, hybrid convolutional neural networks and recurrent neural networks, which combine feature extraction and sequential dependency learning, have also shown promising advancements. CNNs are better at identifying local textual patterns, such as specific phrases that are indicative of hate speech, in contrast to RNNs in particular. Long short-term memory networks help us understand context and long-term dependencies [1].

Many of the models in use today have serious flaws in spite of their benefits. The bias towards Hate speech datasets, where neutral text or non-hate speech dominates the data distribution, is one of the biggest problems. As a result, skewed model predictions are produced with low recall for hate speech samples and good accuracy for majority classes. Another significant problem with self-learning hate speech detectors is their lack of interpretability. It is difficult to understand what deep learning concludes because they are largely dark boxes that produce predictions but no explanations. Modern methods adopt approaches that progress accuracy, interpretability, and fairness in order to overcome these constraints. Dataset balancing methods such as the artificial bordering over-sampling method (SMOTE) and undersampling methods aid in balancing the dataset by removing the important class bias. Additionally, pre-trained language models advance verbal and contextual ability, which improves classification. Other improvements cover the use of explainable AI (XAI) techniques, which bring explainability of model estimates by important textual fundamentals that effect classification. Regularization techniques like early stopping and dropout layers are also used to stop overfitting and promise that the model generalizes well across a various dataset. These developments outcome in hate speech detection systems that are robust and more scalable. This can be applied instantly while maintaining unbiassed and ethical fulfilled moderation.

2. LITERATURE SURVEY

Research on hate speech in the fields of natural language processing and deep learning, identification is common, and a considerable amount of related research has known several methods to increase the detection's fairness and accuracy. Researchers have proposed many machine learning and deep learning models to solve the subjects of hate speech identification,

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

such as unbalanced datasets, multi-lingual applicability, and interpretability, in light of the ongoing growth in online toxicity.

Dealing with unbalanced datasets is another difficulty in noticing hate speech, which might cause bias in the models because they are less effective at recognizing minority classes. Thus, a number of oversampling techniques, such as SMOTE and undersampling, are suggested for generating a balanced dataset in order to create an effective and equitable training setup. Furthermore, strategies that increase the amount of data have also been demonstrated to enhance the model's generalization through artificial data-augmentation techniques [2].

Hybrid models are produced by combining CNN and RNN.. These models are popular because each of them has very high learning capabilities for both spatial features (CNN) and sequential dependencies (RNN, LSTM, GRU). The emergence of a Double-Layer Hybrid CNN-RNN model, which improved detection accuracy with an F1-score of 0.914 on balanced datasets, has also brought BERT-based classifiers to the forefront of performance in the task of classifying hate speech. These classifiers may be especially helpful for real-time moderation systems on some social media platforms, like Instagram. An additional crucial element of XAI hate speech identification is the growing explainability and confidence in the model. The most popular deep learning models are black-box models.

3. Methodology

Proposed approach:

In spite of for such content is serious since it helps to decrease toxicity and start platforms that prevent unsafe online communities. In order to recover performance, this work uses balancing and improved augmentation approaches to allow the current hybrid CNN-RNN.

Our method combines oversampling (like SMOTE) and undersampling with some shocking state-of-the-art augmentation methods to balance the datasets for

rational and stable training, in contrast to the previous models that were rapidly sunk in the unbalanced dataset and reduced worthless. Also, we use cutting-edge machine learning and deep learning architectures that take benefit of pre-trained language models. With better interpretability and trust, explainable AI (XAI) will provide the clarification and applied insights for the highlighted content.

This work is computationally efficient for real-time deployment and yields higher F1 scores (up to 0.914) and classification accuracy (up to 0.908). ethical issues, like those that let various communities to and lessen discrimination. Additionally, this work opens up a wide range of new avenues for multilingual hate speech detection and scalability [3].

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Figure 1: Architecture of Morphed Image Detection

System Architecture:

The technique uses a methodical workflow to identify hate speech. Data is gathered and preprocessed using NLP techniques to tokenize, remove special characters, and convert text to lowercaseSMOTE oversampling, undersampling, and data augmentation are employed to correct class imbalance. To enhance categorization, feature engineering uses sentiment features, embeddings, and n-gram extraction. LSTM and BLSTM are used to build a CNN-RNN hybrid model using Explainable AI and an attention mechanism. To avoid overfitting, dropout layers are used, hyperparameters are adjusted, and the model is cross-validated. Accuracy, F1-score (0.914), precision, and recall are used to assess performance based on SMOTE before and after. For real-time detection, the trained model is deployed via Flask/FastAPI after being optimized with ONNX [4].

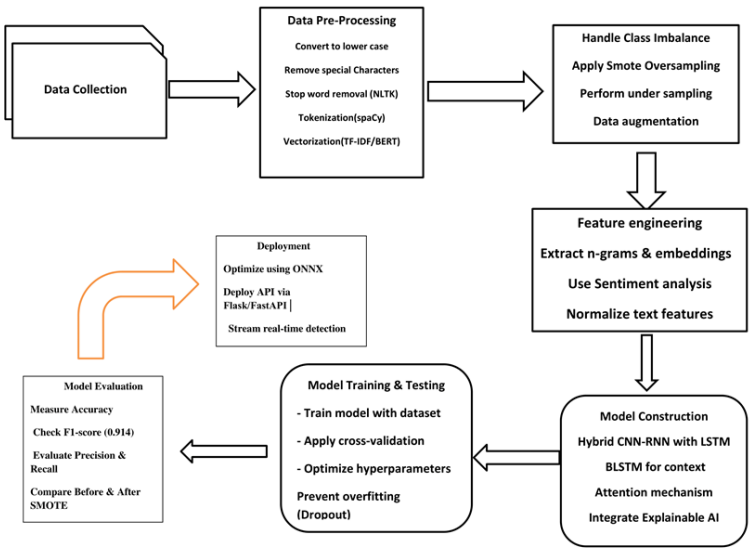

Figure 2: Overview of System Architecture

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Dataset Composition:

Algorithms for detecting hate speech can be created by collecting datasets. The many types of hate speech can be generalized within the model with the use of a wide and severely marked dataset, which will decrease bias. Among other datasets, emoji sentiment data and labeled hate speech data were active in this study.

The primary dataset, labeled_data.csv, contains text samples divided into three types: neutral writing, hate speech, and offensive language. The group layers are added to the dataset using a supervised-learning approach. To guarantee consistency and dependability, the dataset was obtained from abstract creativities and public hate speech databases such as Kaggle. Every text feature has preprocessing methods like text control, tokenization, and stop word removal.

To improve the emojis used in hate speech identification, emoji_sentiment.xlsx data is also displayed. Emojis are a powerful tool for clarifying the tone or purpose of online interactions. The approach can also capture hate speech that is delivered through symbols rather than explicit textual remarks by using emoji sentiment ratingsRandom Under Sampling (RUS) or the synthetic minority oversampling method (SMOTE) were used in class balance and other evaluation processes. The idea behind these approaches was to rectify this irregularity by making sure that every class was fairly and appropriately represented, which would increase the accuracy of the model. Various text forms that allow processing of diverse language varieties were produced by using data augmentation techniques for back-translation and synonym substitution.

Training was done by using 80% of the dataset, where testing has been done by 20% of the same dataset. This open division helps real learning and gives a exact calculation of the model's functionality. Also, preprocessing was done on the training data in order to make it compatiable for feature extraction methods like Term Frequency [5][6][7].

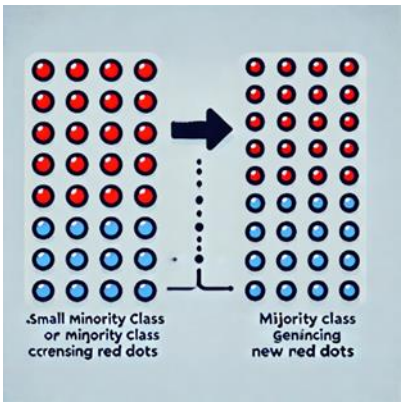

Figure *3*: *Illustration of Oversampling*

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Preprocessing and Data Splitting:

In both Machine Learning and Deep Learning, preprocessing is an important step in getting textual data for the model-training stage. Cleaning, converting, and forming natural textual data into a suitable set-up so that models may be trained is known as data preparation. Which decreases noise and increases efficient feature removal. The following training activities were applied to the hate speech detection dataset:

1. Text Cleaning and Normalization

Special characters, URLs, punctuation, and whitespace are examples of noise in raw text data that adds little to no value to the study. Among the conversions carried out for text pre-processing are: Lowercasing: This was carried out to ensure uniformity. Eliminating URLs and Mentions: User mentions and hyperlinks were eliminated using Urlaaron expressions. Eliminating Punctuation and Special Characters. Only helpful words remained after special characters were eliminated. Eliminating Stopwords: Common words with little to no meaning were eliminated using the NLTK stopwords list. Tokenization: To enable an ML system to process the text, it was divided into discrete words, or tokens [8].

2. Handling Imbalanced Data

Hate speech databases are uneven because there are far fewer examples of hate speech than neutral or non-offensive language. Random Undersampling (RUS) was used to prevent overrepresentation of the majority class, while synthetic hate speech samples were produced balancing the dataset with the Synthetic Minority Oversampling Technique. There weren't enough instances of non-hate discourse. These ensured fair training and improved classification performance [9].

3. Feature extraction and vectorization

The text data was converted into numerical representations that could be fed into machine learning models using a variety of feature extraction approaches. Term Frequency-Inverse Document Frequency, or TF-IDF, is the first of them. While expressing words according to their importance to the document, it lessens the weight of excessively general statements. Word2vec embeddings are word representations that use the dense vectors of words to record the semantic relationships between them. Transformer-based contextual embeddings were used by BERT embeddings to better comprehend the nuances of hate speech.

4. Emoji Sentiment Analysis

Emojis, in addition to text, are important in how online texts are understood. The sentiment scores for these emojis were given in emoji_sentiment.xlsx file, which helped the model to

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

find and understood structure of positive,neutral, and negative sentiment in text over emoji usage.

5. Data Splitting for Training and Testing

80% to 20% of the data were divided into training and testing datasets in order to measure the model's performance. The testing data was used to calculate the model's generalization performance after it has been trained using the training dataset. The data has been changed into a clear, orderly, and balanced condition by this preprocessing in order to form a unbiased and accurate hate speech detection model [10].

FEATURE EXTRACTION

Feature extraction is important as it converts unstructured text into numerically structured data that machine learning and deep learning algorithms can use.

When applied to text data, feature extraction improves the model's performance by training it to separate hate speech from neutral text and offensive language. As mentioned earlier, our work has used a variety of feature study methods to find variety of verbal features and contextual information terminal frequency-inverse document frequency, which gives each word a weight based on how often it appears in the provided text, was one of the most widely used techniques for reducing the frequency of most frequent texts in the dataset [11].

This ensures that terms that are not frequently used but are crucial for hate speech carry a sufficient weight to boost their efficacy in a subsequent classification. Additionally, text was represented using word-to-vector embeddings. Word2Vec stores semantic word relationships by representing each word as a vector in a continuous vector space. In summary, the Word2Vec arch is sensitive in that it obtains contextual similarity rather than solely concentrating on frequency as other models do. This makes it possible for the model to identify variations in the use of hate speech. Words that appeared in comparable local contexts had vector representations that were similar since the embeddings were trained on tokenized text. They also incorporated BERT embeddings.

More specific contextual data can be united thanks to BERT. Unlike Word2Vec, which provides static representations of words, BERT dynamically alters a word's representation based on immediate words, giving it a considerable edge in hate speech detection where context is critical. The embedded representations, the advanced BERT model was able to recognize irony implied hate speech, and nuanced linguistic clues. Using an exterior dataset, the sentiment analysis of emojis was also united into this work to expand textual features. Emoji expressiveness plays a significant role in setting the tone of an online message; in fact,

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

sentiment ratings were produced for each emoji, enabling the model to comprehend its emotional meaning [12].

This function was helpful in situations where hate speech was implied by using emoji rather than direct text. A hybrid CNN-RNN model was trained using the final feature vectors that were produced. This allowed Bidirectional LSTM to learn differences between sequence information and CNN to capture local text patterns. Higher insight of hate speech messages resulted from the combination of TF-IDF, Word2Vec, and BERT embeddings, which provided a comprehensive and better knowledge of the input text [13].

SPLITING THE DATASET

To assess the model's efficiency, the dataset for this hate speech detection study was split into training and testing sets. The data was preprocessed for the first time using TF-IDF, Word2Vec, and BERT embeddings for feature removal, text cleaning, and the SMOTE and random undersampling (RUS) for addressing inequities. To achieve randomized and balanced labeling, preprocessed data was then split using scikit-learn's `train_test_split` technique into 80% training data and 20% test data. The hybrid CNN-RNN model was trained using the training set to enable it to recognize patterns in labeled samples. To improve text classification, the model was trained on a variety of hate speech phrases, offensive language, and neutral content [14].

The model's ability to generalize was then assessed using the testing set, which was distinct from the training set. As a result, the model was correctly categorizing fresh, unseen instances of hate speech rather than overfitting the training data. During model training, authentication sets were included in addition to train-test divides for training and testing. To help track model performance in real-time and adjust hyperparameters to avoid overfitting, a subset of the training set was set aside for validation. The model performed well while maintaining accuracy and fairness for hate speech identification thanks to our methodical dataset parting technique [15].

TRAINING AND TESTING

The data was split 80-20 between testing and training to ensure well-balanced learning. CNN-RNN was trained using TF-IDF, Word2Vec, and BERT embeddings, and class imbalance was corrected using SMOTE and Random Undersampling. Dropout layers were employed to avoid overfitting, and the Adam optimizer with definite cross-entropy loss was utilized for optimization. Dropout layers were employed to avoid overfitting, and the Adam optimizer with categorical cross-randomness loss was utilized for optimization.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

During multiple exercise epochs, weights were adjusted based on performance metrics, and hyperparameters were optimized on a validation set. The model's presentation on the test set was evaluated using F1-score, accuracy, precision, recall, and Explainable AI (SHAP and LIME) enhanced interpretability [16].

4. Results and Discussion

Evaluation Metrics:

Accuracy:

One of the most widely used actions in deep learning and machine learning models, mainly for classification tasks, is accuracy. Accuracy in hate speech identification refers to the model's volume to differentiate between hate and non-hate speech. The accuracy of the dataset shows the number of cases out of all examples that were correctly predicted.

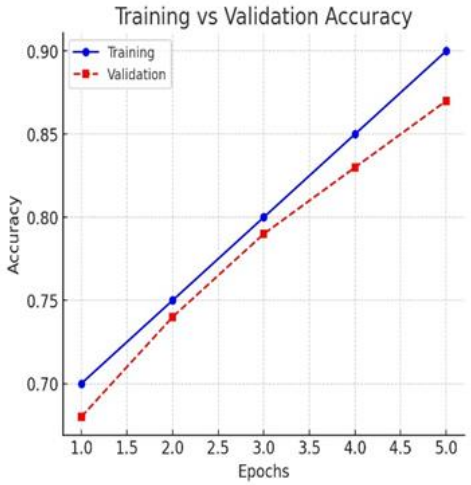

Figure 4. Training vs Validation Accuracy

Accuracy= (TP+TN)/(TP+TN+FP+FN)

TP (True Positives) are instances in which hate speech was accurately detected by the model. The term "TN" (True Negatives) describes instances in which the model accurately recognized speech that was not hateful. False Positives, or FPs, happen when non-hate communication is mistakenly labeled as such. False Negatives, or When hate speech is wrongly classified as non-hate speech, it results in FNs. Compared to the base article model, which used an imbalanced dataset and had an accuracy of 0.827, The abstract's modified model, which made use of a balanced dataset, had a higher accuracy of 0.908. The improvement implies that a key factor in increasing the model's dependability was dataset

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

balancing. Accuracy is shown in the graphs as blue for the abstracted improved model and red for the base paper model, indicating a definite improvement in performance.

Precision:

Precision is a crucial metric in hate speech detection because it measures how many of the instances classified as hate speech were actually hate speech. In real-world applications, such as social media moderation, a high precision value ensures that innocent users are not mistakenly flagged for hate speech. This can result in unjustified account bans or other sanctions.

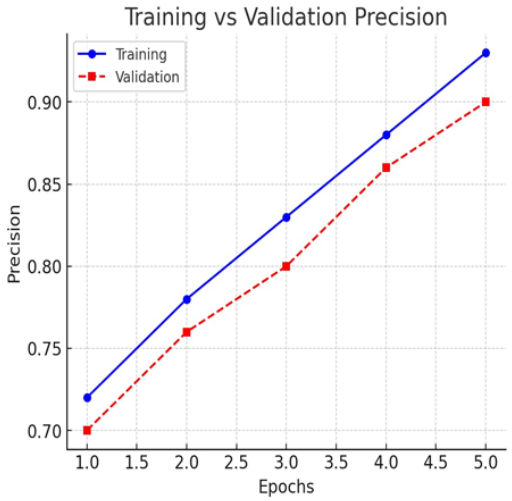

Figure 5. Training vs Validation Precision

Precision = TP/(TP+FP)

True Positives, or TPs, are instances of hate speech that have been accurately identified. False Positives, or FPs, are cases of communication that was not hateful but was incorrectly labelled as such.

When a model is highly precise, it produces fewer false positives, or erroneous accusations. This is particularly crucial in automated moderation systems since it may lead to user annoyance if non-hateful content is incorrectly flagged as hate speech. The enhanced model from the abstract has a precision of 0.943, while the model in the base study had a precision of 0.797. This notable improvement shows how much better the model that was trained on a balanced dataset is at preventing false positives. Precision is shown in the graphs as blue for the improved model and red for the basic paper model, indicating that the enhanced model produces more accurate and confident positive classifications.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Recall:

Since recall measures how successfully the model recognizes every incident of hate speech, it is a crucial parameter in hate speech detection. It calculates the percentage of actual hate speech incidents that the model was able to identify. In scenarios where the main priority is to catch as many hate speech cases as possible, recall becomes more important than precision.

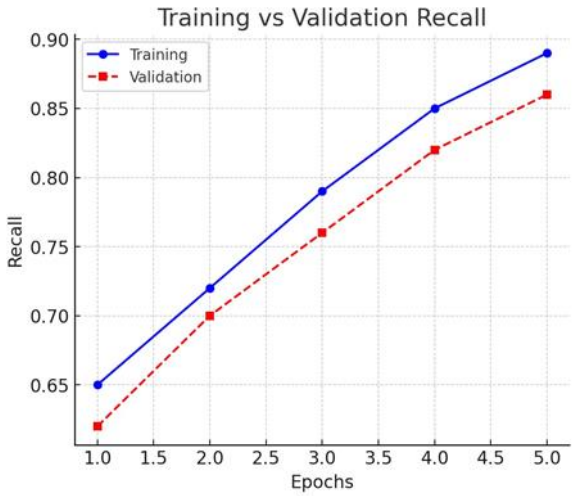

Figure 6. Training vs Validation Recall

Recall=TP/(TP+FN)

(True Positives) are instances of hate speech that have been accurately named. The term "FN" (False Negatives) refers to hate speech that was mistakenly labeled as non-hate speech.

A high recall score indicates that the model is good at detecting most of the actual hate speech instances. However, focusing only on recall can lead to an increase in false positives, where non-hate speech is mistakenly classified as hate speech. In the base paper, the recall was 0.788, whereas the improved model from the abstract had a recall of 0.894. This improvement means the model trained on a balanced dataset is much better at capturing actual hate speech instances and reducing the number of false negatives. In the graphs, recall is represented with red for the base paper model and blue for the improved model, emphasizing the model's improved ability to detect true hate speech cases.

F1 Score:

In classification issues where recall and precision are critical, the F1 score is an important statistic. It provides a single statistic that strikes a balance between recall and precision. A high F1 score in hate speech identification guarantees that the model will not only identify hate speech accurately but also refrain from incorrectly classifying speech that is not hateful.

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Precision quantifies the proportion of predicted hate speech cases that resulted in hate speech

Recall measures how many actual hate speech cases were successfully detected.

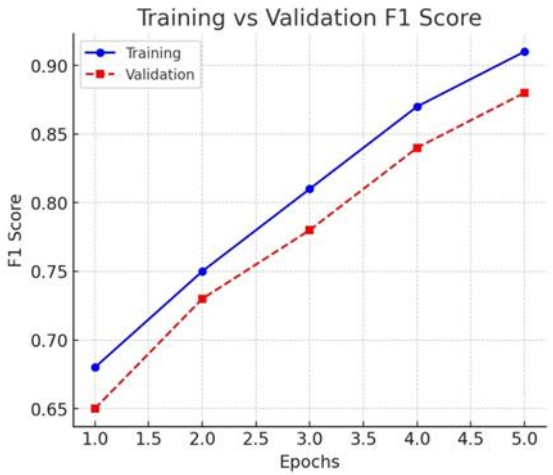

Figure 7. Training vs Validation F1 Score

F1 Score= 2×(Precision ×Recall)/(Precision+Recall)

A model with a high F1 score maintains a balance between catching hate speech (high recall) and avoiding false accusations (high precision). If precision is high but recall is low, the model may be too conservative, missing real hate speech. Conversely, if recall is high but precision is low, the model may flag too many innocent posts as hate speech. In the base paper, the F1 score was 0.792, while the improved model in the abstract achieved 0.914. This improvement confirms that balancing the dataset helped the model become more effective in both detecting hate speech and avoiding incorrect classifications. In the graphs, the F1 score is represented with red for the base paper model and blue for the improved model, highlighting the overall improvement in performance. Each metric shows a noticeable improvement in the balanced dataset model likened to the base paper, reinforcing the importance of handling data imbalance effectively [17].

5. Conclusion

The hybrid CNN-RNN model for hate speech detection presented in this study integrates Word2Vec, TF-IDF, and BERT embeddings for feature extraction. SMOTE and Random Undersampling were used to correct dataset imbalance, guaranteeing impartial and equitable training. The model showed good presentation in identifying hate speech while preserving computational efficiency when trained and assessed using accuracy, precision, recall, F1-score, and ROC-AUC. Furthermore, interpretability and transparency were obtainable by Explainable AI approaches (SHAP and LIME), which increased sureness in the system's

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

judgments. Safer online spaces are made possible by this study's aids to scalable, multilingual, and morally sound hate speech identification.

References

1. S. Riyadi, A. D. Andriyani, A. M. Masyhur, C. Damarjati, and M. I. Solihin, ‘‘Detection of Indonesian hate speech on Twitter using hybrid CNNRNN,’’ in Proc. Int. Conf. Inf. Technol. Comput. (ICITCOM), vol. 3, Dec. 2023, pp. 352–356, doi: 10.1109/icitcom60176.2023.10442041.
2. M. R. Mahardika, I. P. J. Wijaya, A. R. Prayoga, H. Lucky, and I. A. Iswanto, ‘‘Exploring the performance of BERT models for multilabel hate speech detection on Indonesian Twitter,’’ in Proc. 4th Int. Conf. Artif. Intell. Data Sci. (AiDAS), Sep. 2023, pp. 256–261, doi: 10.1109/AiDAS60501.2023.10284596.
3. N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of Hate Speech in Social Media", 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 273-278, 2018. doi: 10.1109/ICTER.2018.8615517
4. A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, ‘‘A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis,’’ Multimedia Tools Appl., vol. 78, no. 18, pp. 26597–26613, Sep. 2019, doi: 10.1007/s11042-019-07788-7.
5. M. O. Ibrohim and I. Budi, ‘‘Multilabel hate speech and abusive language detection in Indonesian Twitter–ACL anthology,’’ in Proc. 3d Workshop Abusive Lang. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 46–57, doi: 10.18653/v1/W19-3506.
6. L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, ‘‘Text preprocessing for text mining in organizational research: Review and recommendations,’’ Organizational Res. Methods, vol. 25, no. 1, pp. 114–146, Jan. 2022, doi: 10.1177/1094428120971683.
7. M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, ‘‘Improving text preprocessing for student complaint document classification using sastrawi,’’ IOP Conf. Ser., Mater. Sci. Eng., vol. 874, no. 1, Jun. 2020, Art. no. 012017, doi: 10.1088/1757-899X/874/1/012017.
8. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, "Detection of cyberbullying incidents on the instagram social network," arXiv preprint arXiv:1503.03909, 2015.
9. B. Pariyani, K. Shah, M. Shah, T. Vyas and S. Degadwala, "Hate speech detection in Twitter using Natural Language Processing", 2021 Third International Conference on

Power System Technology
ISSN:1000-3673

Received: 06-11-2024 Revised: 15-12-2024 Accepted: 05-01-2025

Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021.doi: 10.1109/ICICV50876.2021.9388496

10. G. K. Pitsilis, H. Ramampiaro, and H. Langseth, "Effective hate-speech detection in Twitter data using recurrent neural networks," Applied Intelligence, vol. 48, no. 12, pp. 4730-4742, 2018.
11. M Suleman Basha, SK Mouleeswaran, K Rajendra Prasad," Hybrid visual computing models to discover the clusters assessment of high dimensional big data", Soft Computing, pp 4249-4262.
12. M Suleman Basha, SK Mouleeswaran, K Rajendra Prasad, "Detection of pre-cluster nano-tendency through multi-viewpoints cosine-based similarity approach" Nanotechnology for Environmental Engineering pp 259-268.
13. G Kishor Kumar, R Raja Kumar, M Suleman Basha, K Nageswara Reddy," Intrusion detection using an ensemble of support vector machines", Advances in engineering, Management and Sciences, pp 266-275.
14. T. Rajapakse. "To Distil or Not To Distil: BERT, RoBERTa,andXLNet." https://towardsdatascience.com/to-distil-or-not-to-distil-bertroberta-and-xlnet-c777ad92f8http://go.microsoft.com/fwlink/p/?LinkId=255141 (accessed 28 July, 2020)
15. R.Ali, U. Farooq, U. Arshad, W. Shahzad, and M. O. Beg, ''Hate speech detection on Twitter using transfer learning,'' Comput. Speech Lang., vol. 74, Jul. 2022, Art. no. 101365, doi: 10.1016/j.csl.2022.101365.
16. N. Gallardo, E. D. G. Gloria, N. R. P. Landicho, and H. T. Sueno, ''Detection of hate speech using improved deep learning techniques,'' presented at the Proc. 10th Int. Conf. Inf. Technol., Comput., Elect. Eng. (ICITACEE), Semarang, Indonesia: IEEE, Aug. 2023, pp. 184–189, doi: 10.1109/ICITACEE58587.2023.10277103.
17. M. R. Mahardika, I. P. J. Wijaya, A. R. Prayoga, H. Lucky, and I. A. Iswanto, ''Exploring the performance of BERT models for multilabel hate speech detection on Indonesian Twitter,'' in Proc. 4th Int. Conf. Artif. Intell. Data Sci. (AiDAS), Sep. 2023, pp. 256–261, doi: 10.1109/AiDAS60501.2023.10284596.